UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Industrial

Caracterización de la variabilidad temporal en bases de datos médicas y estudio de su impacto en modelos predictivos basados en inteligencia artificial

Trabajo Fin de Grado

Grado en Ingeniería Biomédica

AUTOR/A: Fernández Narro, David

Tutor/a: Sáez Silvestre, Carlos

Cotutor/a: Ferri Borredà, Pablo

CURSO ACADÉMICO: 2022/2023

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ESCUELA TÉCNICA SUPERIOR INGENIERÍA INDUSTRIAL VALENCIA

TRABAJO FIN DE GRADO EN INGENIERÍA BIOMÉDICA

CARACTERIZACIÓN DE LA VARIABILIDAD TEMPORAL EN BASES DE DATOS MÉDICAS Y ESTUDIO DE SU IMPACTO EN MODELOS PREDICTIVOS BASADOS EN INTELIGENCIA ARTIFICIAL

AUTOR: DAVID FERNÁNDEZ NARRO

TUTOR: CARLOS SÁEZ SILVESTRE

COTUTOR: PABLO FERRI BORREDÀ

Curso Académico: 2022-2023

AGRADECIMIENTOS

En primer lugar, me gustaría agradecer a mis tutores, el Dr. Carlos Sáez Silvestre y Pablo Ferri Borredà por haberme guiado en la realización de este trabajo y por los conocimientos transmitidos en todos los ámbitos, tanto a nivel educativo como personal, así como por la confianza y el apoyo proporcionado en cada fase del proyecto.

En segundo lugar, agradecer a toda mi familia, pero especialmente a mis padres, Resti y Dalia, y a mi hermana, Alicia, quienes han estado toda la vida ahí dispuestos a ayudarme y sin los cuales nada de esto habría sido posible. Gracias por transmitirme valores tan importantes como la constancia o el esfuerzo, y por hacer que merezca la pena luchar cada día por ser un poco mejor que el anterior.

Por último, agradecer a Gema, por convertirse en uno de los pilares más importantes de mi vida y en mi apoyo en mi día a día.

A todos vosotros, gracias de corazón.

RESUMEN

La variabilidad temporal en distribuciones de datos biomédicos es uno de los principales problemas hacia una inteligencia artificial basada en aprendizaje automático generalizable. La investigación y toma de decisiones en entornos biomédicos están centradas en el uso de los datos almacenados en los sistemas de información. Como consecuencia, una baja calidad en estos datos puede afectar negativamente a los procesos y resultados asociados a ellos, pudiendo dar lugar a decisiones subóptimas. En este trabajo, se lleva a cabo una caracterización exhaustiva de la variabilidad temporal de datos, uno de los problemas de calidad de datos más significativos para el aprendizaje automático, centrándose en el relevante conjunto de datos MIMIC-IV, con el fin de caracterizar diferentes tipos de variabilidad o "dataset shifts" como cambios de concepto, en las distribuciones marginales de las variables a predecir o en las covariables, tanto de forma gradual como abrupta. Para ello, se utilizará el *Information Geometric Temporal plot*, para obtener una visualización de la evolución temporal de las distribuciones de datos, basada en la proyección de la variedad estadística de las relaciones entre lotes temporales. Asimismo, los resultados serán contrastados con cambios temporales en el rendimiento de diferentes modelos predictivos basados en árboles de decisión, entrenados con subconjuntos de datos procedentes de distintos espacios temporales, con el fin de evaluar el impacto de la variabilidad temporal. Los resultados de este trabajo demuestran la relevancia de realizar un análisis de la variabilidad temporal previo al desarrollo de modelos predictivos en medicina basados en inteligencia artificial.

Palabras Clave: calidad de datos, variabilidad temporal, inteligencia artificial, aprendizaje automático, enfoque generativo, enfoque discriminativo.

RESUM

La variabilitat temporal en distribucions de dades biomèdiques és un dels principals problemes vers a una intel·ligència artificial basada en aprenentatge automàtic generalitzable. La investigació i presa de decisions en entorns biomèdics estan centrades en l'ús de les dades emmagatzemades als sistemes d'informació. Com a conseqüència, una baixa qualitat en aquestes dades pot afectar negativament als processos i resultats associats a ells, esdevinguent en decissions subòptimes. En aquest treball es du a terme una caracterització exhaustiva de la variabilitat temporal de dades, un dels problemes de qualitat de dades més significatius per a l'aprenentatge automàtic, centrant-se en el rellevant conjunt de dades MIMIC-IV, amb la finalitat de caracteritzar distints tipus de variabilitat o "dataset shifts" com a canvis de concepte, en les distribucions marginals de les variables a predir o en les covariables, tant de forma gradual com abrupta. A fi d'abastar aquest objectiu, s'utilitzarà l'Information Geometric Temporal plot, per obtindre una visualització de l'evolució temporal de les distribucions de dades, basada en la projecció de la varietat estadística de les relacions entre lots temporals. Així mateix, els resultats seran contrastats amb canvis temporals en el rendiment de distints models predictius basats en arbres de decisió, entrenats ambs subconjunts de dades procedents de distints espais temporals, amb la finalitat d'evaluar l'impacte de la variabilitat temporal. Els resultats d'aquest treball demostren la rellevància de realizar una anàlisi de la variabilitat temporal previa al desenvolupament de models predictius en medicina basats en intel·ligència artificial.

Paraules Clau: qualitat de dades, variabilitat temporal, intel·ligència artificial, aprenentatge automàtic, enfocament generatiu, enfocament discriminatiu.

ABSTRACT

The temporal variability in distributions of biomedical data is one of the main challenges towards Artificial Intelligence based on generalizable machine learning. Research and decision-taking in biomedical environments are focused on the use of data stored in information systems. As a consequence, low quality in data can have a negative impact on the processes and results associated with them, leading to suboptimal decisions. In this project, an exhaustive characterization of temporal data variability is carried out, which is one of the most important data quality issues for machine learning, focusing on the relevant dataset MIMIC-IV, with the final purpose of characterizing different types of variability or "dataset shifts," such as concept drifts, prior probability shifts or covariate shifts, both in a gradual and abrupt manner. To accomplish it, the Information Geometric Temporal plot will be used to get a visualization of the temporal evolution of data distributions, based on the projection of the statistical variety of relationships between temporal batches. Additionally, the results will be compared with temporal changes in the performance of different predictive models based on decision trees, trained with subsets of data from different temporal spaces with the aim of evaluating the impact of the temporal variability. The results of this project show the importance of doing an analysis of the temporal variability prior to the development of predictive models in medicine based on Artificial Intelligence.

Keywords: data quality, temporal variability, artificial intelligence, machine learning, generative approach, discriminative approach.

ÍNDICE GENERAL

II Presupuesto

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ACRÓNIMOS

BDSLab Biomedical Data Science Lab

DTH Mapa de Calor Temporal

EWC Consolidación Elástica de Pesos

FN Falsos Negativos

FP Falsos Positivos

FPR Tasa de Falsos Positivos

GPD Desviación Global Probabilística

GWR Crecimiento Cuando es Requerido

HCE Historia Clínica Electrónica

IA Inteligencia Artificial

ICD Clasificación Internacional de Enfermedades

IGT Información Geométrica Temporal

ITACA Institute of Information and Comunication Technologies

MCA Análisis de Correspondencia Múltiple

PCA Análisis de Componentes Principales

PDF Función de Distribución de Probabilidad

PR Precisión Sensibilidad

PSI Indice de Stabilidad de la Población

SPO Fuente de Exclusión Probabilística

TN Verdaderos Negativos

TP Verdaderos Positivos

TPR Tasa de Verdaderos Positivos

UPV Universitat Politècnica de València

Parte I

Memoria

Capítulo 1

Introducción

En este capítulo se expone la motivación que ha inspirado la realización del trabajo y los objetivos que se pretenden cumplir, así como las hipótesis planteadas y la estructura de las secciones que lo componen.

1.1 Motivación

El *Biomedical Data Science Lab* (BDSLab) es un grupo de investigación multidisciplinar perteneciente al *Institute of Information and Comunication Technologies* (ITACA) de la *Universitat Politècnica de Valènci*a (UPV). Este grupo se centra en el desarrollo y la investigación de tecnologías basadas en la ciencia de datos y la inteligencia artificial (IA), mediante el uso de herramientas de aprendizaje automático (o *machine learning* en inglés) procesado de datos biomédicos o modelos predictivos, con el objetivo de abordar diversos desafíos en el ámbito sanitario. Actualmente, las líneas de investigación se centran en diferentes ramas, como por ejemplo el análisis de la calidad de datos biomédicos, la imagen médica y los sistemas de apoyo a la decisión clínica.

El presente trabajo de fin de grado se enmarca en el creciente interés del autor por el campo de la inteligencia artificial y los requisitos necesarios para que estos sistemas puedan estar basados en un aprendizaje automático generalizable en el sector biomédico. La oportunidad de unirse al BDSLab ha permitido al autor explorar el potencial de la ciencia de datos y la inteligencia artificial en el ámbito médico, incorporándose al sector del análisis de la calidad de datos biomédicos y desarrollo de IA confiable.

La motivación de este proyecto reside en la necesidad no solo de corroborar los conocimientos ya existentes sobre el tema, sino también proporcionar información valiosa y perspectivas novedosas que ayuden a mejorar la eficacia y confiabilidad de los modelos de inteligencia artificial por medio de comprender la variabilidad temporal de repositorios de datos biomédicos, uno de los principales problemas de calidad de datos para el desarrollo de la IA, así como su impacto en modelos predictivos.

La variabilidad temporal en los datos biomédicos es un fenómeno de gran relevancia, debido a que los datos pueden cambiar con el tiempo como consecuencia a diferentes factores, como la evolución de las tecnologías, los avances en los protocolos de recolección de datos o las variaciones en los perfiles de los pacientes. Comprender y caracterizar esta variabilidad es fundamental para asegurar la confiabilidad y la generalización de los modelos de inteligencia artificial.

Además, se espera que los hallazgos y conclusiones obtenidos en este proyecto proporcionen una base sólida para futuras investigaciones y puedan contribuir al diseño de estrategias más robustas y confiables a la hora de desarrollar modelos de inteligencia artificial en la medicina.

1.2 Objetivos

Los objetivos del presente trabajo los podemos dividir en objetivo principal y objetivos secundarios:

- Objetivo principal (**OP**): Evaluar la asociación entre los resultados de análisis de la variabilidad temporal basados únicamente en las distribuciones de los datos, y los resultados de modelos predictivos basados en *machine learning* desarrollados a partir de los datos. El enfoque busca el potencial beneficio de realizar una evaluación previa de los posibles problemas de modelado y generalización a la hora de desarrollar modelos de IA.
- Objetivos secundarios:
    - **OS1**. Demostrar el objetivo principal utilizando el repositorio de datos biomédicos MIMIC-IV, un *benchmark* ampliamente aceptado en la literatura biomédica, que incluye datos de más de 40000 pacientes de la unidad de cuidados intensivos del *Beth Israel Deaconess Medical Center*, gracias a caracterizar exahustivamente su variabilidad temporal usando representación de la Información Geométrica Temporal, que genera proyecciones de los datos divididos en lotes temporales en función de su distribución de probabilidad.
    - **OS2.** Evaluar el objetivo principal con dos modelos predictivos diferentes, incluyendo *Random Forest* y *Gradient Boosting* y llevando a cabo la clasificación de los datos del MIMIC IV y obteniendo tablas de resultados para las diferentes métricas.
    - **OS3.** Desarrollar una técnica que permita comparar la trayectoria temporal de variabilidad generada por la representación de la IGT del análisis de la variabilidad temporal con las tablas de rendimiento de modelos predictivos, y cuantificar estadísticamente su asociación.

1.3 Contribuciones

Este trabajo de final de grado presenta hallazgos importantes a la hora de avanzar hacia una inteligencia artificial basada en aprendizaje automático generalizable, mediante el uso de técnicas y métodos novedosos, los cuales serán descritos a continuación:

C1. En primer lugar, se ha realizado un estudio con el objetivo de evaluar la variabilidad temporal en los datos del repositorio biomédico MIMIC-IV debido a diferentes causas, como cambios en el protocolo de registro de los datos o diferencias en el registro. Para llevar a cabo esta evaluación, se ha desarrollado un algoritmo que permite realizar un análisis temporal tanto univariante como multivariante por medio de la herramienta de la representación de la Información Geométrica Temporal (IGT)(Sáez Silvestre, C., 2016). Este método ha permitido realizar un análisis exhaustivo realizando un informe sobre cómo se distribuyen los datos a lo largo del tiempo, así como la influencia que tiene cada variable y cada clase del repositorio sobre la variabilidad final gracias al estudio de los mapas de calor temporales y la representación de la IGT, pudiendo descubrir de esta forma tendencias temporales en los datos, periodos de tiempo relacionados conceptualmente, cambios abruptos o anomalías temporales.

C2. En segundo lugar, se ha llevado a cabo la implementación de dos modelos predictivos que permiten llevar a cabo una clasificación de los datos bajo técnicas de aprendizaje supervisado. Tras una revisión del estado del arte de los modelos de clasificación de datos biomédicos en la actualidad, se ha tomado la decisión de diseñar dos modelos basados en árboles de decisión, tanto *Random Forest* como *Gradient Boosting*. Con el objetivo de poder relacionar posteriormente el resultado de los modelos con los resultados obtenidos previamente en análisis de la variabilidad temporal, se ha desarrollado un método para entrenar y evaluar los modelos con datos provenientes de diferentes lotes temporales, obteniendo para cada uno de los resultados las siguientes métricas: área bajo la curva ROC, sensibilidad, precisión, exactitud y *F1-Score*.

C3. Se ha cumplido con el objetivo principal del trabajo (**OP**) demostrando la influencia y la relación de los *data shifts* encontrados en los repositorios sobre los modelos de clasificación. Para ello, se ha creado un algoritmo que permite obtener la IGT de las métricas descritas en la **C2** por medio de la aplicación de un escalado multidimensional, para posteriormente poder compararlo con la IGT del análisis temporal por medio de un proceso de agrupación mediante un *cluster* jerárquico, y su validación correspondiente por medio de un test chi-cuadrado entre las dos agrupaciones.

Con el fin de documentar y difundir los resultados obtenidos en este estudio, se ha comenzado a redactar un artículo científico para su posterior publicación en una revista especializada (como por ejemplo *npj Digital Medicine* o *The Lancet Digital Health*), con el objetivo de compartir los hallazgos y contribuciones significativas de este trabajo en el campo de la inteligencia artificial aplicada a la salud.

1.4 Estructura

En la Figura 1.1 se muestra de manera esquemática los capítulos de los que se compone este trabajo, así como un breve resumen de ellos expuesto a continuación:

- **Capítulo 1.** Se expone la motivación para la realización de este trabajo, así como los objetivos que se pretenden cumplir, las contribuciones y la estructura que presenta.
- **Capítulo 2.** Estudio profundo de la problemática de la variabilidad temporal en los repositorios biomédicos, así como del estado actual de los métodos de análisis de dataset shifts y del aprendizaje continuo, justificando qué análisis de la variabilidad temporal beneficiaría al desarrollo de modelos de IA.
- **Capítulo 3.** Descripción de los materiales utilizados durante este trabajo: los repositorios biomédicos utilizados y las herramientas empleadas.
- **Capítulo 4.** Descripción de la metodología empleada para el desarrollo del presente trabajo: el preprocesado de los datos, los análisis de variabilidad temporal, el desarrollo de los modelos de clasificación y la validación del resultado final.
- **Capítulo 5.** Recopilación de los resultados obtenidos aplicando la metodología expuesta anteriormente.
- **Capítulo 6.** Discusión de los resultados, exposición de las limitaciones encontradas y la relevancia del proyecto, así como posibles investigaciones futuras.
- **Capítulo 7.** Conclusiones extraídas de la realización del trabajo.

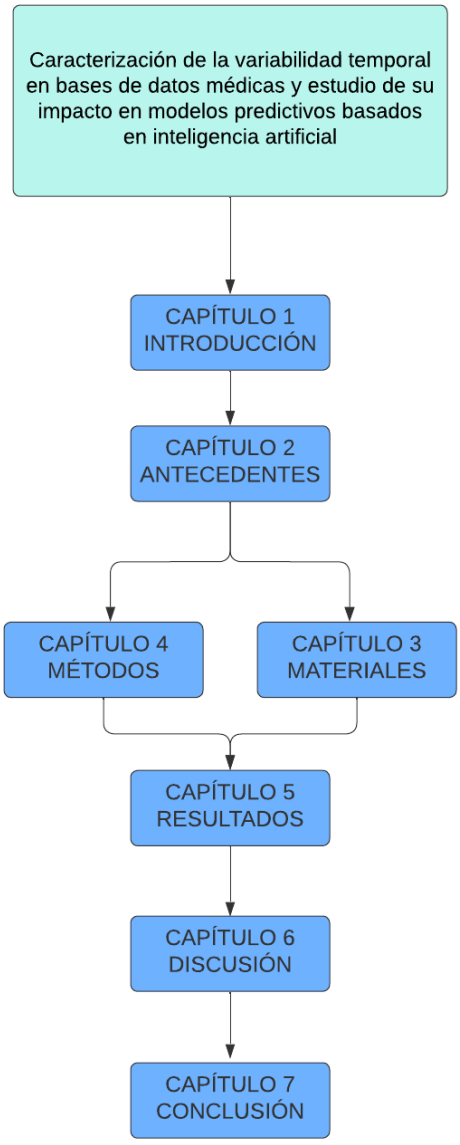


Figura 1.1: Esquema de la estructura en capítulos en la que se basa el presente trabajo

Capítulo 2

Antecedentes

En este capítulo se presentan los conceptos relacionados con el problema que se pretende abordar. En primer lugar, se describirá la problemática asociada a la calidad de datos biomédicos. En segundo lugar, se realizará una revisión teórica del estado del arte de los temas a tratar, con el objetivo de comprender de mejor forma los métodos desarrollados más adelante.

2.1 Problema de la variabilidad temporal en IA médica

En el ámbito biomédico, tanto los sistemas de toma de decisiones como los resultados de las investigaciones se fundamentan en la información disponible en el momento de su realización. Esta información, recopilada por humanos y dispositivos, no está exenta de imperfecciones y errores inherentes al proceso. Sin embargo, en un contexto en el que la vida de las personas está en juego, resulta crucial garantizar que los modelos y técnicas empleados no conduzcan a decisiones erróneas o resultados subóptimos debido a la utilización de datos de baja calidad. Por tanto, es imprescindible abordar la relevancia de utilizar fuentes de información confiables y de alta calidad en la generación y aplicación de los modelos de inteligencia artificial en el ámbito biomédico.

En los cuidados primarios de pacientes, una baja calidad de los datos puede dar lugar tanto a errores directos, como por ejemplo aplicar terapias inapropiadas o anticuadas, errores quirúrgicos, el suministro de medicación incorrecta o errores en la dosis; como a errores indirectos, como la falta de precauciones, el no uso de las pruebas o test indicados, demoras en los tiempos de diagnóstico evitables o no actuar en base a los resultados de las pruebas (Institute of Medicine (US) Committee on Data Standards for Patient Safety, 2004).

Es por esto por lo que detectar y caracterizar esta deficiencia en la calidad de los datos debe ser una prioridad antes de tomar cualquier decisión clínica o realizar cualquier estudio o tarea en el ámbito médico relacionado con la extracción de información a partir de datos existentes.

Sin embargo, este es un reto difícil de abordar. Esto es debido a que, por una parte, la falta de calidad de los datos puede estar debida a numerosas causas; por otro lado, los repositorios biomédicos presentan un alto nivel de complejidad, debido a que comprenden una gran cantidad de datos, generados por múltiples fuentes, y con diversos tipos de datos, cuyas distribuciones muchas veces muestran la presencia de subpoblaciones y cuya interoperabilidad es una tarea complicada de conseguir (Sáez Silvestre, C., 2016).

Durante las últimas décadas, se han introducido un gran número de sistemas informáticos para ayudar a la práctica cínica. Esta digitalización, ha permitido a los investigadores encargados del desarrollo de modelos de *machine learning* acceder a una gran cantidad de datos etiquetados sobre las Historias Clínicas Electrónicas (HCE) de los pacientes. Sin embargo, esta digitalización no ha sido completamente beneficiosa en todos los aspectos, y ha provocado que muchos de los problemas en la calidad de la información clínica estén relacionados con dos causas principales (Cruz-Correia RJ et al., 2010): en

primer lugar, las HCE están diseñadas para proporcionar una atención a los pacientes (uso primario), sin tener en cuenta que el posterior uso de esos datos con fines de investigación pueden requerir diferentes grados de calidad y, en segundo lugar, las HCE no han sido diseñadas con la previsión de abordar de manera proactiva los desafíos relacionados con la calidad de los datos.

De hecho, incluso si los sistemas de recogida de datos de las HCE y los diseños de sus interfaces de usuario fuesen perfectos, seguiría habiendo un alto nivel de incertidumbre en la calidad de los datos obtenidos (Cruz-Correia RJ et al., 2010). Esto se debe a que los datos son recogidos por diferentes usuarios, y son introducidos en distintos sistemas de registro casi simultáneamente o durante el mismo acontecimiento del paciente, provocando problemas de vinculación de registros y acontecimientos, o que se puedan producir cambios en el protocolo de registro de los mismos a lo largo del tiempo.

Es por ello por lo que parece razonable que, para realizar una evaluación adecuada de los problemas en la calidad de los datos, haya que centrarse en diferentes aspectos por separado. A estos diferentes aspectos se los conoce como Dimensiones de Calidad de Datos (definidas en las tablas 2.1 y 2.2) (Wang, R. Y., & Strong, D. M., 1996) , teniendo cada uno sus propias herramientas de detección para poder evaluar así sus pros y contras.

Como ya se ha mencionado anteriormente, este trabajo se ha centrado en el estudio de uno de los problemas de la calidad más importantes a la hora de reutilizar datos procedentes de las HCE con fines de investigación por medio de herramientas existentes, y este es el problema de la variabilidad en la distribución de los datos a lo largo del tiempo.

La variabilidad temporal, puede ser clasificada dentro de las categorías de las dimensiones de calidad de datos como intrínseca o contextual (Wang, R. Y. y & Strong, D. M., 1996) .

Tabla 2.1: Definiciones de las dimensiones de calidad de datos (Wang, R. Y. y & Strong, D. M., 1996)

Categoría	Dimensión	Atributos
Intrínseca:	Credibilidad	Los datos son creíbles
Los datos tienen calidad por sí mismos	Precisión	Los datos son certificados como libres de errores, precisos, correctos, fiables y con la integridad intacta
	Objetividad	Los datos no tienen sesgos y son objetivos
	Reputación	La fuente de datos tiene una buena reputación
Contextual:	Valor añadido	Los datos brindan una ventaja competitiva y agregan valor a las operaciones
La calidad de los datos de ser considerada teniendo en cuenta el contexto de la tarea	Relevancia	Los datos son aplicables, relevantes, interesantes y utilizables
	Puntualidad	La antigüedad de los datos
	Integridad	El alcance y la profundidad de la información contenida en los datos, y la cantidad apropiada de datos
	Cantidad adecuada de datos	La cantidad de datos
Representacional:	Interpretabilidad	Los datos son interpretables

Los datos están presentados de una manera clara e inteligible	Facilidad de entendimiento	Fácilmente entendibles, claros y legibles
	Coherencia de representación	Los datos se presentan continuamente en el mismo formato, de manera consistente y compatible con los datos anteriores
	Representación concisa	Los datos se presentan de manera concisa, organizada y estéticamente agradable
Accesibilidad:	Accesibilidad	Los datos son accesibles, recuperables, están disponibles y actualizados
Los datos son accesibles y seguros		
	Seguridad de acceso	Los datos no pueden ser accedidos por competidores, pueden restringirse los accesos y la seguridad está garantizada

Tabla 2.2: Definiciones de las dimensiones de calidad de datos (Liaw, S. T. et al. 2013)

Dimensión	Descripción
Integridad	El grado en que la información no falta y tiene suficiente amplitud y profundidad para la tarea en cuestión. La capacidad de un sistema de información para representar cada estado significativo del sistema del mundo real representado. Grado en que la información es suficiente para representar cada estado posible de la tarea. Se registran todos los valores para una variable. Disponibilidad de un número mínimo definido de registros/pacientes.
Corrección	La dimensión libre de errores. Credibilidad de la fuente y nivel de experiencia del usuario. Los valores, formato y tipos de datos son válidos y apropiados; por ejemplo, la altura está en metros y dentro del rango para la edad. El valor registrado está en conformidad con el valor real. La precisión de los datos incluye exactitud y completitud.
Precisión	El valor registrado está en conformidad con el valor real.
Consistencia	La representación de los valores de datos es la misma en todos los casos. Incluye valores y representación física de los datos. El grado en que la información es fácil de manipular y aplicar a diferentes tareas. La equivalencia y el proceso para lograr la equivalencia de la información almacenada o utilizada en aplicaciones y sistemas. El grado de uso de un tipo y formato de datos uniforme (por ejemplo, entero, cadena, fecha) con una etiqueta de datos uniforme (consistencia interna) y códigos/términos que se pueden asignar a una terminología de referencia (consistencia externa).
Puntualidad	Los datos no están desactualizados; la disponibilidad de los resultados es puntual. El grado en que la información está actualizada para la tarea. El retraso entre un cambio en el estado del mundo real y la modificación resultante del estado del sistema de información.

El tiempo es un factor que ha sido estudiado comúnmente como parte de las dimensiones de calidad de datos en muchas investigaciones, sobre todo dentro de las dimensiones de puntualidad, actualidad o volatilidad. Estas dimensiones están relacionadas con el grado de actualidad de los datos y con el grado de cambio en comparación a los valores del mundo real. Sin embargo, este trabajo se ha centrado en el estudio del tiempo desde el punto de vista de la concordancia de la información a lo largo del tiempo, no en la dimensión de la actualidad como se ha mencionado en los estudios anteriores. Muchos estudios del factor temporal llevados a cabo en la actualidad asumen una distribución estacionaria de la distribución de los datos en el dataset, pero estos enfoques pasan por alto los cambios en las características causados por cambios subyacentes en los conceptos a lo largo del tiempo

cuando se recogen datos en largos periodos de tiempo y por diferentes usuarios o dispositivos (Sáez, C. et al., 2012), demostrando así la importancia del presente trabajo.

Para demostrar la influencia de la variabilidad temporal, se ha llevado a cabo un análisis del conocido benchmark MIMIC-IV por medio de la representación de la IGT y de los mapas de calor de la información temporal gracias al uso del paquete de R "*EHRtemporalVariability"* (Sáez, C. et al., 2020).

2.2 Estado del arte en métodos de análisis de *dataset shifts*

Los repositorios biomédicos a menudo están compuestos por datos provenientes de diferentes fuentes o adquiridos durante largos periodos de tiempo, provocando variaciones en las distribuciones de los datos de entrenamiento y de test, lo que provoca que a la hora de desarrollar modelos de aprendizaje automático con estos datos los resultados sean subóptimos. Estas variaciones en las distribuciones de los datos son los denominados *dataset shifts* (Moreno-Torres, J. G. et al., 2012), (Quiñonero-Candela, J. et al., 2008), y son uno de los principales impedimentos a la hora de desarrollar modelos de inteligencia artificial basados en aprendizaje automático generalizable. Durante los últimos años, el objetivo de muchos investigadores ha sido desarrollar un método genérico para la obtención de métricas que permitan evaluar la variabilidad de los datos médicos.

En esta sección, se hablará de los tipos de *dataset shifts* más comunes, así como de sus posibles causas y las técnicas de detección existentes. Los *dataset shifts* se pueden definir como cambios en la distribución de los datos de entrenamiento y test (Quiñonero-Candela, J. et al., 2008). Estos cambios pueden estar causados por diversos factores, como los definidos en el apartado 2.2.4. Asimismo, hay que tener en cuenta de que existe la posibilidad de que los *dataset shifts* aparezcan de forma repentina, en forma de cambios abruptos en los datos o, por el contrario, que aparezcan de una forma más gradual. Además, existe la posibilidad incluso de que se manifiesten siguiendo patrones recurrentes debido a cambios con un componente estacionario (Gama, J. et al., 2014).

Normalmente en el contexto de los problemas de clasificación donde se tienen unas variables de entrada x, y otras variables de salida y, los *dataset shifts* ocurren cuando la distribución de probabilidad conjunta de los datos entrenamiento y test difiere:

$$P_{entrenamiento}(x, y) \neq P_{test}(x, y) \quad (2.1)$$

Esta variación en la distribución de probabilidad conjunta puede estar relacionada con múltiples fuentes (Moreno-Torres, J. G. et al., 2012), (Gama, J. et al., 2014), pero en este trabajo se han considerado tres fuentes principales de cambio: *covariate shift, prior probability shift* y *concept shift* (Moreno-Torres, J. G. et al., 2012).

Cuando se lleva a cabo el análisis de *dataset shifts,* las relaciones entre las covariables y las etiquetas de cada clase son muy relevantes. Moreno-Torres, J. G. et al. (2012) describen la relación causal entre las covariables y las etiquetas de las clases como una propiedad de los datos para clasificar en dos los problemas que pueden suceder, dando lugar a los *dataset shift*. Sin embargo, en el presente trabajo se

ha considerado una definición propia para caracterizar los *dataset shifts* debido al carácter bidireccional de los problemas en el ámbito médico.

A continuación, se presenta una breve descripción de cada tipo de *dataset shift*.

2.2.1 *Covariate shift*

Matemáticamente el *covariate shift* se define como:

$$P_{train}(x) \neq P_{test}(x) \text{ (2.2)}$$

Se refiere a cambios en la distribución en las variables de entrada, como por ejemplo cambios en la distribución de la población a lo largo del tiempo o variaciones en la distribución de los datos de entrenamiento y de test. Para llevar a cabo la identificación de los *covariate shifts* en el presente trabajo, se ha establecido como criterio la variación de la distribución de las variables de entrada p(x) (Kelly, M. G. et al. 1999). En la figura 2.1 se pudede observar un ejemplo de *covariate shift* obtenido del trabajo de Moreno-Torres, J. G. et al., (2012).

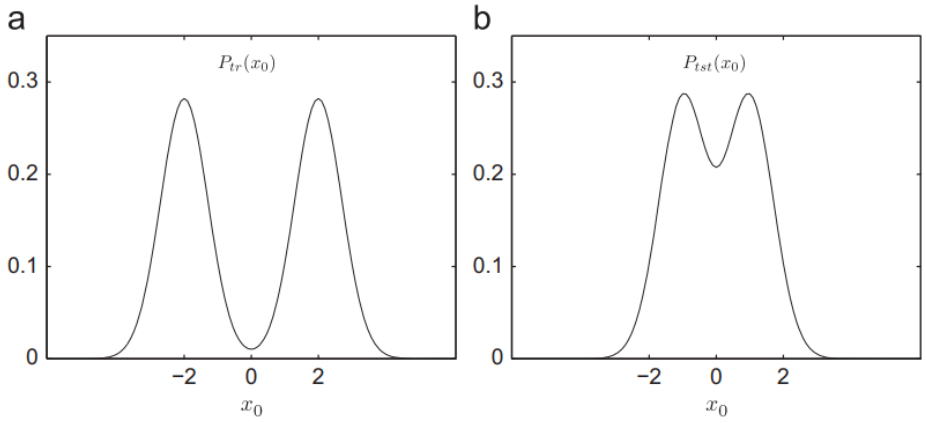


Figura 2.1: Ejemplo de *covariate shift*. (a) Datos de entrenamiento y (b) test de Moreno-Torres, J. G. et al., (2012).

2.2.2 *Prior probability shift*

Matemáticamente el *prior probability shift* se define como:

$$P_{train}(y) \neq P_{test}(y) \text{ (2.3)}$$

Hace referencia a cambios en la distribución de cada clase. En este caso, el criterio de identificación de *prior probability* shifts utilizado ha sido la variación en la distribución de cada clase p(y) en los datos de training y test (Alaiz-Rodríguez, R., y Japkowicz, N. 2008). Esto ocurre por ejemplo cuando cambia la prevalencia de una enfermedad en una población objetivo, pero se manifiesta de la misma manera. En la figura 2.2 se pudede observar un ejemplo de *prior probability shift* obtenido del trabajo de Moreno-Torres, J. G. et al., (2012).

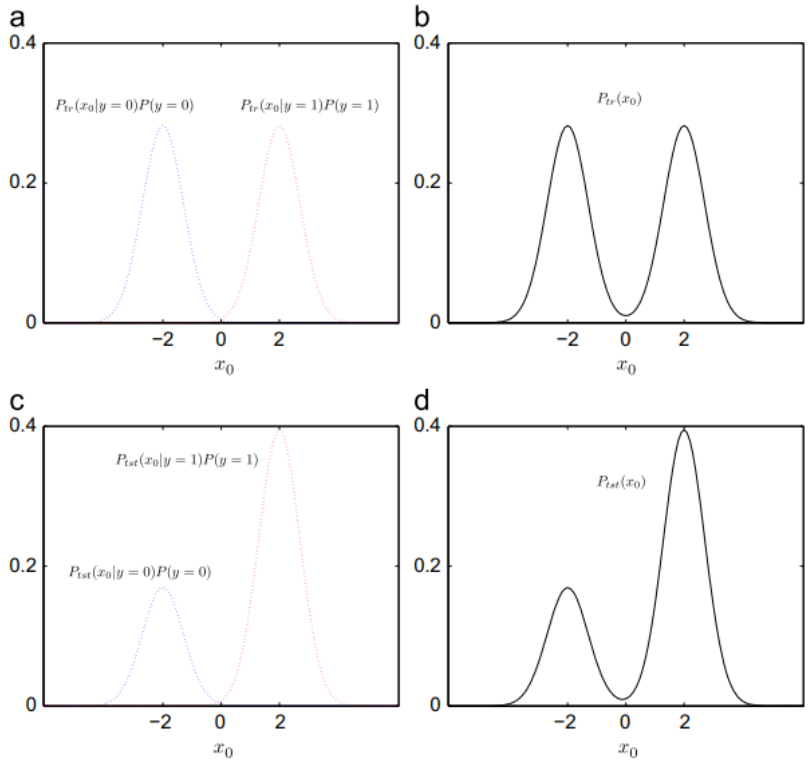


Figura 2.2: Ejemplo de *Prior probability shift*. (a) Datos de entrenamiento, (b) densidad de los datos de entrenamiento, (c) datos de test y (d) densidad de los datos de test de Moreno-Torres, J. G. et al., (2012).

2.2.3 *Concept shift*

Matemáticamente el *concept shift* se define como:

$$P_{train}(y|x) \neq P_{test}(y|x) \quad (2.4)$$

Se produce cuando hay un cambio en el contexto que puede provocar cambios en los conceptos objetivo (Widmer, G., y Kubat, M. 1996), cambios en la definición de las clases, o cambios en la distribución condicionada p(y|x) entre las fases de entrenamiento y test. En la figura 2.3 se pudede observar un ejemplo de *concep shift* obtenido del trabajo de Moreno-Torres, J. G. et al., (2012).

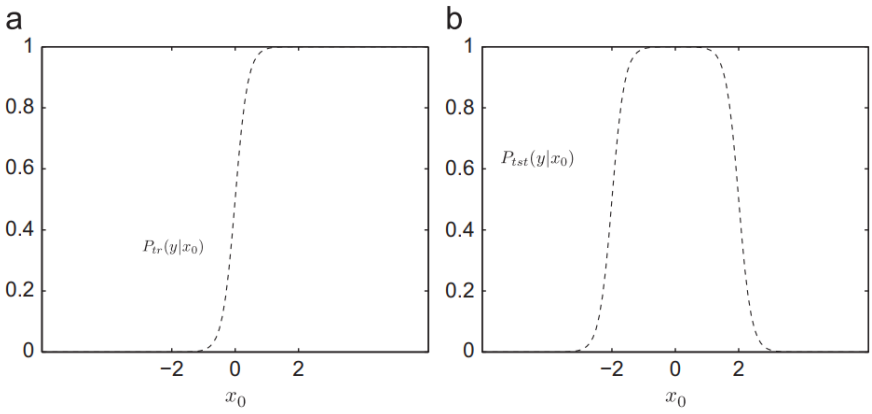


Figura 2.3: Ejemplo de *Concept shift*. (a) Datos de entrenamiento y (b) datos de test de Moreno-Torres, J. G. et al., 2012).

2.2.4 Causas de los *dataset shifts*

Existen numerosas causas que pueden dar lugar a la variabilidad de los datos, pero las dos que se han considerado más importantes son (Moreno-Torres, J. G., et al. 2012) : (1) el sesgo en la selección de las muestras, (2) entornos no estacionarios.

El sesgo de selección de la muestra no es un defecto de ningún algoritmo ni del manejo de los datos. Es puramente un fallo sistemático en el proceso de recogida o etiquetado de datos que provoca una selección no uniforme de ejemplos de entrenamiento de una población, lo que hace que se formen sesgos durante el entrenamiento. Este caso puede dar lugar tanto a la presencia de *covariate shift* como de *concept shift*, debido a que se está influenciando la distribución de las muestras. Los efectos de la selección sesgada de muestras son especialmente relevantes en casos en los que tenemos un gran desbalanceo en la prevalencia de las clases, debido a que la clase minoritaria es particularmente sensible a los errores de clasificación singulares, debido al bajo número de muestras.

La segunda causa mencionada se debe a que, en el mundo real, los datos no son estacionarios en el espacio ni en el tiempo, como puede ser el caso de que se den cambios en las poblaciones, protocolos, prácticas médicas, o la aparición de eventos inesperados, como pandemias o sucesos por el estilo.

2.2.5 Métodos de identificación de *dataset shifts*

Existen numerosos métodos que pueden ser utilizados para determinar la presencia de *dataset shifts* y su gravedad, como podemos ver de forma esquemática en la figura 2.4

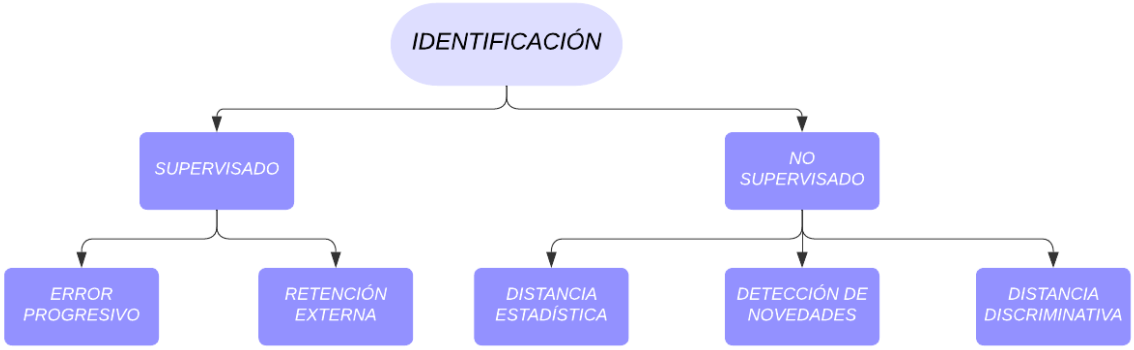


Figura 2.4: Métodos de identificación de *dataset shifts*

A continuación, se explicarán las técnicas de identificación no supervisadas, siendo estas las más utilizadas.

Distancia estadística:

Es un método útil para detectar variaciones en las predicciones de los modelos de *machine learning* a lo largo del tiempo. Para ello, se obtiene el histograma y la distribución de densidad de los datos a lo largo del tiempo y se compara para observar la presencia de cambios. Para realizar la comparación y determinar la presencia de variabilidad, se pueden utilizar métricas como el *Population Stability Index* (PSI), la estadística *Kolmogorov-Smirnov*, o la divergencia *Kullback-Lebler* o la intersección entre histogramas (Matthew,S., 2019).

Detección de novedades:

Es un método que se adapta mejor a espacios complejos. La idea es crear un modelo para representar la distribución de las fuentes, de tal forma que, dado un nuevo punto en los datos, se intenta probar cual es la probabilidad de que ese punto proceda de las distribuciones fuente. Es un método eficaz para datos con interacciones complejas. La principal desventaja del método es que no permite conocer qué ha cambiado en los datos, sino simplemente que ha habido un cambio (Matthew,S., 2019) .

Distancia discriminativa:

Es un método más inusual pero muy efectivo en entornos de alta dimensionalidad o con datos muy dispersos. El método consiste en entrenar un clasificador para detectar si un dato pertenece o no al dominio de origen o al de destino. Para ello, se puede utilizar el error de entrenamiento como indicador de la distancia entre ambas distribuciones, de tal forma que cuanto mayor sea el error, más cerca estarán, o lo que es lo mismo, el clasificador no puede discriminar entre el domino de origen y el de destino (Matthew,S., 2019).

2.3 Estado del arte en aprendizaje continuo

Como se ha estado viendo hasta el momento, a la hora de realizar modelos de inteligencia artificial

basados en el uso de datos reales, corremos el riesgo de que los datos utilizados para el entrenamiento no tengan las mismas características que los datos usados para test, haciendo que los modelos desarrollados solo sirvan para datos e instantes temporales concretos, de tal forma que siempre que queramos emplear nuevos datos o desarrollar tareas nuevas deberíamos de desarrollar nuevos modelos. El aprendizaje continuo es un paradigma del *machine learning* que intenta poner solución a este problema, entrenando modelos en el tiempo de forma que puedan adquirir conocimiento para desarrollar tareas nuevas, así como mantener conocimiento para solucionar tareas antiguas (Parisi, G. I. et al., 2019), (Chen, Z., y Liu, B. (2018) . Otro paradigma asociado con el aprendizaje continuo es el aprendizaje por transferencia, que hace referencia al hecho de aprovechar el conocimiento de un modelo preentrenado para una tarea original para una nueva tarea (Tercan, H. et al., 2022). Sin embargo, al entrenar modelos de *machine learning*, como pueden ser las redes neuronales, por medio de aprendizaje por transferencia, pueden sufrir los denominados olvidos catastróficos, provocando que cuando se entrenan para realizar los nuevos objetivos, su rendimiento para tareas aprendidas previamente cae debido a cambios en los pesos de la red. Las técnicas de aprendizaje continuo tienen en cuenta este hecho e intentan ponerle solución encontrando un equilibrio entre la estabilidad y la plasticidad de los parámetros de la red cuando se entrena para realizar nuevas funciones. Los métodos de aprendizaje continuo se pueden dividir en 3 categorías: estrategias de ensayos basados en memoria, arquitecturas dinámicas y estrategias de regularización (Parisi, G. I. et al., 2019).

Los métodos de ensayo utilizan una memoria de tamaño fijo para almacenar muestras de tareas entrenadas previamente. Estas muestras son luego utilizadas durante el entrenamiento de nuevas tareas para mitigar el olvido catastrófico. Por ejemplo, el trabajo de Rebuffi, S.-A. et al., (2017) guarda una memoria episódica con muestras representativas para cada tarea, de tal forma que cuando se entrenan nuevas tareas, se calcula una pérdida en la destilación adicional para evitar que las predicciones de la red para esas muestras cambien significativamente.

Por otro lado, el uso de arquitecturas dinámicas cambia la arquitectura de la red cuando se van a entrenar nuevas tareas. A menudo, se expande dinámicamente la capacidad de la red con el objetivo de aprender nuevos patrones sin causar conflictos. El trabajo de Parisi, G. I. et al., (2019) usa un enfoque de Crecimiento Cuando es Requerido (GWR en inglés) para entrenar redes neuronales autoorganizativas recurrentes que son extendidas jerárquicamente para nuevas tareas.

Por último, las estrategias de regularización reducen el olvido catastrófico restringiendo la actualización de los parámetros de la red mientras se entrenan nuevas tareas. En la Consolidación Elástica de Pesos (EWC en inlgés) de Kirkpatrick, J. et al. (2017), esto se realiza penalizando cambios en los parámetros que son importantes para la predicción de tareas previas. La importancia de los parámetros se estima por medio de la densidad de probabilidad usando la matriz de información de Fisher. A diferencia de este método, los enfoques de sinapsis con memoria propuestos por Aljundi, R. et al. (2018) y Aljundi, R. et al. (2019) representan la importancia de los parámetros de la red por medio de la sensibilidad de la salida de la red ante los cambios en los parámetros. Al incorporar cambios en parámetros importantes en la función de pérdidas, los pesos de las redes se adaptan a tareas nuevas que todavía no son importantes.

Los métodos mencionados anteriormente han mostrado resultados muy prometedores y la mayoría de ellos podrían ser utilizados con funciones predictivas en escenarios con condiciones cambiantes. Sin

embargo, la mayoría no cumplen los 3 requisitos establecidos por el trabajo de H.Tercan et al., (2022) que deberían de cumplir los sistemas de aprendizaje continuo:

- Aprender sin olvidar: cuando se está entrenando una red neuronal existente para cumplir con nuevas tareas, el conocimiento de las anteriores tareas no debería de ser olvidado y la calidad del modelo no debe empeorar.
- Los modelos no deben tener la necesidad de acceder a datos de tareas previas: debido a la limitación del hardware y el acceso restringido a los datos, el entrenamiento de nuevas tareas no debería de requerir el uso de datos previos.
- Buen rendimiento con datos dispersos: el entrenamiento para nuevas tareas es eficiente y debería implicar el uso de menos datos que para entrenar una red desde cero.

Si nos guiamos por estas directrices, podemos intuir cuáles de los métodos vistos hasta el momento son los de mayor interés. El caso de los métodos de ensayo, son los que mejores resultados dan, pero requieren el acceso a datos antiguos y una gran cantidad de memoria para el entrenamiento. Las arquitecturas dinámicas consiguen obtener un aprendizaje sin olvidar, pero son bastante ineficaces en la mayoría de los casos y no se adaptan bien al número de tareas. Los enfoques de sinapsis con memoria son los que mejor cumplen estas directrices, ya que pueden ser utilizados para regresiones, no como el método EWC, y no requieren de acceso a datos antiguos para entrenar nuevas tareas. La importancia de los pesos de la red para una tarea se calcula una sola vez y se retiene para futuros entrenamientos (H.Tercan et al., 2022).

A pesar de que, a la fecha de realización del presente trabajo, el objetivo no es desarrollar modelos de aprendizaje continuo, el desarrollo de redes neuronales para realizar este mismo estudio sobre estos modelos de *machine learning* sí que se plantean como una idea de futuro, tal y como se describe en el apartado 6.3. Además, la información que aporta la realización de este trabajo sí que permite adelantarse a los momentos en los que sí que serían necesarias este tipo de técnicas por medio de la caracterización de los puntos de corte en los datos, bien abruptos o bien graduales, debidos a cambios en los protocolos, olvidos, variabilidad entre fuentes… proporcionando de esta manera una valiosa información tanto para el aprendizaje continuo como para la medicina en general.

2.4 Beneficios de delineación de dataset shifts para el desarrollo, evaluación y regulación de IA médica

Como ya se ha mencionado previamente en el presente trabajo, los repositorios de datos biomédicos generalmente están compuestos por datos de múltiples fuentes y adquiridos durante largos periodos de tiempo. Por ello, se defiende la relevancia de realizar un análisis de ambas características, con el objetivo de identificar variaciones de las distribuciones de probabilidad de los datos antes de desarrollar cualquier modelo de inteligencia artificial. Esto es importante debido a que las diferencias en las distribuciones de los datos pueden llevar a hipótesis erróneas o perjudicar los resultados posteriores obtenidos con nuevos datos, como se demuestra en Sáez, C. et al., (2021), donde se utilizan repositorios con información correspondiente al COVID-19 para evidenciar cómo la variabilidad multifuente de los datos afecta a los resultados de los modelos de *machine learning*. Sin embargo,

detectar estas variaciones puede ser una tarea complicada debido precisamente a la heterogeneidad de los datos en el sector biomédico: (1) variables de diferentes tipos (categóricas, ordinales o no numéricas; discretas o continuas), (2) datos provenientes de distribuciones uni-modales o multi-modales, (3) datos univariantes o multivariantes, (4) actualizaciones en el protocolo de registro a lo largo del tiempo. Proporcionar información precisa sobre estos tipos de variabilidad de los datos puede ayudar a los investigadores en gran medida a la hora de tomar decisiones sobre la definición y el desarrollo de modelos de IA con el fin de obtener los mejores resultados posibles. Además, la creación de métodos que permiten obtener métricas generalizables comparables entre distintos estudios ha dado lugar en los últimos años (Sáez Silvestre, C. 2016) a la posibilidad de conocer la calidad de los datos de repositorios biomédicos en función de su grado de variabilidad, para poder así conocer cuáles son los más óptimos con fines de investigación y para desarrollar modelos de IA. Este hecho en un entorno como es el sector médico supone un gran beneficio, debido a la alta multimodalidad por el uso de diferentes dispositivos para obtener información de cada paciente y su manipulación por diferentes usuarios, así como por el uso de datos altamente codificados, como por ejemplo mediante el uso de códigos ICD, o la anonimización de los mismos por cuestiones legales.

Este estudio aporta beneficios no solo a la hora de desarrollar los modelos de IA en el ámbito médico, sino también para su evaluación y regulación. En el primer caso, realizar un análisis de los *dataset shifts* permite evaluar la robustez y generalización de los modelos de IA ante cambios en los datos. De esta forma, al comprender la forma en la que varían las distribuciones y patrones a lo largo del tiempo, es posible ajustar y adaptar los modelos para que mantengan un rendimiento óptimo a lo largo del tiempo, permitiendo que sean capaces de adaptarse a nuevas situaciones y seguir brindando resultados precisos y confiables en entornos clínicos cambiantes (Dockès, J. et al., 2021).

Por otro lado, la regulación de la IA médica también se ve beneficiada por este tipo de estudios, debido a que la comprensión de las variaciones temporales de los datos médicos permite que se pueda evaluar de forma más precisa la seguridad y eficacia de los modelos utilizados en la toma de decisiones clínicas, facilitando de esta forma la actualización de las políticas y marcos reguladores para adaptarse a los cambios en la práctica médica según la Comisión Europea (2021).

A continuación, se describen algunas de las técnicas existentes para analizar la variabilidad de los datos tanto multi-fuente como temporal para poder alcanzar los beneficios descritos:

2.4.1 Variabilidad multi-fuente

Cuando se utilizan datos provenientes de diferentes fuentes, se pueden llevar a cabo varios procedimientos: en primer lugar, si las distribuciones de los datos son similares, estas pueden tratarse como un conjunto; sin embargo, si son muy diferentes, es mejor considerar cada distribución de forma individual.

Existen algunas herramientas estadísticas que pueden ser utilizadas para la detección de las inhomogeneidades previamente mencionadas, desde técnicas univariantes y de tipo único, como por ejemplo un ANOVA, hasta multivariantes y multitipo, como un PCA (Levman, J., y Takahashi, E. 2015) , pero la mayoría de ellas están restringidas por las suposiciones que conllevan, las cuales no siempre se verifican en datos biomédicos (ej. Distribuciones gaussianas, homocedasticidad, unimodalidad,

etc.)(Sáez Silvestre, C. 2016). Sin embargo, hay métricas, como la desviación global probabilística (GPD) o la fuente de exclusión probabilística (SPO), propuestas en Sáez Silvestre, C. et al., (2017) las cuales no suponen ninguna distribución subyacente y permiten tratar con datos multitipo.

La GPD proporciona un grado limitado de la variabilidad global multifuente, diseñado como un estimador equivalente a la noción de desviación típica normalizad de las PDF. La SPO a su vez, proporciona un grado limitado de la disimilitud de cada fuente a una distribución central latente. Estas métricas están basadas en la proyección de una estructura geométrica simple la cual está construida a partir de las distancias Jensen-Shannon sobre las fuentes de las PDFs (Sáez, C. et al., 2013) .

2.4.2 Variabilidad temporal

El descubrimiento de conocimientos a partir de datos biomédicos se puede aplicar sobre el análisis de flujos de datos online, o sobre conjuntos de datos retrospectivos, con fecha y hora offline. En ambos casos, la variabilidad producida en el proceso de generar los datos o en las características de la calidad a lo largo del tiempo pueden alterar los resultados obtenidos. Para poder obtener una comparación de la calidad de los datos de diferentes repositorios basada en la variabilidad temporal, en el trabajo de Sáez , C. et al., (2015) se proponen dos métodos para evaluar cambios tanto en la monitorización como en la caracterización, tendencias y detección de subgrupos temporales.

El primero es un algoritmo probabilístico de detección de cambios basado en el Control Estadístico de Procesos de la distribución Beta posterior de la distancia Jensen-Shannon, con un mecanismo de olvido sin memoria. Este algoritmo (PDF-SPC) clasifica el grado del cambio actual en tres estados: bajo control, alerta y fuera de control.

El segundo, es un método que permite visualizar y caracterizar los cambios temporales de los datos basados en la proyección de un colector estadístico geométrico de información no paramétrica de ventanas temporales. Esta proyección, facilita la exploración de tendencias temporales usando la representación de la Información Geométrica Temporal, y por medio de técnicas de aprendizaje no supervisado, permite descubrir subgrupos temporales relacionados conceptualmente. Esta IGT se obtiene por medio de calcular la matriz de disimilitud de las PDFs en diferentes paquetes temporales, para posteriormente aplicar una proyección del escalado multidimensional de la misma en un espacio de dimensionalidad reducida de 2 o 3 dimensiones. Esta representación de la IGT proporciona una poderosa herramienta analítica para explorar, caracterizar y entender los cambios de los datos desde una perspectiva probabilística. De esta forma, la IGT es un colector estadístico temporal en el que las PDFs se representan como puntos representados por un índice temporal o con una fecha determinada. El presente trabajo se centrará a partir de este punto en complementar los estudios realizados sobre el MIMIC-IV en Yao, H. et al., (2023) y Sáez, C. et al., (2016), en los cuales se llevan a cabo tareas predictivas y de análisis de variabilidad temporal de estos datos, llevando a cabo la caracterización de *dataset shifts* utilizando la técnica de la IGT y demostrando su relación con los resultados de los modelos de Inteligencia Artificial aplicándola sobre el repositorio de datos biomédicos ampliamente aceptado MIIMC-IV, por medio del uso del paquete de R “EHRtemporalVariability” (Sáez, C. et al. 2020).

Capítulo 3

Materiales

En este capítulo, se exponen los materiales utilizados en el trabajo, describiendo los repositorios biomédicos utilizados, así como las herramientas para llevar a cabo el análisis temporal y los modelos predictivos.

3.1 Repositorios de datos biomédicos

Para la búsqueda de repositorios biomédicos sobre los que poder aplicar los procedimientos teóricos explicados hasta el momento, se han seguido los siguientes criterios :

- Datos tabulares.
- Variaciones temporales naturales: se han buscado repositorios con datos reales del sector biomédico que consistan en datos registrados a lo largo del tiempo durante más de un año. De esta forma se han podido realizar pruebas entrenando modelos con datos del pasado y testeándolos con datos del futuro y viceversa.
- Disponibilidad de una variable de fecha de adquisición de los datos a nivel de mes o a nivel de año en el caso de que los datos de registro sean superiores a 3 años.
- Número elevado de datos en el *dataset* para poder desarrollar modelos de *machine learning.*

Analizando estos criterios, la búsqueda de datos se realizó en: kaggle (https://www.kaggle.com), OpenML (https://www.openml.org), Harvard Dataverse (https://dataverse.harvard.edu), y github (https://github.com). El proceso de búsqueda de *datasets* que cumpliesen con los requisitos especificados previamente fue complejo, encontrando finalmente los mencionados a continuación: *COVID-19 effect on Liver Cancer Prediction Dataset, UCI ML Drug Review dataset* y MIMIC-IV. Sin embargo, tras comenzar con los experimentos iniciales, el *dataset* principal sobre el que se ha decidido centra el trabajo debido a su relevancia y su mejor cumplimiento de los requisitos ha sido el MIMIC-IV, el cual es una de las bases de datos públicas de registros médicos más grandes del mundo, comprendiendo registros deidentificados de más de 40000 pacientes de la unidad de cuidados intensivos *de Beth Israel Deaconess Medical Center*. Este *dataset* contiene información de las EHR de los pacientes a lo largo del tiempo, incluyendo datos personales de los pacientes, admisiones en el hospital, diagnósticos, procedimientos realizados, servicios a los que han acudido, tipo de seguro médico, etc. Para garantizar a anonimidad de los datos del repositorio, la información temporal viene codificada en dos variables: *anchor_year* y *anchor_year_group*, donde la primera es un año deidentificado comprendido entre 2100 y 2200, y la segunda variable es un grupo de 3 años comprendido entre los años 2008 – 2019, que es de cuando constan los registros de la base de datos. Esta información permite a los investigadores inferir de forma aproximada el año en el que se trató a dicho paciente. Por ejemplo, si un paciente tiene un *anchor_year* 2158, y su *anchor_year_group* es 2011 – 2013, todas las hospitalizaciones para el paciente que ocurren en el año 2158 realmente sucedieron en algún momento entre 2011 y 2013. Además, las admisiones en el hospital también tienen una variable de tiempo llamada *admittime* con las mismas características que *anchor_year.*

Con el fin de poder realizar un correcto análisis temporal, se ha aplicado un procesado al *dataset* de la misma forma que el trabajo de Yao, H. et al., (2023), donde se consigue obtener cada admisión de un paciente de la base de datos original como una entrada del nuevo dataset, dando lugar a 239,246 entradas, y se consigue también estimar el año real en el que se ha producido esa admisión. Por un lado, se establecen como variables de entrada los códigos del sistema de Clasificación Internacional de Enfermedades (ICD en inglés) en sus versiones 9 y 10 asociados a los diagnósticos y procedimientos de cada admisión. El sistema ICD permite clasificar los síntomas, diagnósticos y procedimientos llevados a cabo en la práctica médica. El cambio de protocolo para pasar de códigos ICD-9 a ICD-10 se produjo el 1 de octubre de 2015, siendo este nuevo sistema mucho más específico y detallado con el uso de 68000 códigos por los 14000 del protocolo anterior.

Por otro lado, el *dataset* consta de variables características de cada paciente, como la etnia, o los ID de cada paciente y cada admisión. El uso de este repositorio considera dos tareas, por un lado, intentar predecir la mortalidad en el hospital de cada paciente. Por otro lado, intentar predecir el riesgo que tiene cada paciente de ser reingresado una vez pasados 15 días de su salida. Para ello, el *dataset* contiene las etiquetas correspondientes a cada una de estas variables para cada paciente. A la hora de desarrollas los modelos predictivos en el presente trabajo, se han realizado con el objetivo de abordar la tarea de la predicción de la mortalidad de los pacientes durante su estancia en el hospital.

Cabe destacar que el uso del *dataset* MIMIC-IV con fines de investigación requiere de la aprobación por parte de su comité, para lo cual es requisito realizar el curso *CITI Data or Specimens Only Research* el cual trata aspectos importantes sobre el uso de datos humanos con fines de investigación, como por ejemplo las políticas de privacidad y confidencialidad o principios éticos. El autor realizó el curso en la fecha 23 de diciembre de 2022 con código 53312103.

3.2 Herramientas utilizadas

En este apartado, se describirán las herramientas que se han utilizado para llevar a cabo el análisis de variabilidad temporal como los modelos de clasificación.

3.2.1 Herramientas de variabilidad temporal

Como ya se ha mencionado anteriormente, se ha utilizado el paquete de R EHRtemporalVariability para estudiar la variabilidad temporal de los datos del MIMIC-IV. En el presente trabajo, se ha utilizado tanto la versión oficial del paquete como una versión experimental ofrecida por el creador del mismo y tutor del trabajo, Carlos Sáez, para realizar un estudio de las distribuciones condicionadas del dataset. Este paquete ofrece los medios para visualizar y delinear analíticamente los *dataset shifts* de datos multimodales y altamente codificados. Una de las principales ventajas de esta herramienta es que no realiza suposiciones distribucionales, lo que permite un uso directo y una visualización analítica de los datos sin pérdida de información. El uso iterativo y metodológico de la herramienta puede identificar y definir cambios de referencia que puedan impedir posteriores investigaciones (Sáez, C. et al., 2016).

La librería EHRtemporalVariabilty se basa en los métodos probabilísticos de análisis de variabilidad temporal para obtener la IGT, definida en el punto 2.4.2, así como su representación por medio de mapas de calor temporales (DTH, del inglés *Data Temporal Heatmap*). La representación de la IGT de

los paquetes temporales de los datos del MIMIC-IV ha permitido la identificación tendencias, representadas como un cambio continuo en el tiempo de los paquetes de datos; cambios abruptos, representados como grandes espacios entre paquetes; subgrupos temporales, identificados como agrupaciones de los datos; estacionariedad, representada como círculos formados por los paquetes. Los paquetes de datos son representados con el año al que hacen referencia y están unidos por una trayectoria suavizada que representa la evolución de la información a lo largo del tiempo. La representación de la IGT de los datos también proporciona los medios para identificar los cambios para modelar los efectos estacionales o aplicar un método de agrupación para obtener los subgrupos temporales. Además de la representación de la IGT, el uso de los DTHs que permite la librería da la posibilidad de explorar los cambios usando las frecuencias relativas y absolutas a lo largo del tiempo, y simultáneamente, usando valores de múltiples variables. Finalmente, EHRtemporalVariability y la *app Shiny* publicada en www.ehrtemporalvariability.upv.es ofrecen una serie de opciones que permiten: cargar y procesar *datasets*; realizar un análisis de los datos por paquetes temporales por medio de la representación de los DTH y la IGT y visualizar los resultados por medio de gráficos interactivos. La figura 3.1 muestra un esquema del funcionamiento del paquete EHRtemporalVariability y el uso de la *app Shiny* para la representación de los datos.

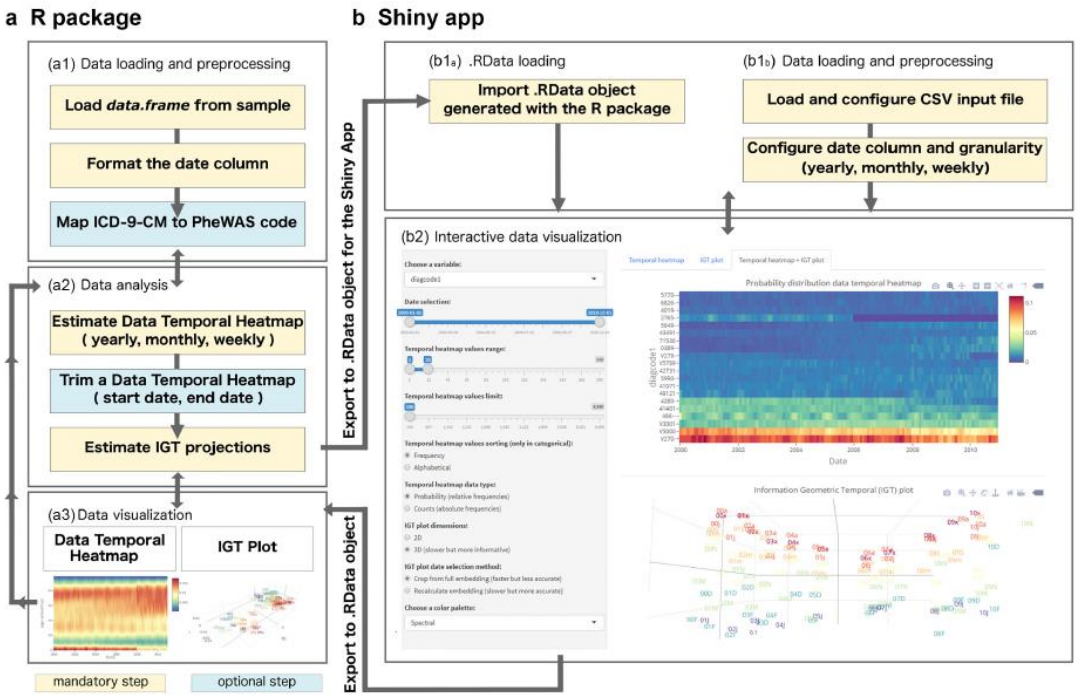


Figura 3.1: Esquema del paquete de R (a) y la *app* de *Shiny* (b) de Sáez, C., Gutiérrez-Sacristán, A., Kohane, I., García-Gómez, J. M., & Avillach, P. (2020) (Sáez, C. et al. 2020)

Para realizar el estudio de las distribuciones condicionadas, se ha utilizado una variante de la librería la cual permite dividir el dataset en función de la variable de interés, en nuestro caso la mortalidad del paciente, y hacer un estudio de la variabilidad temporal condicionada a esa variable.

3.2.2 Modelos predictivos utilizados

En el presente trabajo, se ha decidido utilizar métodos de aprendizaje conjunto—más conocidos como *ensemble learning*—el cual es un enfoque de aprendizaje automático que busca combinar varios modelos predictivos para realizar mejores predicciones a nivel global. Estos métodos buscar reducir tanto el sesgo como la varianza de los resultados obtenidos, y su elección está justificada debido a que, tras una revisión del estado del arte, en tareas de regresión y clasificación permiten obtener mejores resultados que cualquier otro método de *machine learning* individual (Maclin, R., y Opitz, D. 1999) (Ghimire, B. et al., 2012). De todas las técnicas de aprendizaje conjunto existente, en la presente investigación se ha optado por el uso de técnicas de *bagging* y *boosting* (Figura 3.2) basadas en árboles de decisión, en concreto *Random Forest* y *Gradient Boosting*, cuyo funcionamiento se muestra en la Figura 3.3. Las técnicas de *bagging* se basan en generar múltiples muestras de los datos de entrenamiento a partir de procesos de remuestreo, y entrenar distintos modelos de forma independiente en paralelo con las diferentes muestras, obteniendo finalmente un único modelo. Esto mejora la estabilidad y la precisión de los modelos de *machine learning*. Las técnicas de *boosting*, sin embargo, son técnicas iterativas en las que se generan numerosos clasificadores débiles en los que la salida de un clasificador se utiliza como entrada del siguiente con el objetivo de ir actualizando sus pesos y generando clasificadores más fuertes (Cha, G.-W. et al., 2021). Finalmente, se puede concluir que las técnicas de bagging ofrecen una menor varianza, pero mayor sesgo que las de boosting, las cuales ofrecen menor sesgo pero mayor varianza (Polikar, R., 2012).

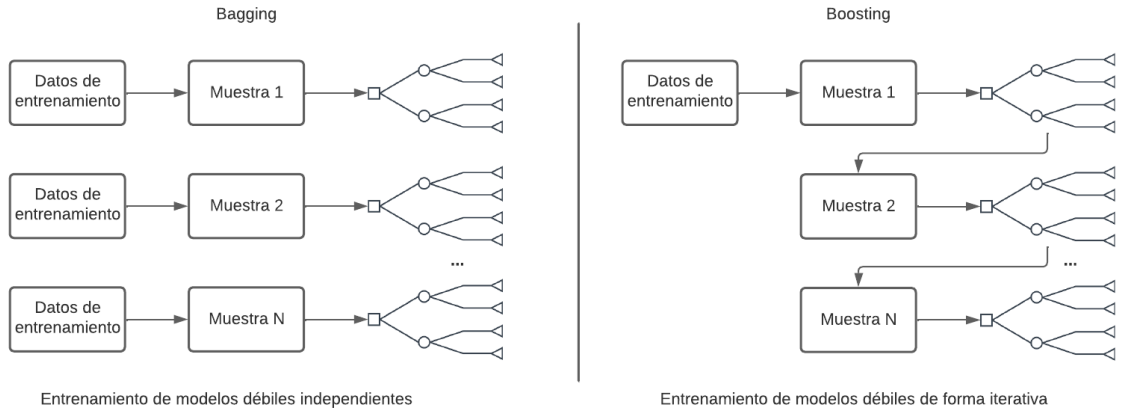


Figura 3.2: Flujo de trabajo de los modelos de *bagging* y *boosting* . Incluye referencia a Figura (me aparece tachada). Aplica para esta figura como para las otras que no sean de creación propia.

3.2.1 Árboles de decisión, *Gradient Boosting* y *Random Forest*

Un **árbol de decisión** es un tipo de modelo utilizado en *machine learning* para tareas de clasificación y regresión(Breiman, L., 2017). Es una estructura con forma de árbol la cual representa una serie de decisiones basadas en los valores de las variables de entrada, en el que cada nodo interno del árbol representa una prueba del valor de la característica y cada nodo hoja representa un valor de salida o clase predicha. La idea es dividir de forma recursiva el espacio de entrada en regiones cada vez más

pequeñas, cada una asociada con una clase o valor de salida específico, de manera que las muestras de cada región sean lo más similares posibles a la variable de salida.

Una vez el árbol está creado, puede ser usado con funciones predictivas recorriendo el árbol desde la raíz hasta los nodos, siguiendo el camino correspondiente a las características de los datos de entrada (Breiman, L., 2017).

El ***Gradient Boosting***, como su propio nombre indica, es una técnica de *boosting* presentada en el trabajo Friedman, J. H. (2001) utilizada para problemas de regresión y clasificación, el cual genera un modelo de predicción fuerte por medio de la unión de modelos de predicción débiles. Es uno de los algoritmos más robustos y utilizados en *machine learning* en todo el mundo (Friedman, J. H. 2001). Es un algoritmo de optimización numérico que tiene como objetivo encontrar un modelo aditivo que minimice la función de pérdidas (Friedman, J. H. 2001). Para ello, el algoritmo va añadiendo de forma iterativa nuevos árboles de decisión reduciendo la función de pérdida en cada iteración mediante el uso de clasificadores débiles, como se puede observar en la figura 3.2. Este algoritmo está especialmente enfocado en la reducción del sesgo de los modelos predictivos, lo cual es especialmente útil cuando se trabaja con *datasets* con pocos datos (Cha, G.-W. et al., 2021).

Random forest, a su vez, es un método de *enseble learning* propuesto en Ho, T. K., (1995), el cual combina múltiples árboles de decisión para mejorar el rendimiento y la robustez del modelo. Se basa en la técnica del *bagging*, dividiendo el conjunto de datos de entrenamiento en muestras y entrenando múltiples árboles con las distintas muestras y obteniendo el clasificador final como la media de todos los entrenamientos (Friedman, J. H. 2001). Algunas de las ventajas de este método es que a medida que aumenta el número de árboles, se evita el efecto del sobreajuste y el modelo está menos afectado por los *outliers*. Además, esta técnica tiene un rendimiento mucho mayor que otras técnicas de *machine learning* cuando las clases de los datos están muy desbalanceadas.

Con el objetivo de implementar estos modelos para llevar a cabo la clasificación de los datos del MIMIC-IV, se ha utilizado la librería de Python *Scikit Learn* ("Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research, 12, 2825-2830.")*(Nestor, B. et al., 2018). Scikit Learn es una librería gratuita lanzada en 2007 por David Cournapeau que contiene algoritmos de clasificación, regresión, clustering y reducción de dimensionalidad. La principal ventaja de esta librería es la variedad de módulos y algoritmos que facilitan el aprendizaje y trabajo en las labores de investigación.

Figura 3.3: Comparación de la estructura y flujo de trabajo de los algoritmos de *Gradient Boosting* y *Random Forest* Me aparece la referencia tachada.

Capítulo 4

Métodos

En este capítulo se presentan los métodos que han sido utilizados para cumplir los objetivos planteados en este trabajo. En primer lugar, se expone el preprocesado que se ha realizado sobre la base de datos MIMIC-IV. A continuación, se describirá el proceso de obtención del análisis temporal y la estrategia de entrenamiento de los modelos, así como la evaluación de los mismos. Por último, se explicará la metodología seguida para validar el cumplimiento del objetivo principal.

4.1 Preprocesado de los datos

Como ya se ha mencionado previamente en el apartado 3.1, para justificar el cumplimiento del objetivo principal, se ha optado por el uso del repositorio MIMIC-IV. Para obtener los datos de interés para la realización del trabajo, se ha realizado un primer preprocesado de los datos siguiendo el procedimiento realizado en el trabajo de Yao, H. et al., (2023), en el cual se ha utilizado la información de los pacientes, admisiones y los códigos ICD de diagnósticos y procedimientos del repositorio original para generar un *dataset* sobre el que poder aplicar un algoritmo de *machine learning* con el objetivo de llevar a cabo las tareas descritas en el apartado 3.1. Para decodificar la fecha y obtener el año real de cada admisión de forma aproximada, en el trabajo de Yao, H. et al., (2023) lo realizan de la siguiente manera donde la variable *real_anchor_year* procede de obtener el primero de los 3 años que forman el *anchor_year_group*:

$$Fecha\_final\_aprox = admittime - anchor\_year + real\_anchor\_year$$

Además, ha sido necesario realizar un nuevo preprocesado de los datos debido a la estructura que estos presentaban, en el que se ha llevado a cabo una codificación de la etnia de los pacientes, así como de los códigos ICD de diagnóstico y procedimientos. Para ello, se ha aplicado una codificación *one-hot*, en la cual se crea una variable *dummy* (binaria) para cada etnia y cada código ICD, adquiriendo un valor “True” o “False” en función de si presentaban o no ese valor.

Después, con el objetivo de simplificar el dataset obtenido debido al elevado número de códigos ICD únicos representados, se ha realizado la agrupación de los códigos ICD-9 e ICD-10 por capítulos tanto para los procedimientos como para los diagnósticos, para posteriormente unir los capítulos que hiciesen referencia a los mismos términos y tener cada uno los capítulos de códigos de diagnósticos y procedimientos ICD-9 e ICD-10 en una única variable. Esta agrupación ayuda a reducir de partida problemas de actualización temporal de códigos (Nestor, B. et al., 2018).

La agrupación se ha realizado conforme a lo representado en las Tablas 4.1, 4.2 y 4.3.

Tabla 4.1A: Códigos de diagnósticos ICD-9 según https://dexur.com/pcs9/ y https://www.icd10data.com/Convert

ICD-9
001-139- Infectious And Parasitic Diseases
140-239- Neoplasms

240-279- Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders
280-289- Diseases Of The Blood And Blood-Forming Organs
290-319- Mental Disorders
320-389- Diseases Of The Nervous System And Sense Organs
390-459- Diseases Of The Circulatory System
460-519- Diseases Of The Respiratory System
520-579- Diseases Of The Digestive System
580-629- Diseases Of The Genitourinary System
630-679- Complications Of Pregnancy, Childbirth, And The Puerperium
680-709- Diseases Of The Skin And Subcutaneous Tissue
710-739- Diseases Of The Musculoskeletal System And Connective Tissue
740-759- Congenital Anomalies
760-779- Certain Conditions Originating In The Perinatal Period
780-799- Symptoms, Signs, And Ill-Defined Conditions
800-999- Injury And Poisoning
V01-V91- Supplementary Classification Of Factors Influencing Health Status And Contact With Health Services
E000-E999- Supplementary Classification Of External Causes Of Injury And Poisoning

Tabla 4.1B: Códigos de diagnósticos ICD-10 según https://dexur.com/pcs9/ y https://www.icd10data.com/Convert

ICD-10
A00-B99- Certain infectious and parasitic diseases
C00-D49- Neoplasms
D50-D89- Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
E00-E89- Endocrine, nutritional and metabolic diseases
F01-F99- Mental, Behavioral and Neurodevelopmental disorders
G00-G99- Diseases of the nervous system
H00-H59- Diseases of the eye and adnexa
H60-H95- Diseases of the ear and mastoid process
I00-I99- Diseases of the circulatory system
J00-J99- Diseases of the respiratory system
K00-K95- Diseases of the digestive system
L00-L99- Diseases of the skin and subcutaneous tissue
M00-M99- Diseases of the musculoskeletal system and connective tissue
N00-N99- Diseases of the genitourinary system
O00-O9A- Pregnancy, childbirth and the puerperium
P00-P96- Certain conditions originating in the perinatal period
Q00-Q99- Congenital malformations, deformations and chromosomal abnormalities
R00-R99- Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
S00-T88- Injury, poisoning and certain other consequences of external causes
V00-Y99- External causes of morbidity
Z00-Z99- Factors influencing health status and contact with health services

Tabla 4.1C: Agrupación de los códigos de diagnósticos ICD-9 e ICD-10 en un único capítulo según https://dexur.com/pcs9/ y https://www.icd10data.com/Convert

Agrupación ICD-9 ICD-10	Correspondencia ICD-9	Correspondencia ICD-10
Infectious and Parasitic Diseases	001-139	A00-B99
Neoplasms	140-239	C00-D49
Endocrine, Nutritional and Metabolic Diseases, and Immunity Disorders	240-279	E00-E89
Diseases of the Blood and Blood-Forming Organs	280-289	D50-D89
Mental Disorders	290-319	F01-F99
Diseases of the Nervous System and Sense Organs	320-389	G00-G99, H00-H59, H60-H95
Diseases of the Circulatory System	390-459	I00-I99
Diseases of the Respiratory System	460-519	J00-J99
Diseases of the Digestive System	520-579	K00-K95
Diseases of the Genitourinary System	580-629	N00-N99
Complications of Pregnancy, Childbirth, and the Puerperium	630-679	O00-O9A
Diseases of the Skin and Subcutaneous Tissue	680-709	L00-L99
Diseases of the Musculoskeletal System and Connective Tissue	710-739	M00-M99
Congenital malformations, deformations and chromosomal abnormalities	740-759	Q00-Q99
Certain Conditions Originating in the Perinatal Period	760-779	P00-P96
Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	780-799	R00-R99
Injury and Poisoning	800-999	S00-T88
Supplementary Classification of Factors Influencing Health Status and Contact with Health Services	V01-V91	Z00-Z99
External causes of morbidity	E000-E999	V00-Y99

Tabla 4.2A: Códigos de procedimientos ICD-9 según https://dexur.com/pcs9/ y https://www.icd10data.com/Convert

ICD-9
00-00- Procedures And Interventions , Not Elsewhere Classified
01-05- Operations On The Nervous System
06-07- Operations On The Endocrine System
08-16- Operations On The Eye
17-17- Other Miscellaneous Diagnostic And Therapeutic Procedures
18-20- Operations On The Ear
21-29- Operations On The Nose, Mouth, And Pharynx

30-34- Operations On The Respiratory System
35-39- Operations On The Cardiovascular System
40-41- Operations On The Hemic And Lymphatic System
42-54- Operations On The Digestive System
55-59- Operations On The Urinary System
60-64- Operations On The Male Genital Organs
65-71- Operations On The Female Genital Organs
72-75- Obstetrical Procedures
76-84- Operations On The Musculoskeletal System
85-86- Operations On The Integumentary System
87-99- Miscellaneous Diagnostic And Therapeutic Procedures

Tabla 4.2B: Códigos de procedimientos ICD-10 según https://dexur.com/pcs9/ y https://www.icd10data.com/Convert

ICD-10
0- Medical and Surgical
1- Obstetrics
2- Placement
3- Administration
4- Measurement and Monitoring
5- Extracorporeal or Systemic Assistance and Performance
6- Extracorporeal or Systemic Therapies
7- Osteopathic
8- Other Procedures
9- Chiropractic
B- Imaging
C- Nuclear Medicine
D- Radiation Therapy
F- Physical Rehabilitation and Diagnostic Audiology
G- Mental Health
H- Substance Abuse Treatment
X- New Technology

Tabla 4.2C: Agrupación de los códigos de procedimientos ICD-9 e ICD-10 en un único capítulo según https://dexur.com/pcs9/ y https://www.icd10data.com/Convert

Agrupación ICD-9 ICD-10	**Correspondencia ICD-9**	**Correspondencia ICD-10**
Medical and Surgical	01-05, 06-07, 08-16, 18-20, 21-29, 30-34, 35-39, 40-41, 42-54, 55-59, 60-64, 65-71, 76-84, 85-86	0
Obstetrical Procedures	72-75	1
Placement		2
Administration		3

Measurement and Monitoring		4
Extracorporeal or Systemic Assistance and Performance		5
Extracorporeal or Systemic Therapies		6
Osteopathic		7
Other Procedures	17-17, 87-99, 00-00	8, 0W, 0X, 0Y
Chiropractic		9
Imaging		B
Nuclear Medicine		C
Radiation Therapy		D
Physical Rehabilitation and Diagnostic Audiology		F
Mental Health		G
Substance Abuse Treatment		H
New Technology		X

En el caso de los procedimientos, debido a que la mayoría de los capítulos correspondientes a códigos ICD-9 se agrupaban en la categoría “Medical and Surgical”, se ha dividido esa categoría en los subcapítulos correspondientes y se ha realizado la agrupación con dichos subcapítulos.

Tabla 4.3: Agrupación de los códigos ICD 9 correspondientes al capítulo “Medical and Surgical” con los subcapítulos de los códigos ICD-10

Medical and Surgical	Correspondencia ICD-9	Correspondencia ICD-10
Nervous System	01-05	00, 01
Endocrine System	06-07	0G
Operations on the Eye	08-16	08
Ear, Nose, Sinus, Mouth and Throat	18-20, 21-29	09, 0C
Respiratory System	30-34	0B
Cardiovascular System	35-39	02, 03, 04, 05, 06
Lymphatic and Hemic Systems	40-41	07
Digestive System	42-54	0D, 0F
Urinary System	55-59	0T
Male Reproductive System	60-64	0V
Female Reproductive System	65-71	0U
Musculoskeletal System	76-84	0K, 0L, 0M, 0N, 0P, 0Q, 0R, 0S
Integumentary System	85-86	0H, 0J

Por último, se han eliminado las variables correspondientes a los ID de los pacientes y de las admisiones con el fin de reducir el sesgo de los modelos predictivos, ya que no aportaban información relevante a la hora de llevar a cabo la clasificación al ser números establecidos de forma aleatoria.

4.2 Análisis temporal

Con el fin de examinar todos los datos de los que se dispone de la mejor manera posible, se han realizado diferentes técnicas de análisis exploratorio, exponiendo así sus características y las posibles relaciones entre ellos.

4.2.1 Análisis multivariante

Se ha iniciado el estudio de la variabilidad temporal mediante un análisis multivariante con el propósito de comprender las relaciones y patrones que se generan en el dataset a lo largo del tiempo. Este enfoque permite realizar un análisis inicial que sienta las bases para un estudio más detallado de las distribuciones de probabilidad individuales de cada variable. Este enfoque permite examinar de forma detallada la forma en que las variables cambian en conjunto a lo largo de los diferentes periodos de tiempo, permitiendo así la identificación de tendencias, patrones cíclicos o cambios significativos en la estructura de los datos.

Para llevar a cabo el estudio, se ha aplicado un Análisis de Correspondencia Múltiple (MCA en inglés) (Abdi, H., & Valentin, Dominique., 2007) con el objetivo de poder evaluar así las posibles correlaciones entre variables en un espacio de menor dimensionalidad.

El objetivo del análisis de correspondencia es analizar las variables categóricas transformadas en tablas cruzadas y demostrar los resultados de una manera gráfica (Costa, P. S. et al., 2013). En el análisis se han incluido el total de las entradas de la base de datos, *n*=239,246, no siendo necesario descartar ninguna debido a que el repositorio no contaba con datos perdidos.

Una vez realizado el MCA, se han conservado las dos primeras dimensiones en base a los siguientes criterios: (1) a pesar de que el número de dimensiones a conservar no es un parámetro predefinido, algunos autores recomiendan la obtención de dos dimensiones con el objetivo de facilitar la interpretación y representación de los datos (Gifi, A., 1991), (2) a la fecha de la realización del trabajo, el código que se ha utilizado para analizar la variabilidad temporal expuesto en el punto 3.2.1 solo permite la unión de dos mapas de probabilidad con los que generar un nuevo mapa que contenga la información de los dos anteriores, en este caso la información correspondiente a las dos dimensiones del MCA, con el que posteriormente se obtiene la IGT. De esta forma la IGT obtenido será más representativa del conjunto de datos completo que realizándola de cada dimensión de forma independiente.

De esta manera, la primera dimensión representa un 7.44% de la variabilidad total de los datos, y la segunda un 3.93%, dando lugar a una varianza total de un 11.36%. A pesar de que el valor de variabilidad total de los datos no es muy elevado, cumple el objetivo para el que se ha propuesto el método, el cual es realizar un análisis temporal multivariante de los datos, y no encontrar el método de reducción de la dimensionalidad que mejor se ajuste al dataset.

Una vez realizado el MCA, se ha evaluado la variabilidad temporal siguiendo el procedimiento descrito a continuación:

(1) Obtener los índices de los pacientes fallecidos y no fallecidos del dataset original

(2) Generar dos nuevos *datasets* con las coordenadas del MCA para los pacientes fallecidos y no fallecidos para poder llevar a cabo un estudio de la probabilidad condicionada en busca de posibles *concept shifts*
(3) Obtener el mapa de probabilidad de cada dimensión del MCA para todos los pacientes juntos, los pacientes fallecidos y los pacientes no fallecidos usando el paquete EHRtemporalVariability con el objetivo de identificar variaciones en las distribuciones de probabilidad, tanto del total de los datos como de la probabilidad condicionada para identificar los *dataset shifts* existentes
(4) Unir las dos dimensiones de los mapas de probabilidad de los pacientes fallecidos y no fallecidos para generar un único mapa de probabilidad y normalizar para poder obtener posteriormente la IGT correspondiente a las probabilidades condicionadas
(5) Obtener la Información Geométrica Temporal del nuevo mapa de probabilidad utilizando el paquete EHRtemporalVariability para graficar las variaciones en las distribuciones de probabilidad de los datos

4.2.2 Análisis univariante

En segundo lugar y con la finalidad de obtener una información más detallada sobre la variabilidad temporal y ver qué características son las que tienen una mayor influencia en la variabilidad analizada en el apartado 4.2.1, se ha realizado un análisis de la distribución de probabilidad todas las variables de manera individual.

En primera instancia, se ha optado por obtener en un único gráfico el mapa de probabilidad de las frecuencias de cada uno de los códigos ICD, tanto para diagnóstico como para procedimientos, a lo largo del tiempo. Esta información ha sido utilizada para posteriormente obtener la IGT de los códigos. Los códigos ICD constituyen las principales variables de entrada de la base de datos y son las características con mayor relevancia a la hora de realizar la clasificación de los datos, por lo que obtener este resultado aporta información a la hora de estudiar qué variables están generando una mayor variabilidad temporal en los datos analizando las diferencias en las distribuciones de los códigos a lo largo del tiempo.

Una vez obtenido este primer análisis general, el procedimiento ha sido muy similar al realizado en el apartado anterior:

(1) Obtener los subconjuntos correspondientes a los pacientes que han fallecido y que no.
(2) Calcular el mapa de calor de la distribución de probabilidad para cada variable, tanto del conjunto total de datos (p(x)) como de los dos subconjuntos (p(x|y)) mediante el paquete EHRtemporalVariability para poder identificar los diferentes *dataset shifts* en cada variable
(3) Calcular la distribución de probabilidad condicionada de cada clase menos la del dataset original, p(x|y) – p(x), con el fin de realizar un estudio de la variabilidad de la información exclusiva de cada clase.
(4) Unir los mapas de probabilidad de los dos subconjuntos para cada variable con el objetivo de formar un nuevo mapa de probabilidad y normalizar para poder obtener posteriormente la IGT correspondiente a las probabilidades condicionadas

(5) Obtener la IGT mediante el paquete EHRtemporalVariability para graficar las variaciones en las distribuciones de probabilidad de los datos

4.3 Evaluación de modelos y evaluación lote a lote

Con el objetivo de analizar el rendimiento de los modelos de clasificación a lo largo del tiempo y poder comprarlo con los resultados obtenidos del análisis temporal mediante los IGT-plots, se ha dividido el conjunto inicial de datos en lotes en función del año, obteniendo de esta forma 12 conjuntos de datos, con los que posteriormente se han entrenado y validado los modelos.

Una vez realizado el preprocesado de los datos, se han dividido los datos en lotes temporales en función de los años, de tal manera que se obtiene un conjunto de datos correspondiente a cada año, y los datos de cada año se han dividido a su vez en los subconjuntos de entrenamiento, test y validación. Para ello se ha establecido un criterio de dividir los datos iniciales en un 80% para entrenamiento y el 20% restante para test. Del 80% de los datos originales que habían sido reservados para entrenamiento, se han establecido un 70% para el entrenamiento real de los modelos y el 30% restante para llevar a cabo la validación.

Una vez establecida la partición de los datos, se ha llevado a cabo un primer entrenamiento de los modelos para seleccionar los hiperparámetros de cada modelo. Para ello, se han definido diferentes valores de número de árboles y de profundidad máxima (tabla 4.4), para posteriormente establecer todas las combinaciones posibles entre ellos y entrenar un modelo para cada lote temporal de datos con cada posible combinación de hiperparmámetros, siguiendo un procedimiento de búsqueda de los hiperparámetros óptimos conocido como *grid search*. Además, para reducir el coste temporal, los modelos han sido entrenados en paralelo, y no de forma secuencial (Becker, C. et al., 2019):

Tabla 4.4: Opciones de hiperparámetros seleccionados para el entrenamiento de los modelos

	Random Forest	**Gradient Boosting**
Número de árboles	200, 300, 400, 500, 600	4, 5, 6, 7, 8, 9, 10
Profundidad máxima	50, 100, 200, 300, 400	1, 2, 3, 4, 5

Finalmente, para evaluar el resultado de los modelos y seleccionar la mejor combinación de hiperparámetros, se ha obtenido el área bajo la curva ROC de los resultados de la clasificación para los datos de validación, y se ha calculado la media de este parámetro de los 12 lotes temporales para cada posible combinación de hiperparámetros. La elección de esta métrica ha estado justificada por el gran desbalanceo de las clases del dataset, siendo la prevalencia de la clase negativa muy superior a la de la clase positiva, ya que a pesar de no ser una métrica completamente inmune al desbalanceo de clases, no es dependiente de un umbral de saturación, como sí lo pueden ser otras métricas como el F1-Score, la precisión o la sensibilidad. De esta forma, con el uso del área bajo la curva ROC no es necesario llevar a cabo el proceso de ajuste de este umbral y ofrece una visión del rendimiento más amplia (Brabec, J. et al., 2020). La curva ROC es un gráfico que muestra el rendimiento de modelos de clasificación calculando la relación entre la tasa de verdaderos positivos (TPR, ecuación 4.1)—también conocida

como sensibilidad—y la de falsos positivos (FPR, ecuación 4.2) (Hand, D. J., y Till, R. J. 2001) . En las ecuaciones 4.1 y 4.2 aparecen los términos Verdaderos Positivos (TP en inglés), que hace referencia a los sujetos diagnosticados como positivos que realmente lo son; Falsos Negativos (FN en inglés), que hace referencia a los sujetos diagnosticados como negativos cuando realmente son positivos; Falsos Positivos (FP en inglés), que se refiere a los sujetos diagnosticados como positivos cuando realmente son negativos; o Verdaderos Negativos (TN en inglés), refiriéndose a los sujetos diagnosticados como negativos siendo realmente negativos ante dicha prueba:

$$TPR = \frac{TP}{TP+FN} \text{ (4.1)}$$

$$FPR = \frac{FP}{FP+TN} \text{ (4.2)}$$

En la figura 4.1 obtenida del trabajo de Brabec, J. et al., (2020) se puede observar como la curva ROC no tiene en cuenta el parámetro del umbral de saturación, presentando el mismo resultado para dos umbrales diferentes, a diferencia de la curva Precisión-Sensibilidad (PR).

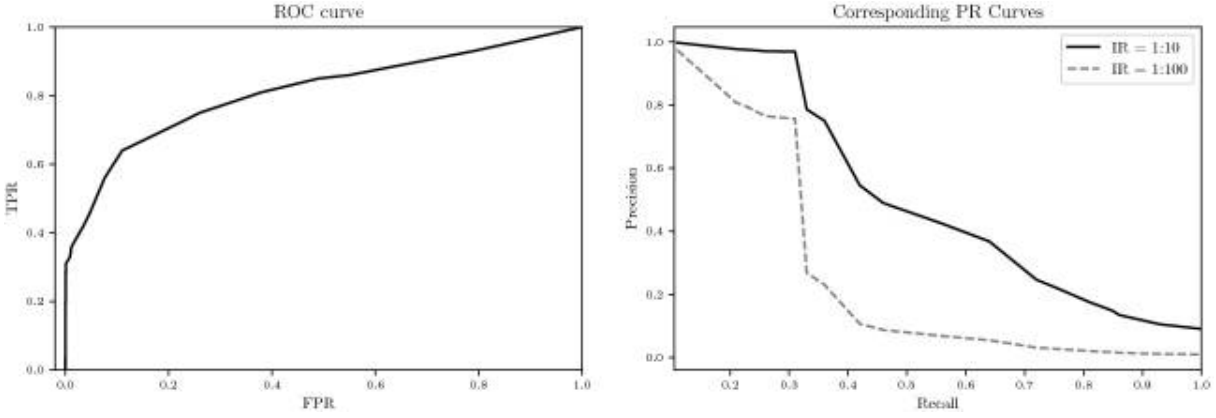


Figura 4.1: Representación de curvas ROC y PR ante dos ratios de desbalanceo distintos de acuerdo al trabajo de Brabec, J. et al., (2020)

Una vez obtenidos los mejores hiperparámetros para los dos modelos mediante el mejor valor de área bajo la curva ROC como se ha descrito, se han usado los modelos de cada año con los hiperparámetros establecidos y se han obtenido las predicciones y las etiquetas usando los datos de entrenamiento anteriores y los datos de test, evaluando cada modelo entrenado con los datos de entrenamiento de un lote temporal con los datos de test cada lote temporal para poder analizar de esta forma el desempeño de cada modelo a lo largo del tiempo.

Con el fin de evaluar este desempeño, se ha utilizado como primera métrica el área bajo la curva ROC, al igual que en el caso anterior, pero también se han calculado la sensibilidad (*Recall* en inglés, proporción de sujetos positivos que realmente lo son), que viene definida por la ecuación 4.1; precisión (*Precision* en inglés, cantidad de sujetos que realmente son positivos entre los que el modelo predice como positivos), descrita en la ecuación 4.3; *F1-Score*, descrito en la ecuación 4.4 y exactitud o tasa de acierto (*Accuracy* en inglés), definida en la ecuación 4.5.

$$Precision = \frac{TP}{TP+FP} \text{ (4.3)}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4.4)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4.5)$$

Sin embargo, de forma previa al cálculo de estas métricas ha sido necesario llevar a cabo una saturación de las predicciones debido al problema del desbalanceo entre la clase positiva y negativa mencionado anteriormente.

Para ello, se ha calculado la curva ROC con el fin de obtener los valores de especificidad (proporción de sujetos negativos que realmente lo son, definida en la ecuación 4.6), sensibilidad (ecuación 4.1) y umbral de saturación para cada predicción de cada clase, para posteriormente ordenar todos los límites obtenidos en función del índice de Youden, descrito en la ecuación 4.7, para quedarnos con un único límite para cada clase que utilizaremos para saturar las predicciones y poder evaluar los modelos con las métricas descritas anteriormente solucionando el problema del desbalanceo de clases.

$$Especificidad = \frac{TN}{TN+FP} \quad (4.6)$$

$$Índice\ de\ Youden = Sensibilidad + Especificidad - 1 \quad (4.7)$$

4.4 Asociación del análisis de variabilidad temporal y resultados de modelos predictivos

La comparación de los resultados de los modelos de inteligencia artificial con el análisis temporal permite aportar información sobre si las variaciones en las distribuciones de probabilidad a lo largo del tiempo tienen influencia en los resultados de los modelos de *machine learning,* de forma que una vez obtenidos tanto los resultados del análisis temporal como de los modelos de clasificación, se ha llevado a cabo un proceso de validación para comprobar que ambos procedimientos dan lugar a resultados similares y justificar de esta forma el objetivo principal del trabajo. Para ello, la estrategia a seguir ha sido calcular una proyección de las métricas obtenidas de los modelos sobre los datos de test descritas en el apartado 4.3, de forma equivalente a la trayectoria ofrecida por el IGT-plot, para de esta forma poder compararla con el mismo IGT-plot obtenido del análisis de variabilidad temporal. Para lograr este resultado, el procedimiento ha sido el siguiente:

En primer lugar, se ha obtenido la trayectoria de información de los resultados de los modelos de clasificación de la misma forma en la que se ha explicado en el apartado 2.4.2, donde en vez de utilizar la matriz de disimilitud de las PDFs, se ha aplicado el escalado multidimensional sobre la matriz resultante de las métricas de los modelos evaluados año a año, para posteriormente representarlo en un espacio de 2 dimensiones.

Una vez obtenida las proyecciones temporales tanto del análisis temporal como de los modelos, se ha evaluado su similitud aplicando en cada grupo los datos por medio del uso de un agrupamiento jerárquico aglomerativo utilizando la distancia de Manhattan debido a que, según Strauss, T., y Maltitz, M. J. von., (2017) es una distancia más robusta y menos sensible a *outliers*; y el método de enlace completo, descrito en la ecuación 4.8.

$$\mathrm{d}(\mathrm{Ci}\,,\mathrm{Cj}\,) \;=\; \underset{\substack{x_l \in \mathrm{Ci} \\ x_m \in \mathrm{Cj}}}{\mathrm{Max}} \{\mathrm{d}(x_l\,, x_m)\} \quad (4.8)$$

Con el objetivo de determinar el número de agrupaciones que se van a realizar sobre cada conjunto de datos, se ha aplicado el método del codo. Para ello, se han calculado la suma de las distancias al cuadrado de cada objeto del agrupamiento a su centroide variando el número de agrupamiento de 1 a 10. Posteriormente se ha representado en una gráfica lineal esta distancia respecto al número de agrupamientos. El punto en el que se observa un cambio brusco en la gráfica será el número de agrupamientos óptimos para ese conjunto de datos. Con el fin de obtener este punto de la forma más objetiva posible, se ha establecido como criterio de corte el valor absoluto de la media de todas las diferencias de cada distancia del agrupamiento con la distancia siguiente, y se ha establecido el número de agrupamientos óptimo como el primer punto correspondiente al valor absoluto de la diferencia de la distancia con la distancia siguiente que sea menor que el criterio de corte. Además, esta selección se ha justificado con el análisis de las distancias del dendrograma.

Finalmente, para justificar la existencia de una asociación significante estadísticamente de los resultados obtenidos, se ha aplicado un test chi-cuadrado entre el resultado de las dos agrupaciones jerárquicas estableciendo un p-valor de 0.05 como límite de rechazo de la hipótesis nula, la cual establece la dependencia de los dos agrupamientos (tablas 5.3 y 5.4).

Capítulo 5

Resultados

En este capítulo, se presentarán los resultados obtenidos durante la realización del proyecto. Se compone en primer lugar de la estructura y contenido del dataset final utilizado, para posteriormente mostrar los resultados de los análisis de la variabilidad temporal y de los modelos de clasificación, y finalmente evaluar la relación entre los resultados anteriores.

5.1 Dataset final

Tras realizar la primera fase del preprocesado el dataset presentaba 239,246 admisiones con 11 variables con la estructura mostrada en las tablas 5.1 y 5.2:

Tabla 5.1: Estructura del dataset después de realizar el primer preprocesado

Nombre de la variable	Tipo de variable	Descripción
subject_id	Categórica	ID de cada paciente
hadm_id	Categórica	ID de cada admisión
admittime	Fecha	Año codificado entre 2100 y 2200 que indica la fecha de admisión
real_admit_year	Fecha	Año real de la fecha de admisión estimado
age	Continua	Edad del paciente de forma estimada debido a la desidentificación de los datos
gender	Categórica	Género del paciente
ethnicity	Categórica	Etnia del paciente
mortality	Categórica	Indicativo de si el paciente ha fallecido durante la estancia en el hospital
readmission	Categórica	Indicativo de si el paciente ha reingresado en el hospital en los 15 días próximos a darle el alta
diagnoses	Categórica	Códigos ICD de diagnósticos asociados a cada intervención
procedures	Categórica	Códigos ICD de procedimientos asociados a cada intervención

Tabla 5.2: Número de pacientes fallecidos y no fallecidos a lo largo del tiempo

	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Pacientes fallecidos	1189	337	194	1119	337	210	1267	357	271	1296	284	109
Pacientes no fallecidos	50513	14869	9021	36917	12642	9105	35190	12296	8983	29875	8609	3556

Una vez aplicado el segundo procesado de los datos con el objetivo de codificar la etnia de los pacientes y los códigos ICD y de agrupar los códigos por capítulos, el dataset pasó a tener 61 variables. La distribución de las frecuencias de los diagnósticos y los procedimientos se puede observar en las figuras 5.1 y 5.2.

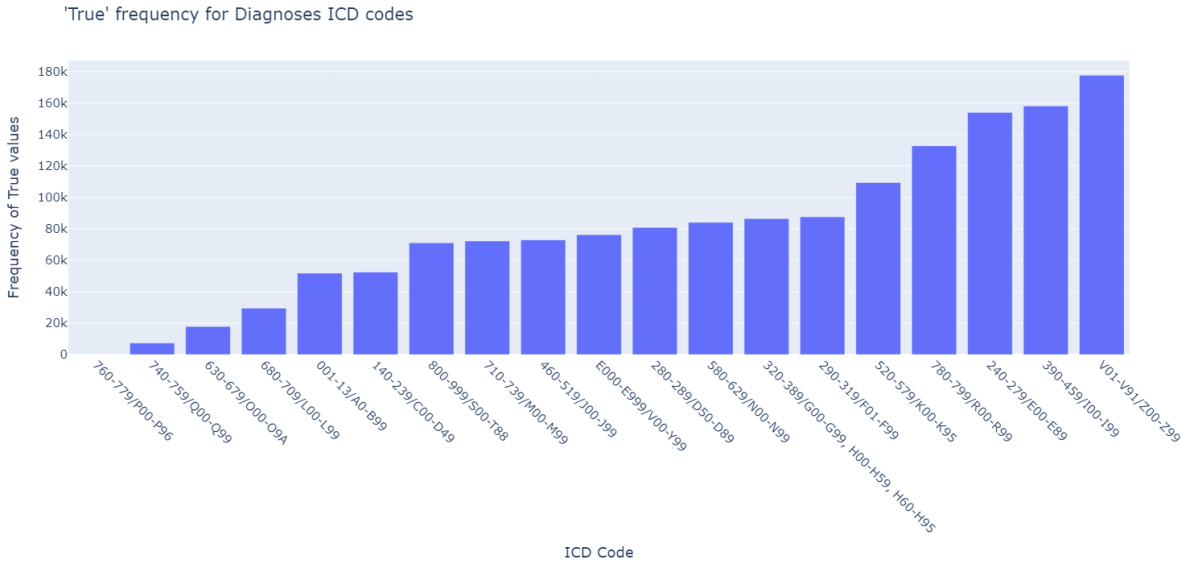


Figura 5.1: Distribución de la frecuencia de códigos ICD de diagnósticos agrupados en capítulos, donde en el eje x se puede observar los códigos ICD-9/ICD-10 correspondientes a cada capítulo según la tabla 4.1

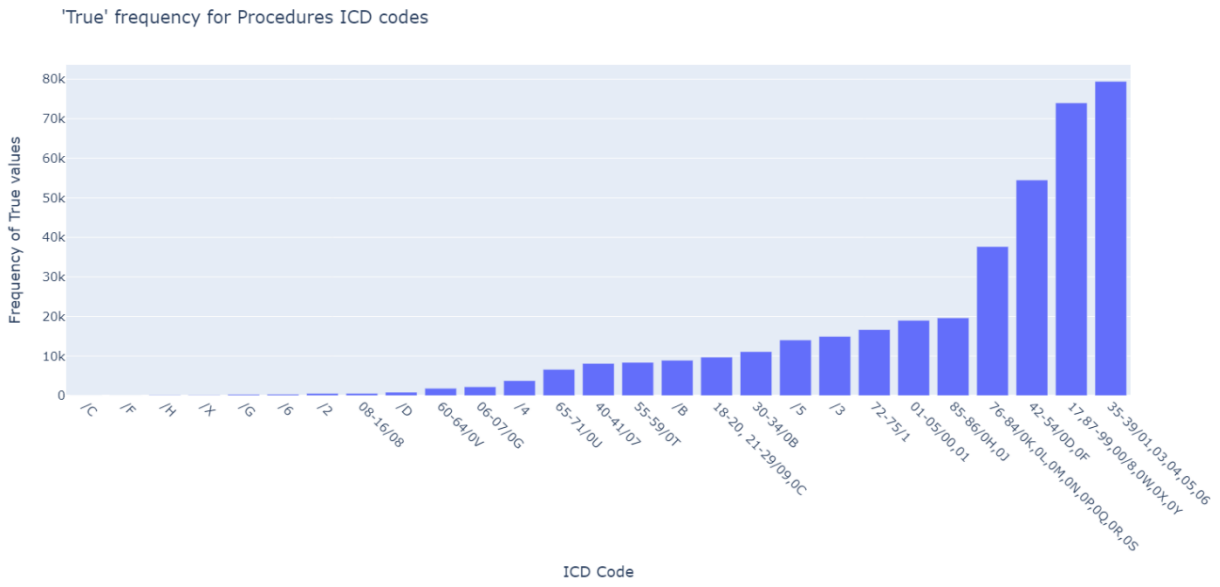


Figura 5.2: Distribución de la frecuencia de códigos ICD de procedimientos agrupados en capítulos, donde en el eje x se puede observar los códigos ICD-9/ICD-10 correspondientes a cada capítulo según las tablas 4.2 y 4.3

De las 239,246 entradas totales que tiene el dataset, 232,276 hacen referencia a admisiones de pacientes no fallecidos, y tan solo 6970 a pacientes fallecidos. Como se puede observar, la clase de pacientes fallecidos (clase positiva) solo representa un 2.91% de los datos totales.

5.2 Resultados de la variabilidad temporal

5.2.1 Análisis multivariante

La figura 5.3 (a) muestra la representación de los puntos del MCA en función de las clases, donde se puede observar una deficiente pero ligera separación entre la clase positiva y negativa a partir de las dos componentes principales, ya que se puede observar que en la parte izquierda del gráfico no hay representación de pacientes fallecidos. Con el objetivo de averiguar de donde procedía la diferenciación entre los grupos de la izquierda y la derecha del gráfico, se ha representado el resultado del MCA en función de los datos anteriores y posteriores a 2014 (figura 5.3 (b)), donde se observa que el origen de la separación no es este. De este gráfico también se puede concluir que en la representación bidimensional que se observa, los datos de ambos grupos tienen una separación deficiente, teniendo los datos posteriores a 2014 valores algo más positivos en la segunda dimensión.

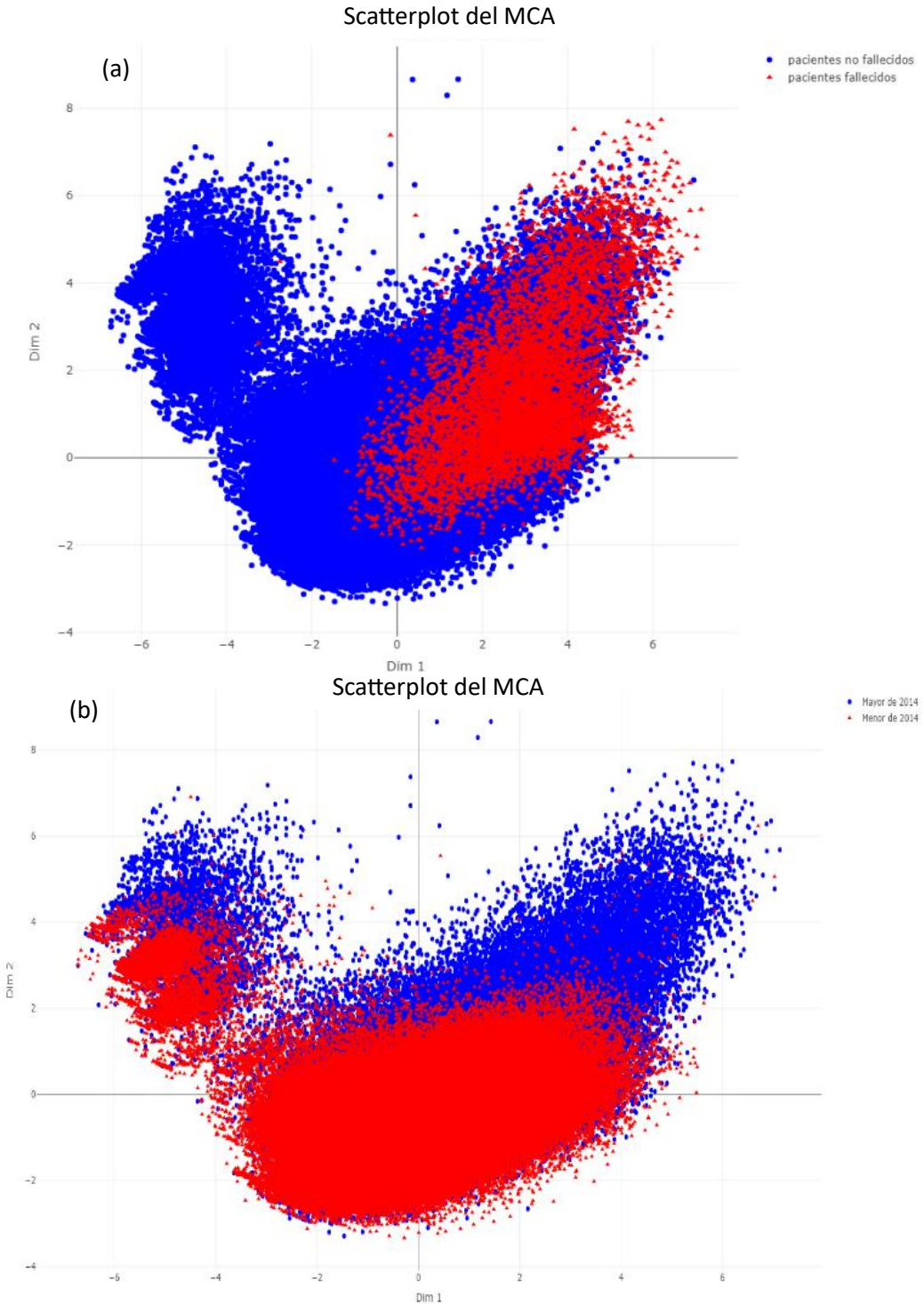


Figura 5.3: Grafico de dispersión del MCA en función de las clases (a) y de las admisiones anteriores y posteriores a 2014 (b)

La figura 5.4 muestra las cargas al cuadrado de cada variable tras realizar el MCA, lo cual nos puede dar pistas de las variables que deberemos de analizar en el análisis temporal univariante.

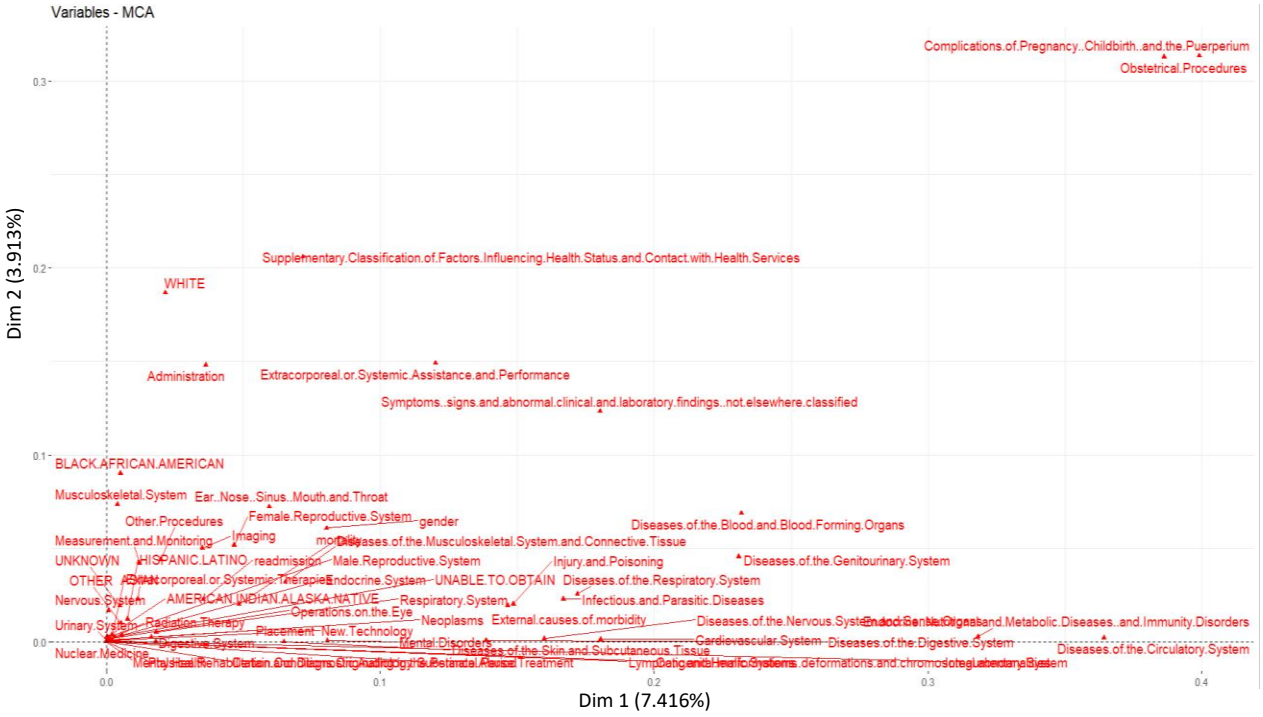


Figura 5.4: Gráfico de las cargas al cuadrado de cada variable tras realizar el MCA

En la figura 5.5 se muestra el DTH de la dimensión 1 correspondiente al conjunto de datos completo (a), las admisiones de pacientes no fallecidos (b) y las admisiones de pacientes fallecidos (c). Lo primero cdque se observa son los cambios bastante notables entre años consecutivos , lo cual está causado por las diferencias en el número de datos de cada año y la desidentificación de los datos originales y el preprocesado aplicado para obtener una fecha aproximada, lo cual supone un problema que se escapa al alcance del presente trabajo. Analizando la distribución de los tres gráficos a lo largo del tiempo, vemos que siguen una distribución similar, presentando datos bastante agrupados con poca presencia de *outliers*, en los casos (a) y (b) agrupados en torno al 0 y en el caso (c) entorno al 3, lo cual la se podía visualizar en la figura 5.3. También se observa una ligera tendencia en la distribución hacia valores más positivos con el paso del tiempo, tendencia la cual también se observaba en las figuras 5.3 (a), y 5.3 (b).

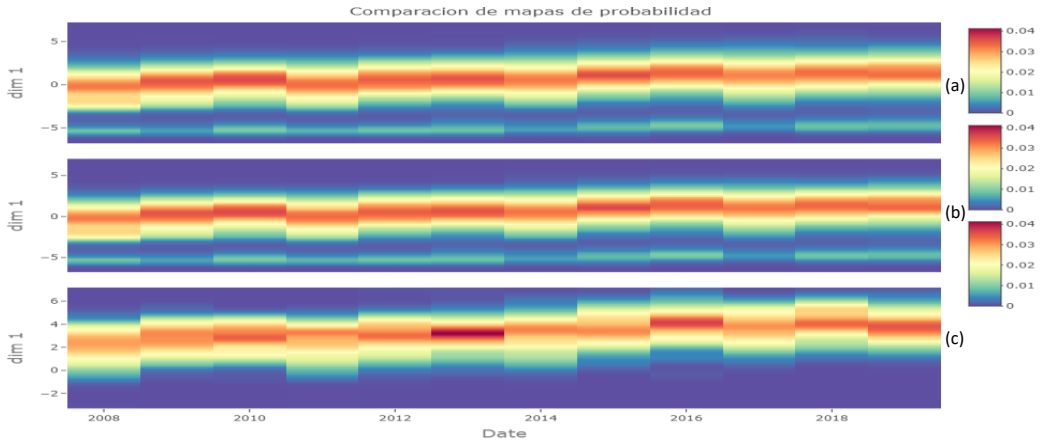


Figura 5.5: Mapa de calor temporal de la dimensión 1 del resultado del MCA en el conjunto total de datos (a), datos de pacientes no fallecidos (b) y datos de pacientes fallecidos (c)

En la figura 5.6 se puede observar el DTH de la misma forma que la figura 5.5 pero en este caso a lo largo de la dimensión 2 del MCA. En este caso se ve como hasta los años 2014-2015 la distribución de los datos es muy similar a la primera dimensión, con datos muy agrupados con una ligera tendencia hacia valores más positivos, pero a partir de estos años se produce una mayor dispersión de los datos, lo cual puede ser indicativo de que a partir de los años en los que se empiezan a introducir el protocolo de códigos ICD-10 se produce una mayor variabilidad dando lugar a cambios tanto en la distribución de los datos originales como en la distribución condicionada a cada clase (cabe destacar que aunque el protocolo de códigos ICD-10 comenzó en 2015, debido a la desidentificación que sufren los datos del repositorio MIMIC-IV la fecha es aproximada en rangos de 3 años, por lo que datos que realmente son de 2015 pueden presentarse como datos del 2014 o 2016, provocando que la variación en la distribución de probabilidad de los datos comience antes de 2015 pero esté relacionada con este suceso). En esta figura se puede apreciar la presencia de dos tipos de *dataset shifts* en los datos. En primer lugar, el cambio en la distribución de probabilidad condicionada a cada clase (figuras 5.6 (b) y 5.6 (c)) se identifica como un *concep shift*, y en segundo lugar, el cambio en la distribución de probabilidad de los datos de entrada del dataset completo (figura 5.6 (a)) se identifica como un *covariate shift*.

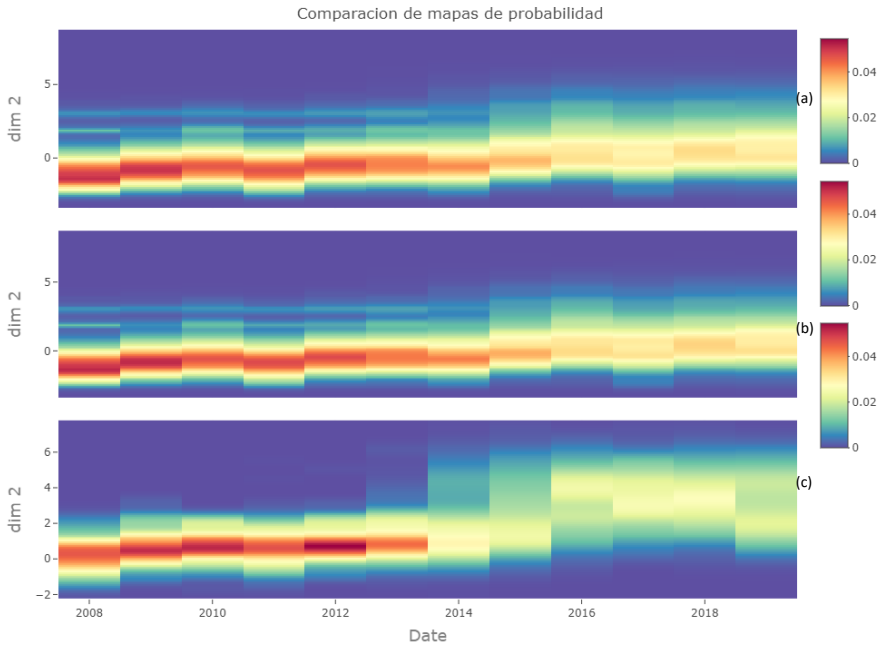


Figura 5.6: Mapa de calor temporal de la dimensión 2 del resultado del MCA en el conjunto total de datos (a), datos de pacientes no fallecidos (b) y datos de pacientes fallecidos (c)

Tras realizar la unión de los mapas de distribución de probabilidad de las dos dimensiones de los datos de pacientes fallecidos y no fallecidos (figura 5.7 (a)) y la unión de las dos primeras dimensiones del MCA del conjunto completo de datos (figura 5.7 (b)), se obtienen resultados similares a los de la figura 5.6, donde a partir de los años donde se empiezan a usar los códigos ICD-10 se produce una variación en la distribución de probabilidad de los datos, generando una mayor variabilidad con el paso del tiempo. En la figura 5.7 (a) se observa la presencia de cambios abruptos en la distribución de

probabilidad de los datos en 2014 y en 2016 desplazándola hacia arriba, para posteriormente tener una tendencia descendente de forma abrupta de nuevo con los datos distribuidos de forma muy dispersa, mostrando de nuevo al igual que en el caso anterior la presencia de un *concept shift*. En la figura 5.7 (b) a su vez, se identifica también un cambio abrupto en el año 2015 con una mayor dispersión de los datos y de nuevo una tendencia ascendente, permitiendo la identificación también en este caso de un *covariate shift*.

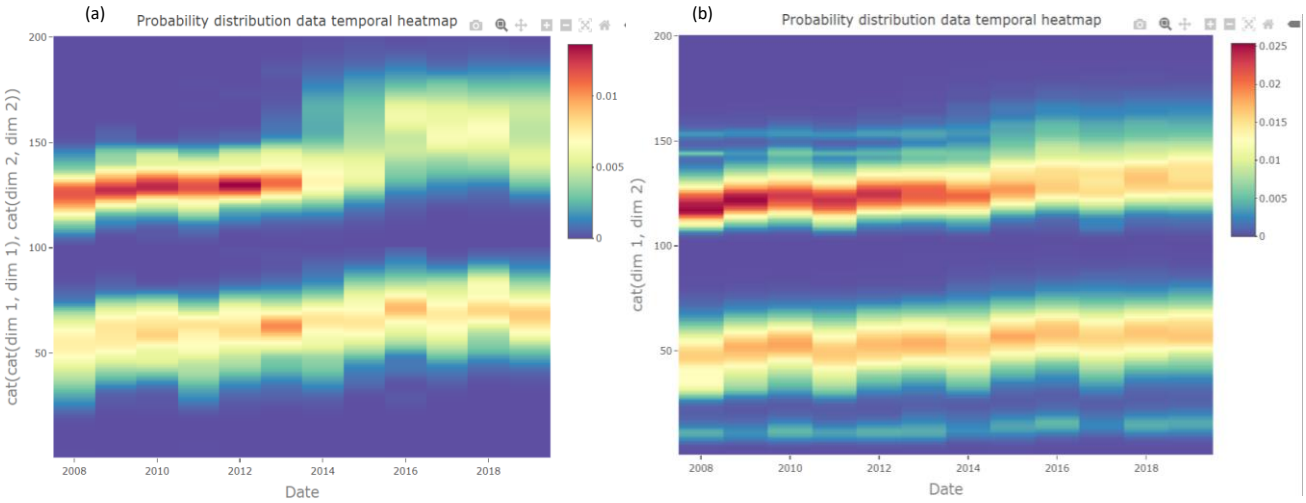


Figura 5.7: Mapa de calor temporal de la unión de las dos dimensiones resultado de la unión de los pacientes fallecidos y no fallecidos (a) y del dataset completo (b)

La obtención la IGT de los resultados obtenidos en la Figura 5.7 (Figura 5.8) muestran los cambios producidos en los datos en el tiempo descritos hasta el momento. Se observan dos agrupaciones claras de los años 2008 hasta el 2012 y del 2016 hasta el 2019 con unos años de transición, la cual es la tendencia que se ha descrito hasta el momento a partir de los DTH.

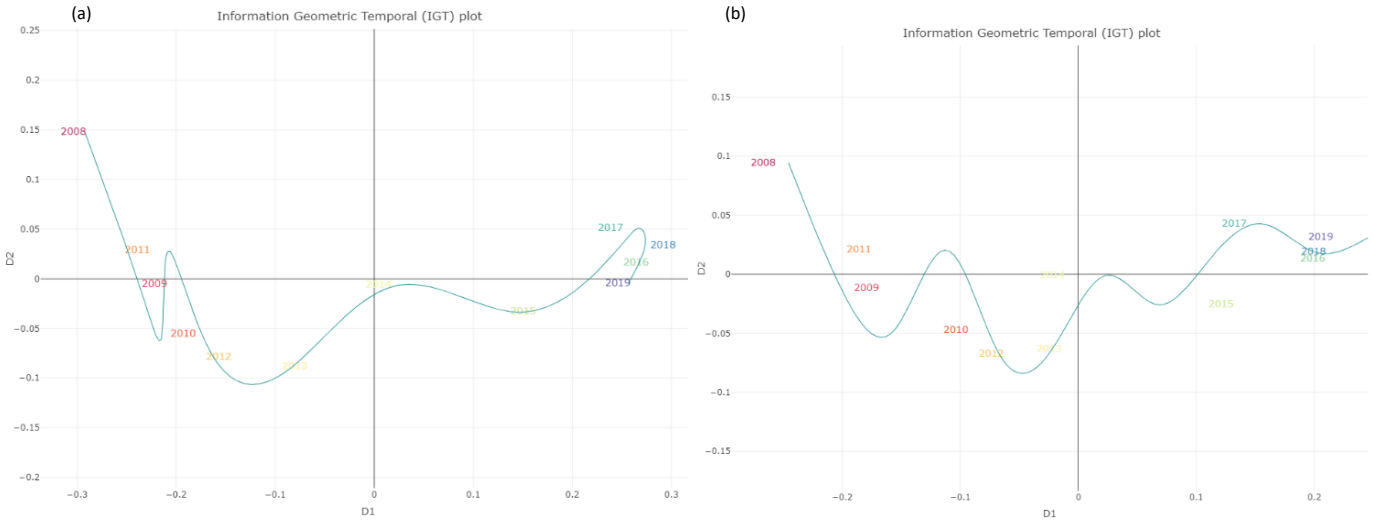


Figura 5.8: IGT del resultado de la unión de las dos dimensiones de los pacientes fallecidos y no fallecidos (a) y del dataset completo (b)

De esta forma queda justificada la presencia de variabilidad temporal en los datos, tanto por la presencia de *concept shifts* y *covariate shifts* descrita anteriormente como por la presencia de diferentes agrupaciones y la distancia entre los lotes temporales en la representación de la IGT. Con el objetivo de analizar más en detalle la naturaleza de esta variabilidad, a continuación, se mostrarán los resultados más relevantes obtenidos del análisis individual de cada variable del *datastet*.

5.2.2 Análisis univariante

Con el fin de obtener un análisis general de la distribución de probabilidad de cada capítulo de códigos ICD se ha obtenido la Figura 5.9, en la que se puede observar la variabilidad de cada capítulo a lo largo del tiempo.

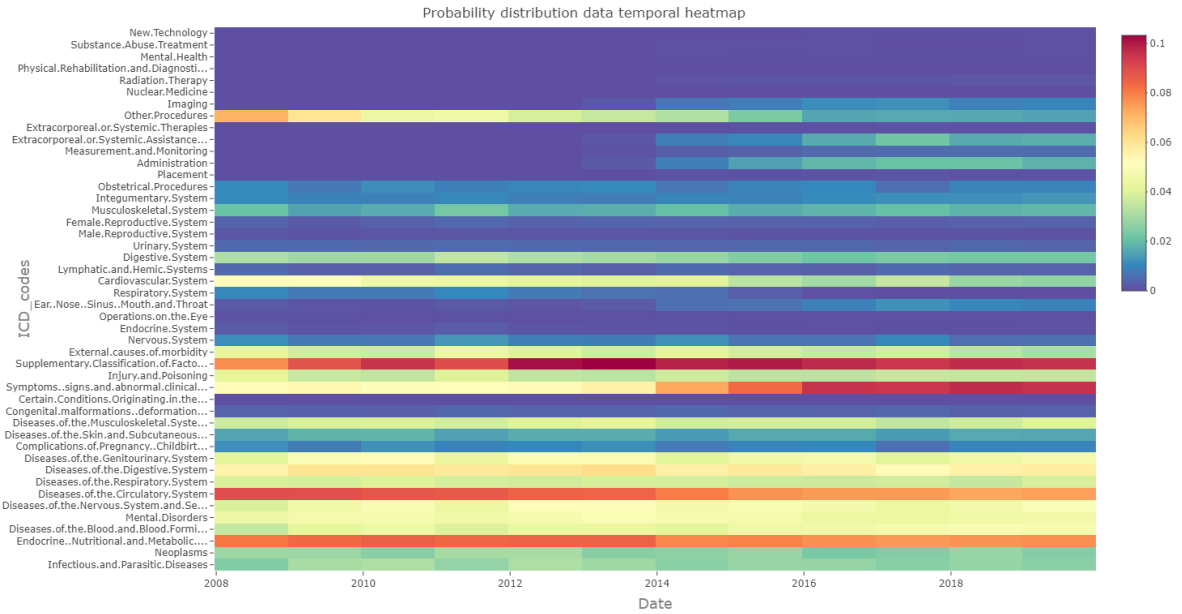


Figura 5.9: DTH de la distribución de la probabilidad de los capítulos de códigos ICD

De esta figura destacan varios aspectos importantes. En primer lugar, en el capítulo de procedimientos *Other procedures* se puede ver una disminución en su distribución a lo largo del tiempo, estando mucho más presente en los años 2008-2015 que en los años posteriores, teniendo su pico máximo en el año 2008. Algo similar sucede con el capítulo de procedimientos *Cardiovascular system*, disminuyendo su función de probabilidad a lo largo del tiempo. En la figura 5.2 se observa que estos dos capítulos son los que presentan una mayor frecuencia dentro de los capítulos de códigos correspondientes a procedimientos, superando ambos las 70,000 entradas, por los que los cambios en la variabilidad temporal de estos códigos van a suponer una elevada relevancia en la variabilidad temporal total del dataset. Los capítulos de diagnósticos *Diseases of the circulatory system* y *Endocrine nutritional and metabolic diseases and immunity disorders* también presentan una disminución en la distribución de probabilidad de sus datos más ligera a partir de los años 2014-2015, y como se puede observar en la figura 5.1 y al igual que sucedía en los capítulos de procedimientos, estos códigos tienen una frecuencia

elevada dentro del dataset, por lo que sus cambios pueden suponer un cambio notorio en la variabilidad del dataset. El hecho común que mantienen todos estos capítulos es que su distribución de probabilidad es mayor en los años en los que solamente se utilizaban códigos ICD-9, lo que sugiere que a partir de la instauración de los códigos ICD-10 algunos de los diagnósticos y procedimientos correspondientes a estos capítulos se han repartido en otros capítulos o que por factores desconocidos se haya disminuido la frecuencia de estos códigos. Lo primero, es lo que puede haber sucedido con el capítulo de diagnósticos *Symptoms signs and abnormal clinical and laboratory findings not elsewhere classified*, donde se observa que a partir del año 2014 se produce un gran aumento en su distribución de probabilidad, siendo este capítulo también uno de los que más frecuencia presenta. En el caso de los capítulos de procedimientos *Imaging, Extracorporeal or systemic assistance and performance y administration*, también se observa un aumento en la distribución de probabilidad de los datos a partir de 2014-2015, aunque en este caso, este hecho era esperable, ya que tal y como se puede observar en las tablas 4.2 y 4.3, la aparición de los códigos ICD 10 da lugar a capítulos los cuales no tienen correspondencia directa con capítulos de códigos ICD 9, como son estos casos, lo cual puede provocar el reparto de códigos ICD 9 en otros capítulos ICD 10 y los consecuentes cambios en las distribuciones de probabilidad mencionados previamente.

Esta variación de las distribuciones de probabilidad de los códigos son indicativos de nuevo de la presencia de nuevo de un *covariate* shift en el repositorio.

Si se representa la IGT de la información anterior (Figura 5.10), se obtiene un resultado muy similar a la Figura 5.8, con dos grupos bien diferenciados y unos años de transición. Resultado el cual coincide con las conclusiones sacadas del análisis del DTH de la Figura 5.9.

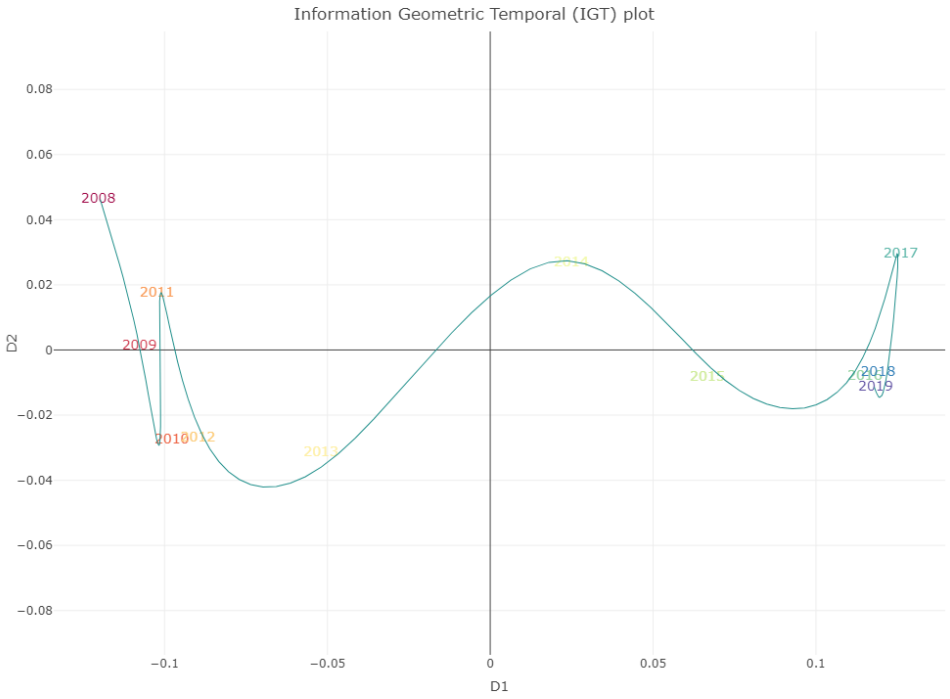


Figura 5.10: Representación de la IGT de los capítulos de los códigos ICD

Con el objetivo de estudiar más en detalle esta variabilidad, se ha analizado la variación de la distribución de probabilidad de los códigos asociada a cada clase (figura 5.11), la diferencia entre la distribución de probabilidad de cada clase con el dataset completo (figura 5.12) y la IGT correspondiente a cada clase (figura 5.13)

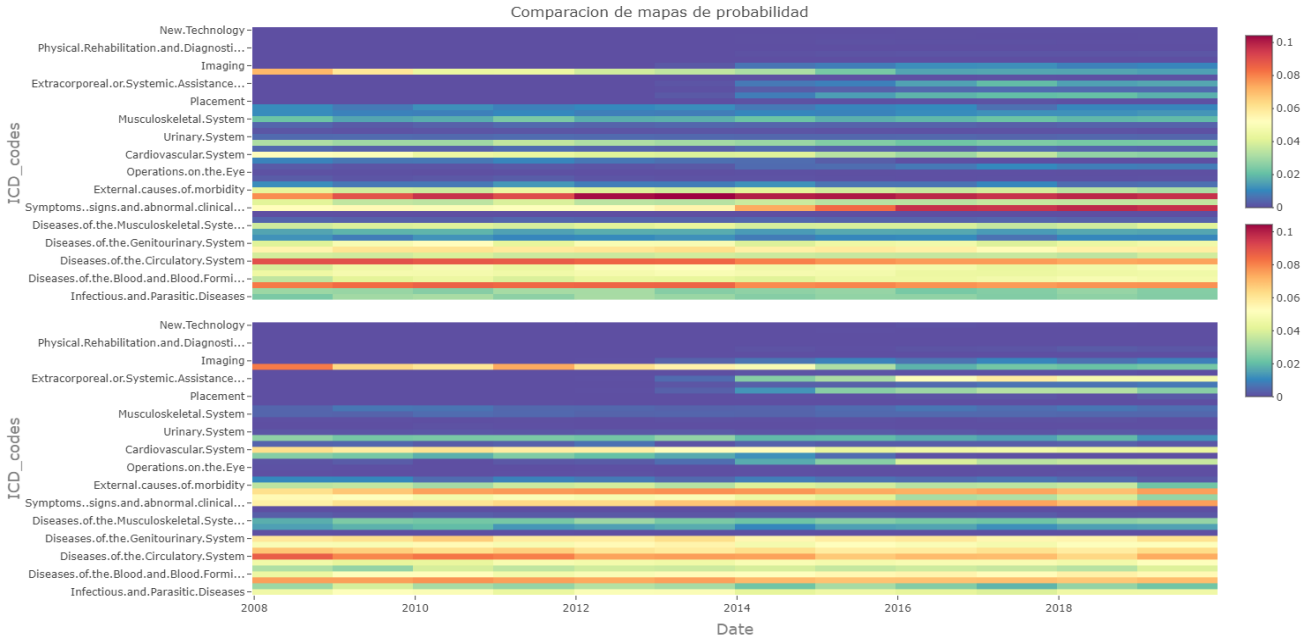


Figura 5.11: Comparación de la variación en la distribución de probabilidad de la clase negativa (gráfica superior) con la clase positiva (gráfica inferior), donde los códigos siguen el mismo orden entre ellas y con la figura 5.9

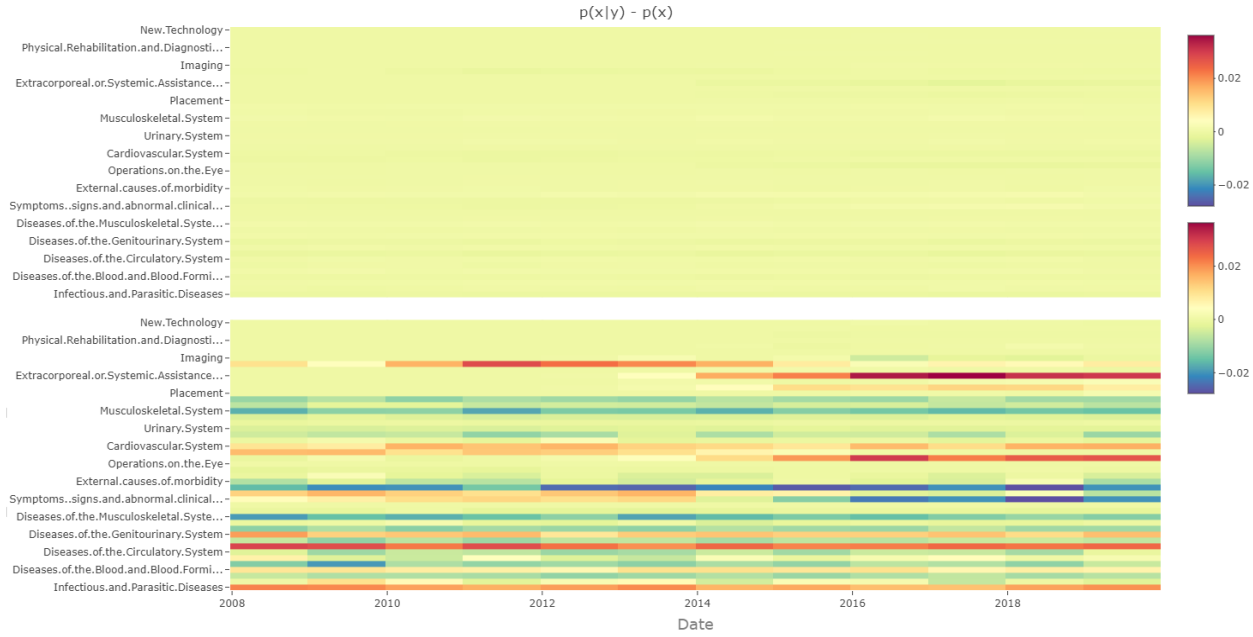


Figura 5.12: Comparación de la resta de la distribución de probabilidad condicionada a cada clase menos la distribución de probabilidad del dataset completo, siendo la gráfica superior la correspondiente a la clase negativa y la gráfica inferior la correspondiente a la clase positiva

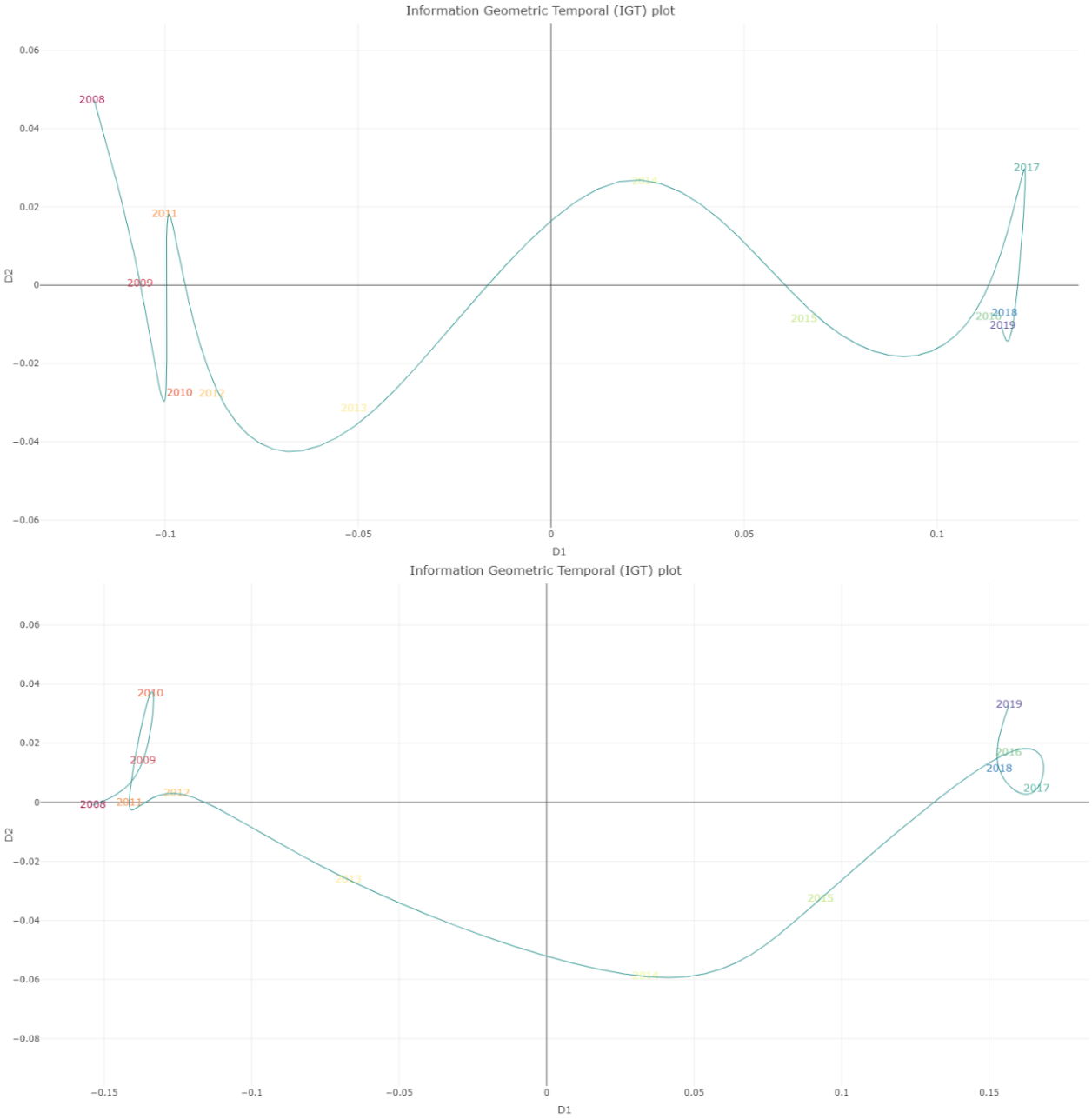


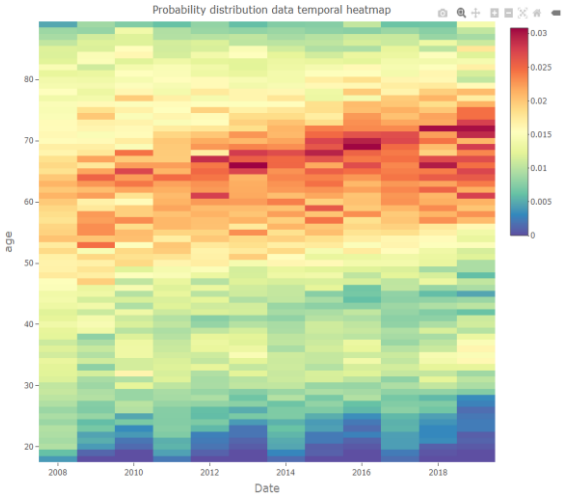
Figura 5.13: Representación de la IGT de los códigos correspondientes a la clase negativa (superior) y a la clase positiva (inferior)

En la figura 5.11 se puede observar cómo los códigos que varían su distribución de probabilidad a lo largo del tiempo son prácticamente los mismos en ambas clases, presentando las principales diferencias en los capítulos *Symptoms signs and abnormal clinical and laboratory findings not elsewhere classified*, *Diseases of the circulatory system* y *Endocrine nutritional and metabolic diseases and immunity disorders*. La presencia de variaciones en la distribución de probabilidad condicionada a

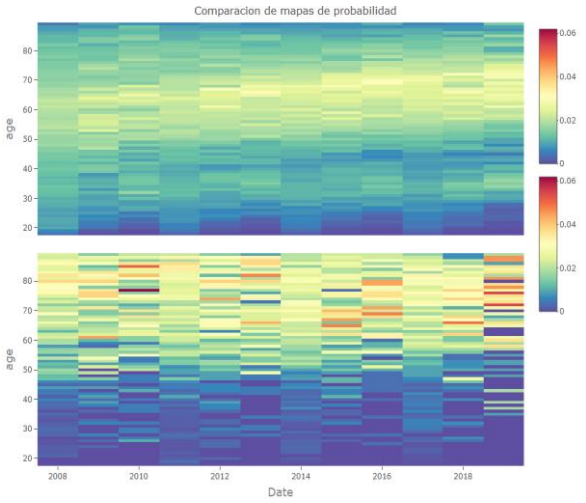
cada clase vuelve mostrar la presencia de *concept shift*. A excepción de estos 3 capítulos, las variaciones en la distribución de probabilidad de los capítulos de ambas clases son muy similares, variando principalmente el valor de la probabilidad de cada clase. Si analizamos como afecta cada capítulo a la variabilidad temporal de cada clase (figura 5.12), se observa que para la clase negativa la influencia de cada capítulo sobre la variabilidad total se distribuye de forma homogénea, mientras que para la clase positiva hay variables que tienen una gran influencia a lo largo de todo el periodo de tiempo, como los capítulos *Diseases of the respiratory system* o *Infectious and parasitic diseases*, y otras variables cuya influencia aumenta con el tiempo, como el caso de *Ear, nose, sinus, mouth and trhoat* o *Extracorporeal or Systemic Assistance and Performance*.

La figura 5.13 vuelve a mostrar la presencia de dos grupos formados por los lotes temporales pre y post-implantación de códigos ICD-10 en ambas clases, tal y comosuceía en los resultados obtenidos hasta el momento.

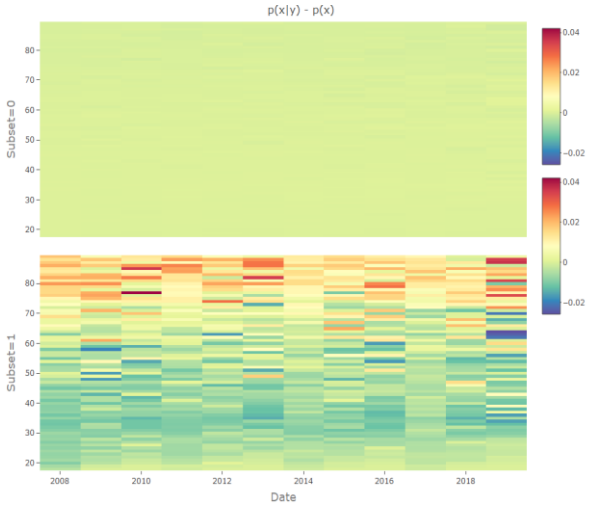
La Figura 5.14 muestra los resultados del análisis temporal de la variable *age*. En ellos se puede observar que la edad de los pacientes admitidos en el hospital tiene una tendencia ascendente con el paso de los años (a), tendencia la cual se observa de forma más clara sobre todo en los pacientes no fallecidos que en los fallecidos, cuya distribución es mucho más heterogénea (b), pudiendo observar también que la edad media de los pacientes fallecidos es mayor que la de los pacientes no fallecidos (b). Se puede concluir también que la edad superior a los 75 años es un factor relevante en los pacientes fallecidos (c), y por último que a pesar de presentar esta variable un ligero cambio en su distribución de probabilidad hacia valores mayores, la IGT no muestra una tendencia clara con el paso de los años, ni la presencia de lotes temporales bien diferenciados, como podíamos observar en resultados anteriores (d).



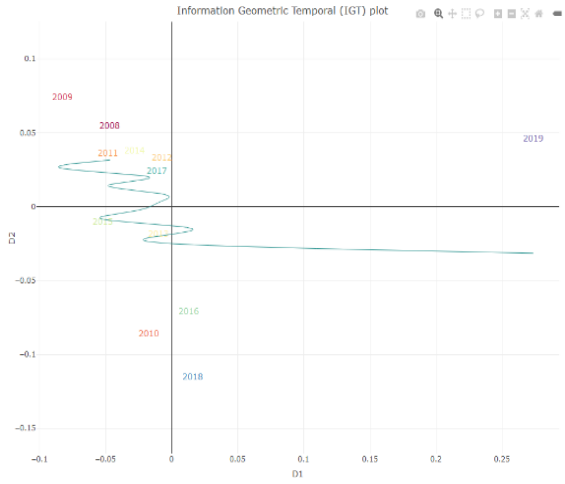
(a) DTH variable *age*



(b) Comparación DTH de pacientes no fallecidos (superior) y fallecidos (inferior)



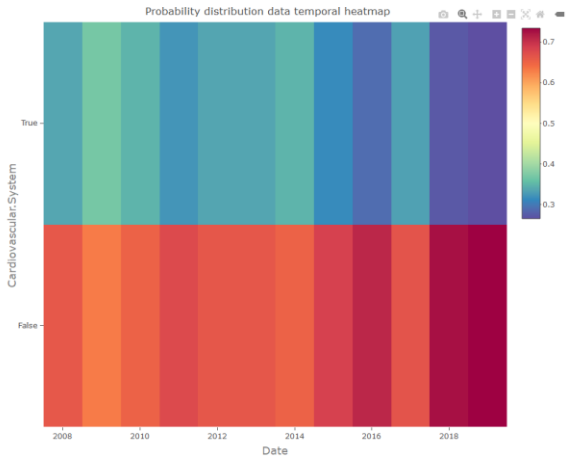
(c) DTH de la distribución de probabilidad condicionada menos la distribución de probabilidad del dataset de la clase negativa (superior) y la clase positiva (inferior)



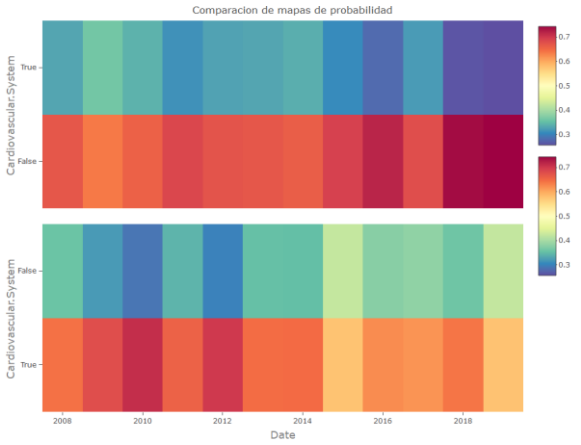
(d) IGT de la unión de las distribuciones de probabilidad condicionadas de cada clase de la variable *age*

Figura 5.14: Análisis temporal de la variable *age*

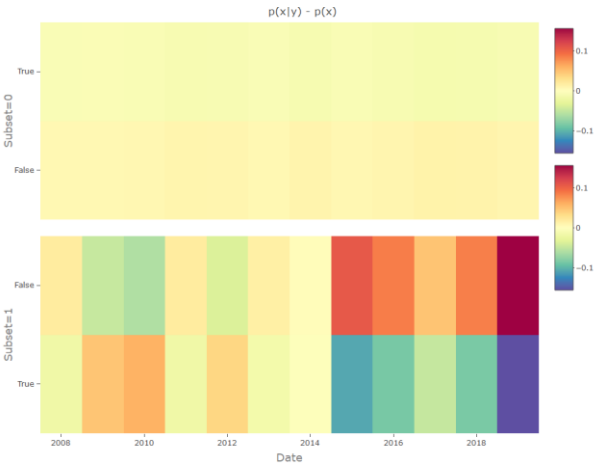
Tras analizar más detenidamente la variable *Cardiovascular system* (Figura 5.15), se observa una ligera tendencia a disminuir el número de casos asociados con estos códigos con el paso del tiempo a partir de los años de aparición de los códigos ICD 10 (a), tendencia la cual se aprecia en las dos clases. También se observa como estos procedimientos están muy presentes en los pacientes fallecidos pero muy poco presentes en pacientes no fallecidos (b). Se observa un cambio en la distribución de probabilidad de la clase positiva hacia valores más negativos debido a la disminución de la frecuencia con el paso del tiempo, disminuyendo así la influencia que tiene la presencia de estos códigos sobre los pacientes fallecidos con el tiempo, mientras que la distribución de la clase negativa se mantiene constante (c). Por último, la IGT muestra una tendencia de los datos hacia valores positivos a lo largo de la dimensión 1, mostrando una mayor diferenciación que el caso anterior, pero siguen sin aparecer agrupaciones claras de lotes temporales (d).



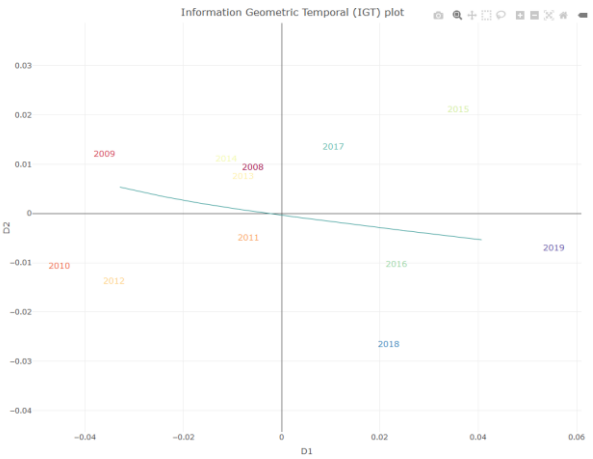
(a) DTH variable *Cardiovascular system*



(b) Comparación DTH de pacientes no fallecidos (superior) y fallecidos (inferior)



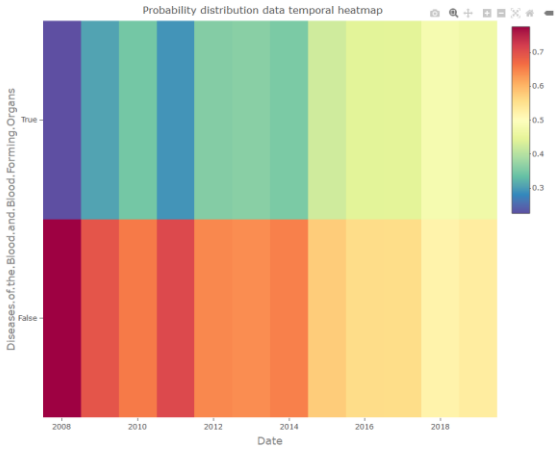
(c) DTH de la distribución de probabilidad condicionada menos la distribución de probabilidad del dataset de la clase negativa (superior) y la clase positiva



(d) IGT de la unión de las distribuciones de probabilidad condicionadas de cada clase de la variable *Cardiovascular system*

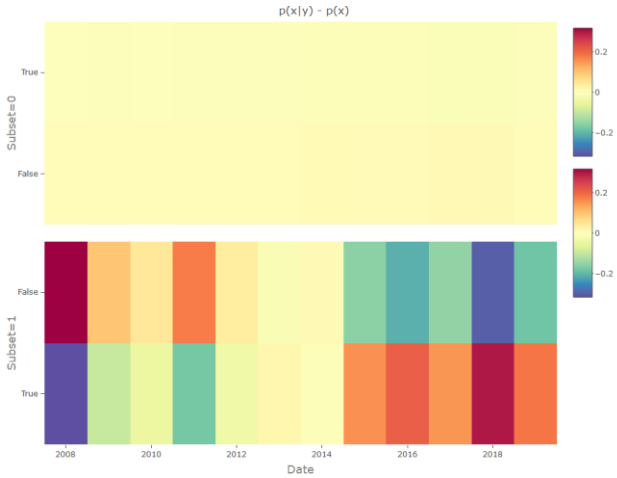
Figura 5.15: Análisis temporal de la variable *Cardiovascular system*

La Figura 5.16 muestra el análisis temporal de la variable *Diseases of the blood and blood forming organs*. En primer lugar, se observa un aumento de la frecuencia de aparición de estos códigos con el paso del tiempo, sobre todo a partir del año 2015 (a). El análisis de la distribución condicionada (b), muestra un aumento en las dos clases, tanto la positiva como en la negativa, y también muestra que la frecuencia de aparición de estos códigos de diagnóstico es mayor en los pacientes fallecidos que en los no fallecidos. En el caso de cómo afecta la variable a la distribución de probabilidad de cada clase (c), vemos como la distribución de probabilidad de esta variable a lo largo de la clase negativa se mantiene constante, mientras que para la clase positiva va aumentando con el tiempo, sobre todo a partir de 2015, llegando prácticamente a invertirse.

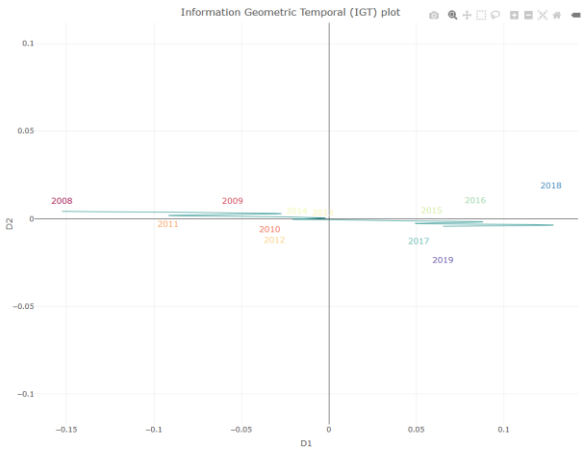


(a) DTH variable *Diseases of the blood and blood forming organs*

(b) Comparación DTH de pacientes no fallecidos (superior) y fallecidos (inferior)



(c) DTH de la distribución de probabilidad condicionada menos la distribución de probabilidad del dataset de la clase negativa (superior) y la clase positiva

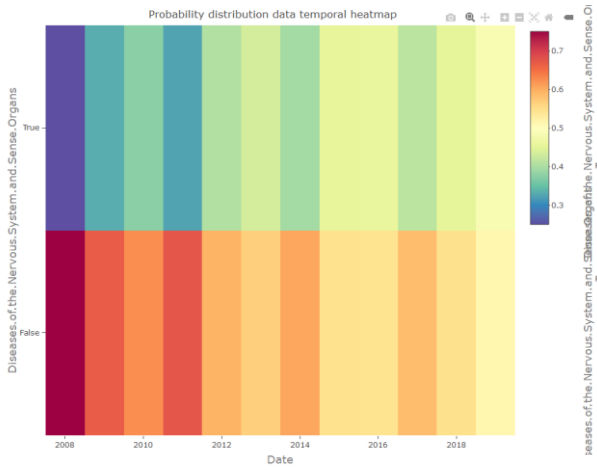


(d) IGT de la unión de las distribuciones de probabilidad condicionadas de cada clase de la variable *Diseases of the blood and blood forming organs*

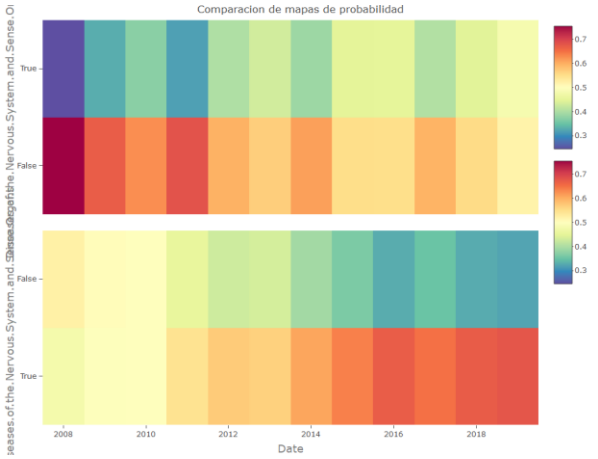
Figura 5.16: Análisis temporal de la variable *Diseases of the blood and blood forming organs*

En la Figura 5.17 se muestran los resultados del análisis del capítulo de diagnósticos *Diseases of the nervous system and sense organs*, el cual tiene una interpretación similar al caso anterior. Se observa

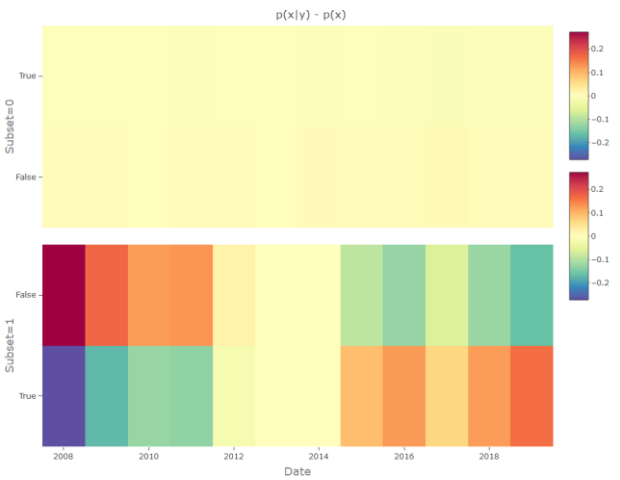
un ligero aumento en la aparición de estos códigos con el paso del tiempo (a), aumento el cual aparece tanto en pacientes fallecidos como en pacientes no fallecidos, aunque la frecuencia de aparición es mayor en los pacientes fallecidos (c). En el caso de la distribución de probabilidad de cada clase se observa como la distribución de la clase negativa se mantiene constante, mientras que la distribución de probabilidad de aparición del capítulo en la clase positiva aumenta con el paso del tiempo, empezando en valores negativos y acabando con valores positivos tras haberse producido prácticamente una inversión en la distribución, lo que indica el aumento en la relevancia de la presencia de estos códigos en pacientes que han fallecido a medida que va pasando el tiempo (c). Por último, la IGT también ratifica estos resultados, mostrando una evolución de los lotes temporales a lo largo de la primera dimensión, formando dos grupos más o menos diferenciados (d).



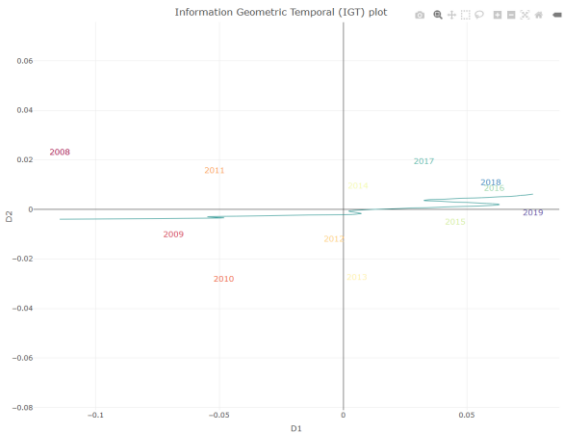
(a) DTH variable *Diseases of the nervous system and sense organs*



(b) Comparación DTH de pacientes no fallecidos (superior) y fallecidos (inferior)



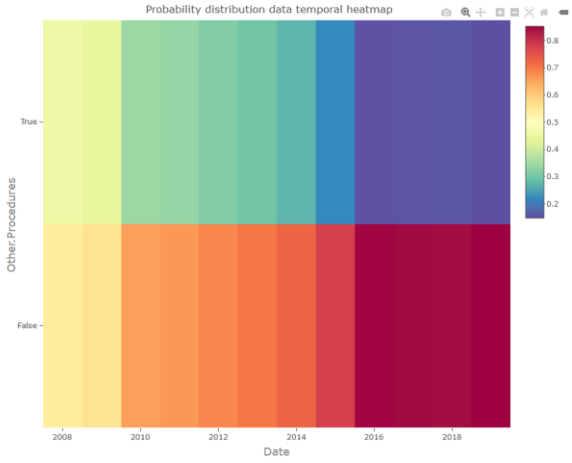
(d) DTH de la distribución de probabilidad condicionada menos la distribución de probabilidad del dataset de la clase negativa (superior) y la clase positiva



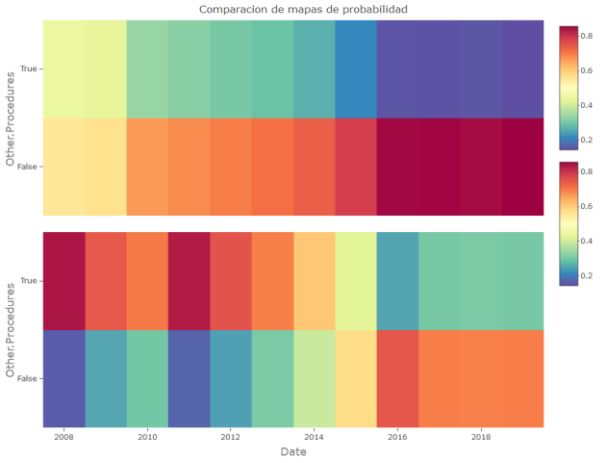
(d) IGT de la unión de las distribuciones de probabilidad condicionadas de cada clase de la variable *Diseases of the nervous system and sense organs*

Figura 5.17: Análisis temporal de la variable *Diseases of the nervous system and sense organs*

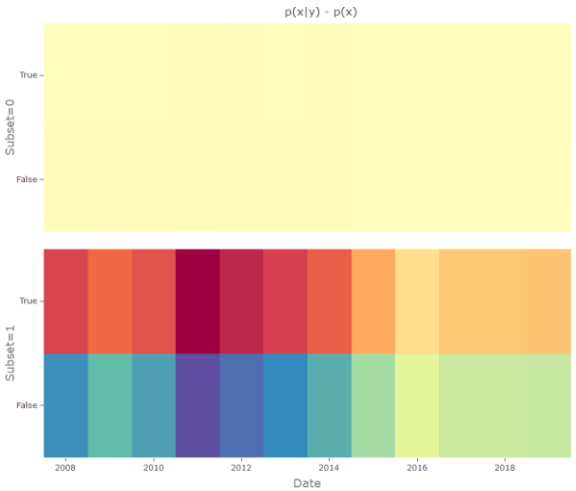
En la Figura 5.18, se observan los resultados de la variable *Other procedures* comentados previamente de una forma más detallada. En primer lugar, se observa una disminución en la distribución de la probabilidad de la variable con el paso del tiempo tanto de forma general en el dataset (a), como en cada clase (b), además se aprecia que la probabilidad de aparición de este tipo de códigos en pacientes fallecidos es mayor que en pacientes no fallecidos. Además, la distribución de la clase positiva va variando hacia valores menores con el paso del tiempo (c). Por último, la representación de la IGT muestra que hay claramente dos grupos bien diferenciados formados por los lotes temporales correspondientes a los años con y sin códigos ICD 10.



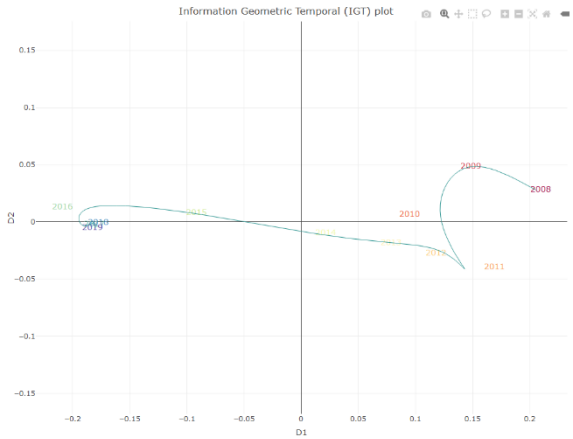
(a) DTH variable *Other procedures*



(b) Comparación DTH de pacientes no fallecidos (superior) y fallecidos (inferior)



(c) DTH de la distribución de probabilidad condicionada menos la distribución de probabilidad del dataset de la clase negativa (superior) y la clase positiva

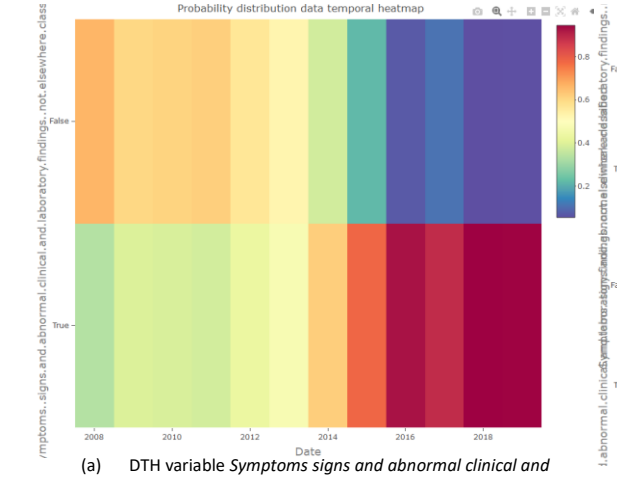


(d) IGT de la unión de las distribuciones de probabilidad condicionadas de cada clase de la variable *Other procedures*

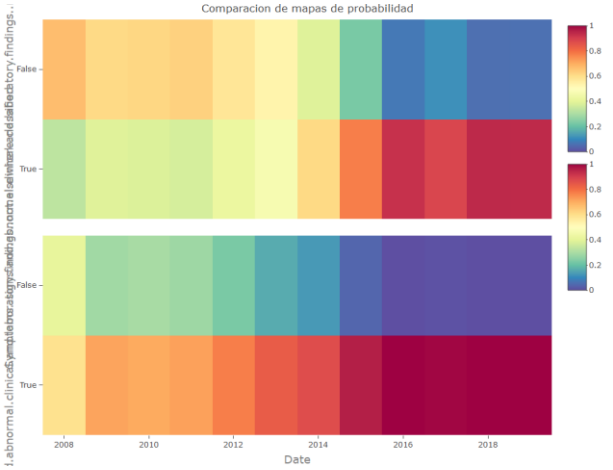
Figura 5.18: Análisis temporal de la variable *Other procedures*

En el caso de la variable *Symptoms signs and abnormal clinical and laboratory findings not elsewhere classified* (Figura 5.19) tal y como se había estudiado en la Figura 5.9, se observa cambio abrupto en la distribución de probabilidad en el año 2015, tanto en el dataset completo (a), como en el análisis de

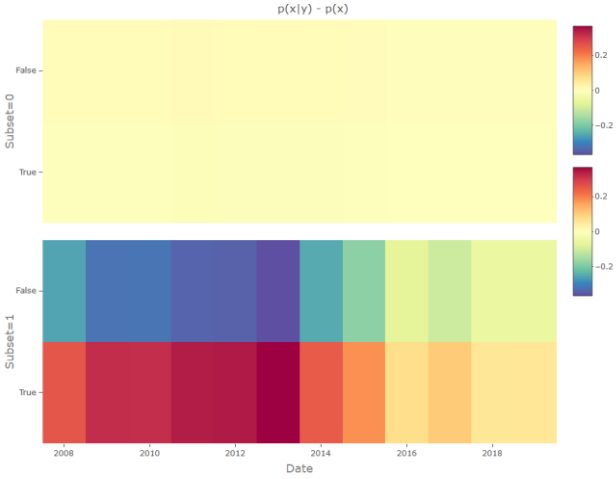
las clases por separado (b). Se observa como después del cambio en la distribución, este capítulo de diagnósticos está muy presente tanto en pacientes fallecidos como en pacientes no fallecidos, pero antes de producirse la probabilidad de que los pacientes fallecidos fuesen diagnosticados con códigos de este capítulo era mayor . Sin embargo, el análisis de la distribución de la clase positiva indica una variabilidad tendiendo hacia valores inferiores a partir de 2015, lo cual puede estar provocado por el aumento de la distribución de probabilidad que se ha llevado a cabo en otras variables, como las algunas de las descritas anteriormente, a partir de la introducción de los códigos ICD10(c). Esto da lugar a que la representación de la IGT muestre un desplazamiento de los datos a lo largo de la dimensión 1 pudiendo identificar en este caso también dos grupos bien diferenciados correspondientes a los lotes temporales pre y post códigos ICD 10 (d).



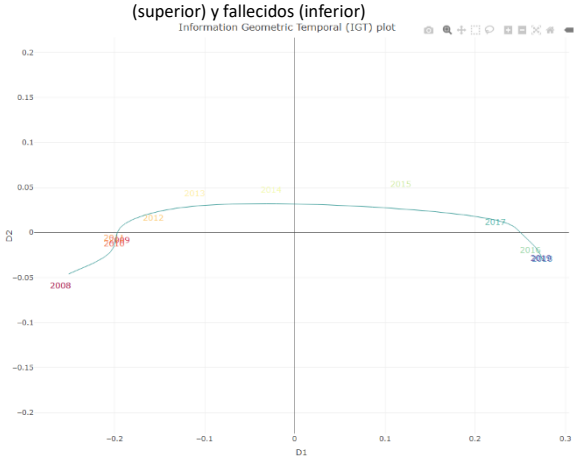
(a) DTH variable *Symptoms signs and abnormal clinical and laboratory findings not elsewhere classified*



(b) Comparación DTH de pacientes no fallecidos (superior) y fallecidos (inferior)



(c) DTH de la distribución de probabilidad condicionada menos la distribución de probabilidad del dataset de la clase negativa (superior) y la clase positiva



(d) IGT de la unión de las distribuciones de probabilidad condicionadas de cada clase de la variable *Symptoms signs and abnormal clinical and laboratory findings not elsewhere classified*

Figura 5.19: Análisis temporal de la variable *Symptoms signs and abnormal clinical and laboratory findings not elsewhere classified*

En última instancia, el estudio de la variabilidad temporal de la variable *mortality* (Figura 5.20), la cual se usa como etiqueta en los modelos de clasificación tal y como se indica en el apartado 3.1, muestra que la distribución de probabilidad de esta variable es constante a lo largo del tiempo, manteniendo una frecuencia muy superior para pacientes que no mueren a lo largo de todos los años, por lo que se puede concluir que no existe la presencia de *prior probability shift*.

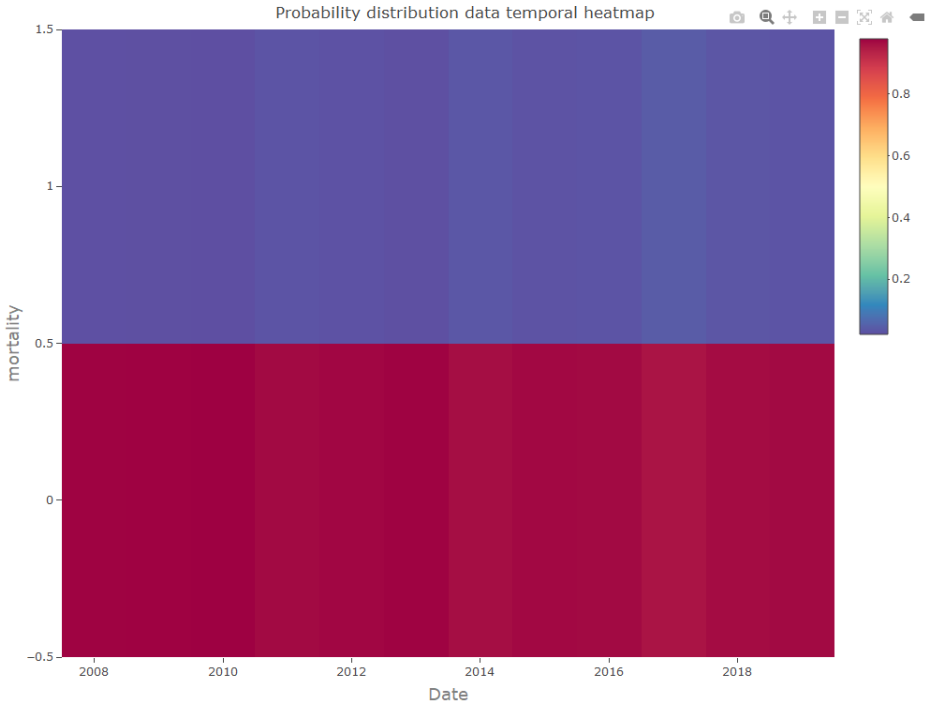


Figura 5.20: Análisis temporal de la distribución de probabilidad de la variable *mortality* utilizada como etiqueta en los modelos predictivos

Tras la realización del análisis temporal tanto multivariante como univariante, ha quedado más que demostrada la presencia de variabilidad temporal de distintos tipos en el dataset.

5.3 Resultados del modelado

Como se ha expuesto en el apartado 4.3, se han llevado a cabo dos modelos de machine learning con el objetivo de llevar a cabo la predicción de la mortalidad. En el siguiente apartado se presentan los resultados obtenidos de la evaluación de los modelos.

5.3.1 Random Forest

En la Figura 5.21 se muestran los resultados del área bajo la curva ROC media de todos los lotes temporales para cada combinación de hiperparámetros establecida en la tabla 4.4 tras realizar el entrenamiento del modelo de *Random Forest* evaluado con los datos de validación.

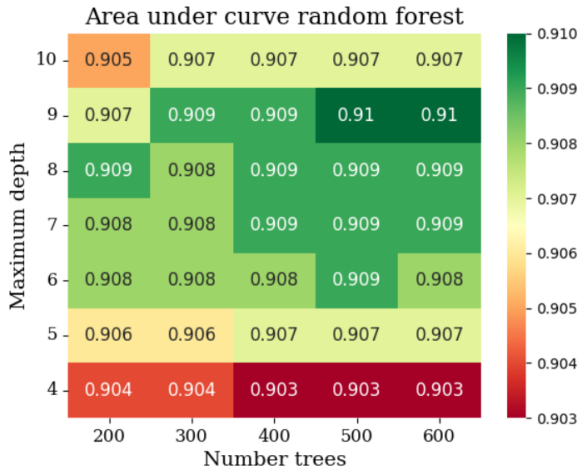
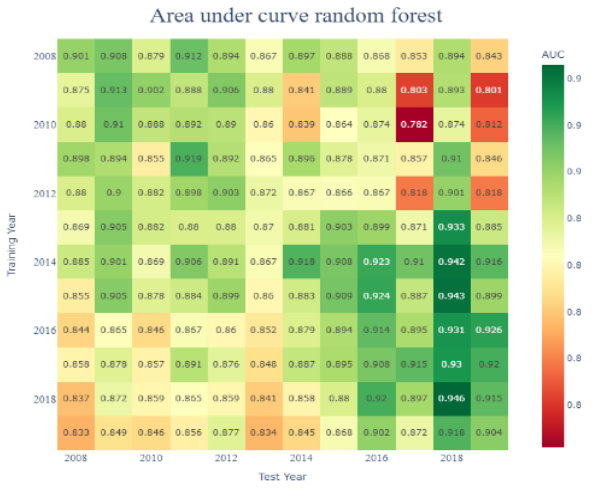


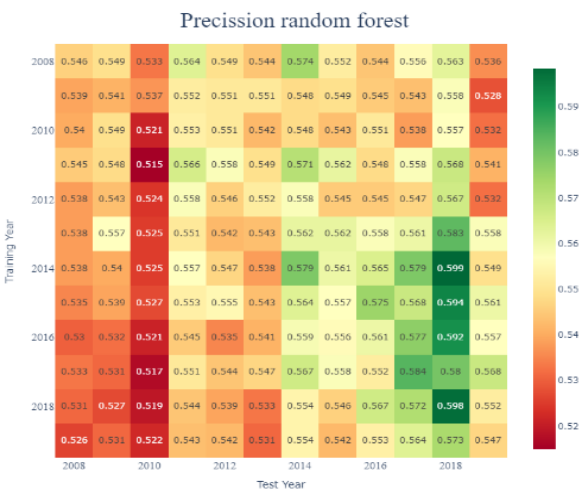
Figura 5.21: Área bajo la curva (AUC) media obtenida con las diferentes configuraciones de hiperparámetros, para el modelo *Random Forest*, utilizando los datos de validación para validar el modelo explicados en el apartado 4.3, donde para cada combinación de hiperparámetros se ha obtenido el área bajo la curva ROC de la validación de los modelos con los 12 lotes temporales y se ha calculado la media.

Los mejores resultados fueron los obtenidos con un número de árboles igual a 500 y una profundidad máxima de 9, y los modelos de todos los lotes temporales entrenados con estos hiperparámetros fueron los que se utilizaron para validarlos con los datos de test de cada lote temporal. Los resultados obtenidos de las métricas descritas en el apartado 4.3 de esta nueva validación se muestran en la Figura 5. 22.

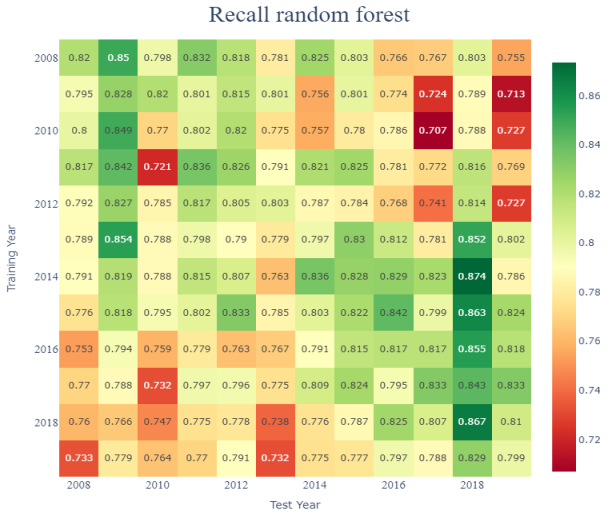
En todas las métricas a excepción de en la precisión (b), se observa la misma tendencia a obtener resultados mejores en los cuadrantes izquierdo superior y derecho inferior, siendo ligeramente mejores los resultados en estos últimos. Estos cuadrantes se corresponden con los modelos que han sido entrenados y validados con lotes temporales pre-instauración de códigos ICD 10 y post-instauración de códigos ICD 10 respectivamente. En el caso de la precisión, la presencia de los resultados obtenidos puede estar justificada debido al gran desbalanceo entre las clases positiva y negativa del que se ha hablado previamente. Este hecho provoca que al dividir el dataset en lotes temporales, la presencia de pacientes fallecidos sea muy pequeña en comparación con la de pacientes no fallecidos, provocando que la precisión de la predicción baje. Aún así, se han obtenido resultados ligeramente mejores para los modelos entrenados y validados a partir de 2015.



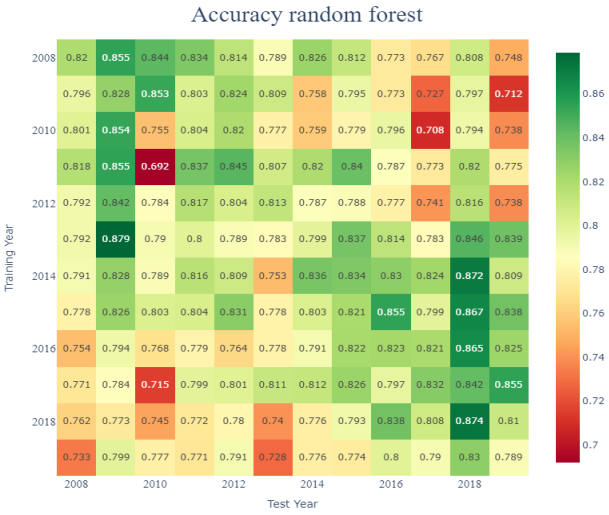
(a) Área bajo la curva ROC del resultado de la predicción



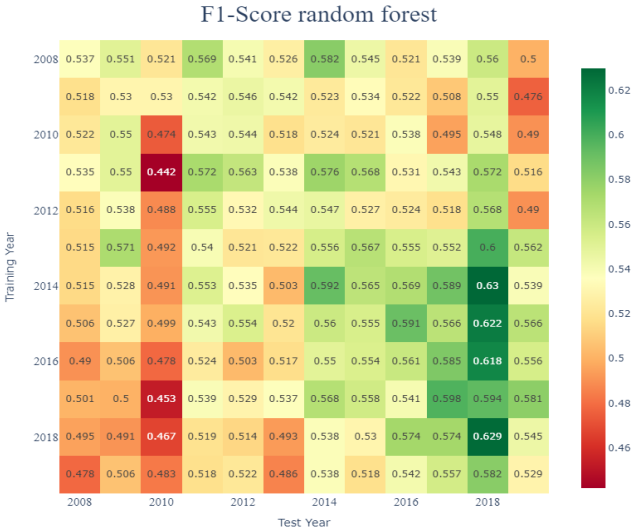
(b) Precisión del resultado de la predicción



(c) Sensibilidad del resultado de la predicción



(d) Exactitud del resultado de la predicción



(e) F1-Score del resultado de la predicción

Figura 5.22: Resultados de las métricas de la predicción de mortalidad con modelo *Random Forest*, considerando la configuración de hiperparámetros *óptima*, en el conjunto de test.

5.3.2 Gradient Boosting

En la Figura 5.23 se muestran los resultados del área bajo la curva ROC media de todos los lotes temporales para cada combinación de hiperparámetros establecida en la tabla 4.4 tras realizar el entrenamiento del modelo de *Gradient Boosting* evaluado con los datos de validación.

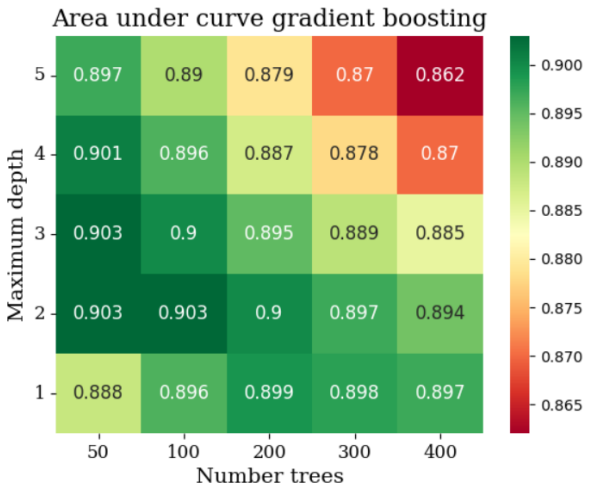
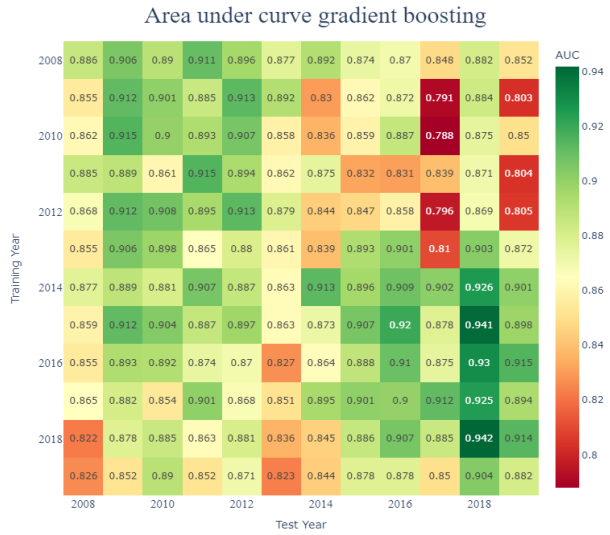


Figura 5.23: Área bajo la curva (AUC) media obtenida con las diferentes configuraciones de hiperparámetros, para el modelo *Gradient Boosting*, utilizando los datos de validación para validar el modelo explicados en el apartado 4.3, donde para cada combinación de hiperparámetros se ha obtenido el área bajo la curva ROC de la validación de los modelos con los 12 lotes temporales y se ha calculado la media.

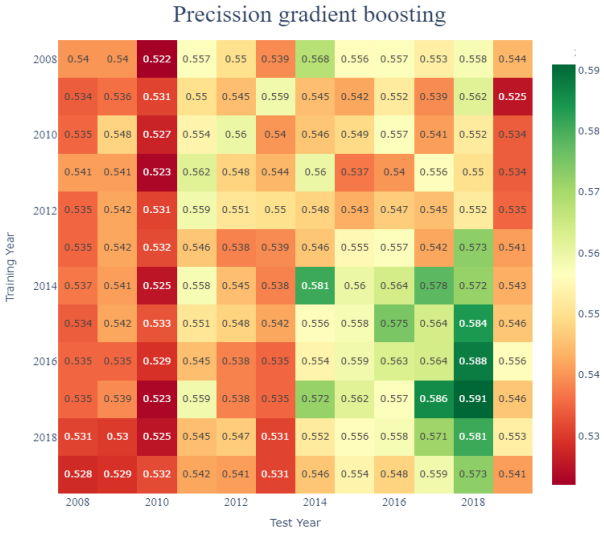
En este caso, se utilizaron los modelos entrenados con un número de árboles igual a 100 y una profundidad máxima de 2. Los resultados obtenidos de las métricas descritas en el apartado 4.3 de esta nueva validación se muestran en la Figura 5.24

De forma similar a como ocurría con el modelo anterior, todas las métricas a excepción de en la precisión (b) presentan una tendencia a obtener resultados mejores en las zonas entrenadas y validadas con los datos pre-códigos ICD 10 y post-códigos ICD 10 por separado, y presentando el mismo problema con la métrica de la precisión.

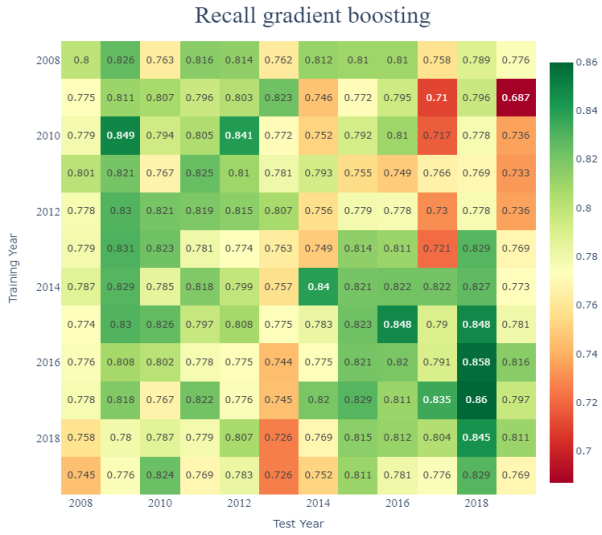
Ambos modelos han presentado un comportamiento y unos valores similares, siendo ligeramente mejor el rendimiento en este caso el modelo *Random Forest*.



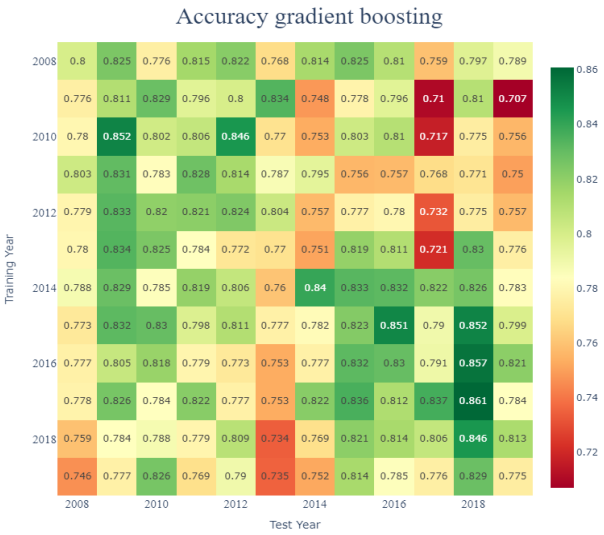
(a) Área bajo la curva ROC del resultado de la predicción



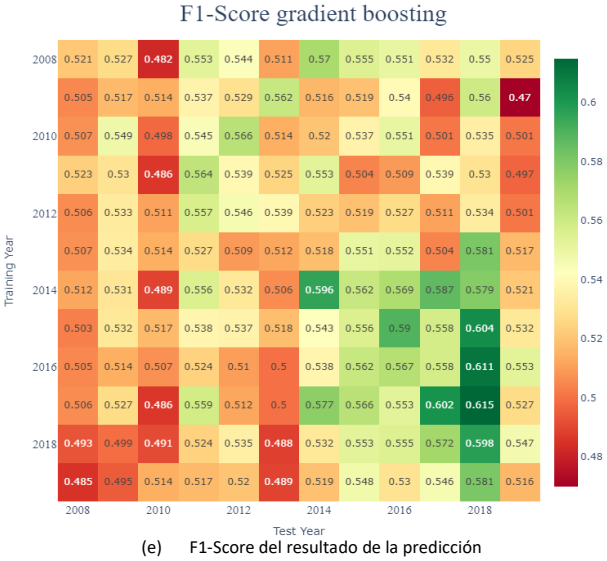
(b) Precisión del resultado de la predicción



(c) Sensibilidad del resultado de la predicción



(d) Exactitud del resultado de la predicción

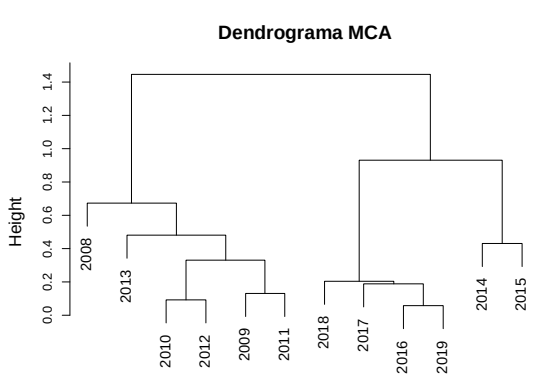


(e) F1-Score del resultado de la predicción

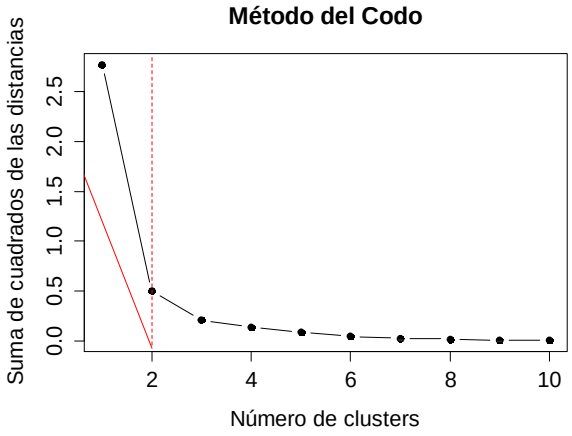
Figura 5.24: Resultados de las métricas de la predicción de mortalidad con modelo *Gradient Boosting*, considerando la configuración de hiperparámetros *óptima*, en el conjunto de test.

5.4 Evaluación de la relación variabilidad temporal-modelado

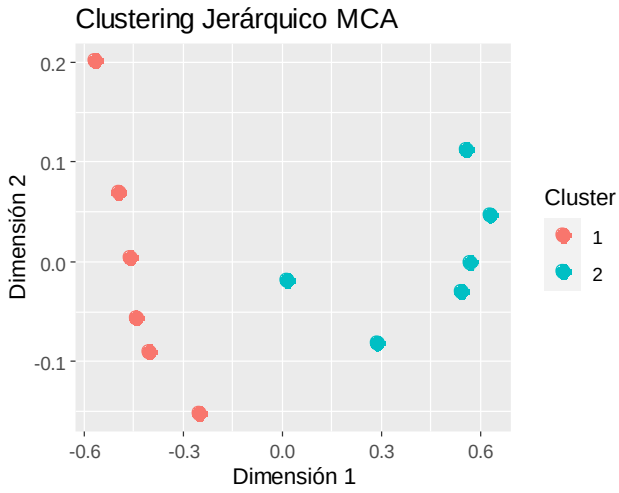
La Figura 5.25 muestra el resultado de aplicar el agrupamiento jerárquico aglomerativo sobre la proyección IGT de la versión multivariante mediante reducción dimensional mediante MCA usando las dos primeras dimensiones y separando los pacientes fallecidos y no fallecidos (Figura 5.8 a).



(c) Dendrograma del agrupamiento jerárquico



(b) Resultado del método del codo para la elección del número de *clusters*



(a) Agrupación de los datos en 2 *clusters*

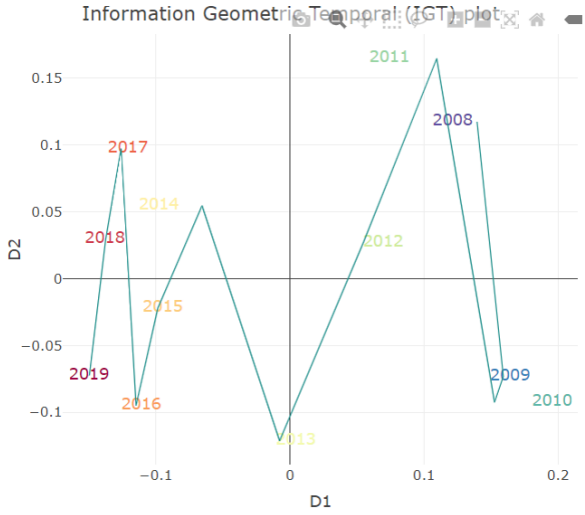
(d) Agrupación de los datos en 3 *clusters*

Figura 5.25: Resultado del agrupamiento jerárquico de la unión de las dos primeras dimensiones de los MCA de las clases positiva y negativa

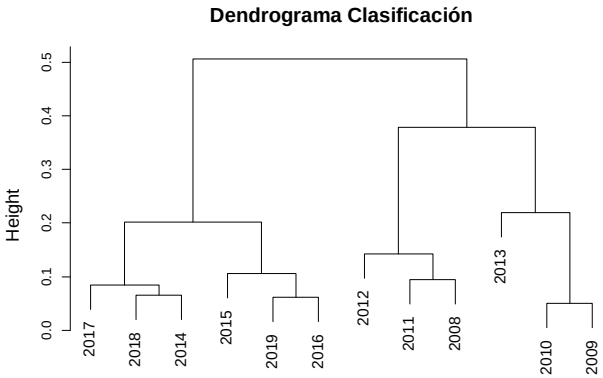
Siguiendo el criterio establecido para el método del codo, se ha dividido el dataset en 2 grupos,los cuales son los dos grupos correspondientes a los años pre y post implantación de códigos ICD 10. Sin embargo, el análisis del dendrograma también da lugar a la posibilidad de formar 3 grupos diferentes. Es por ello que las IGTs obtenidas de las métricas de los modelos predictivos se han agrupado en 2 y 3 grupos diferentes de la misma forma que los datos resultantes de análisis temporal para estudiar la relación entre ellos.

5.4.1 Random Forest

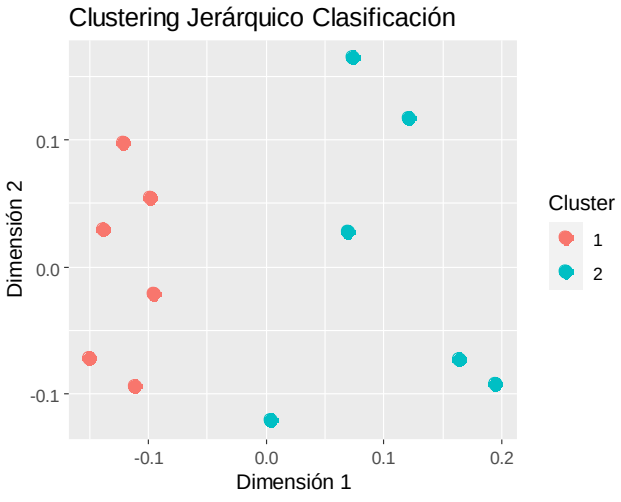
Las figuras 5.26, 5.27, 5.28, 5.29 y 5.30 muestran la representación de la IGT de las métricas de los modelos de *Random Forest* y su agrupación en 2 y 3 *clusters*.



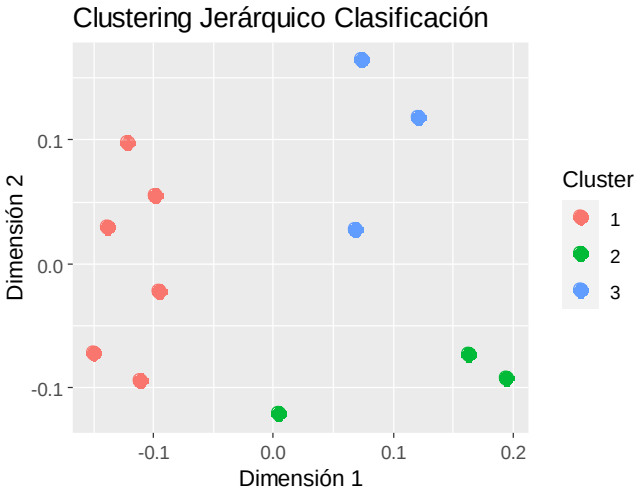
(a) IGT del área bajo la curva ROC en el modelo *Random Forest*



(b) Dendrograma del agrupamiento del área bajo la curva ROC en el modelo *Random Forest*

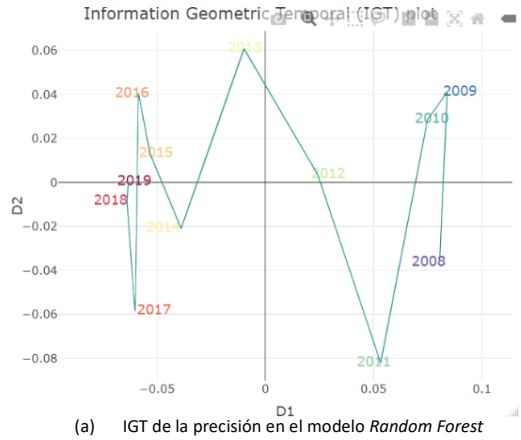


(c) Agrupación de los datos en 2 *clusters*

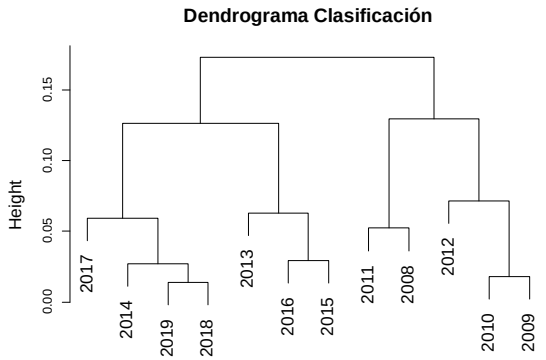


(d) Agrupación de los datos en 3 *clusters*

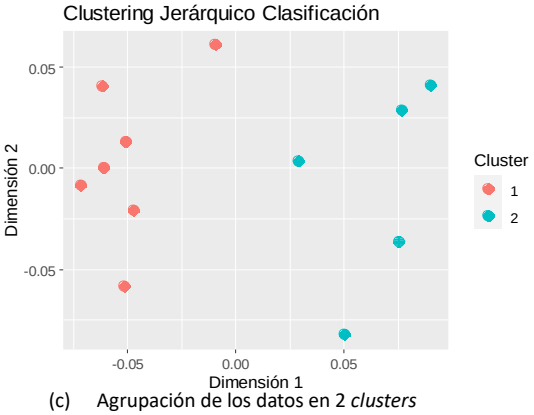
Figura 5.26: Resultados del agrupamiento de la métrica área bajo la curva ROC en el modelo *Random Forest*



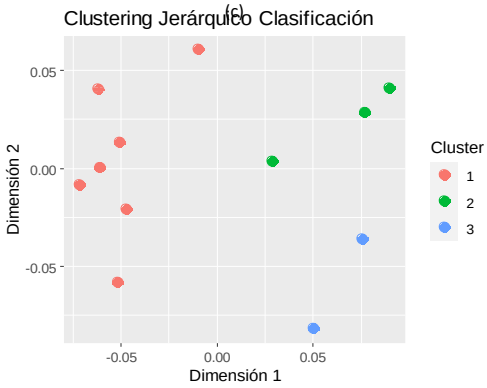
(a) IGT de la precisión en el modelo *Random Forest*



(b) Dendrograma del agrupamiento de la precisión en el modelo *Random Forest*

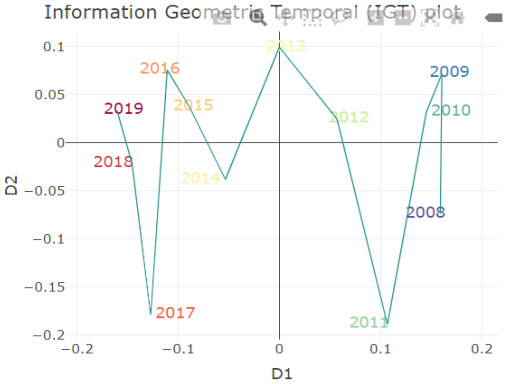


(c) Agrupación de los datos en 2 *clusters*

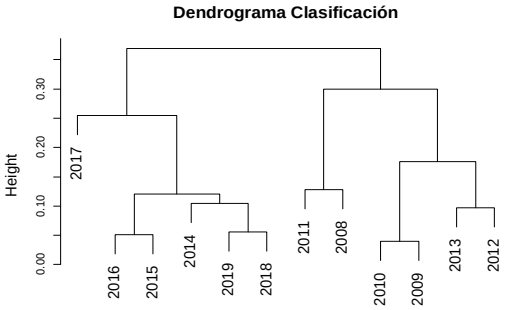


(d) Agrupación de los datos en 3 *clusters*

Figura 5.27: Resultados del agrupamiento de la métrica precisión en el modelo *Random Forest*

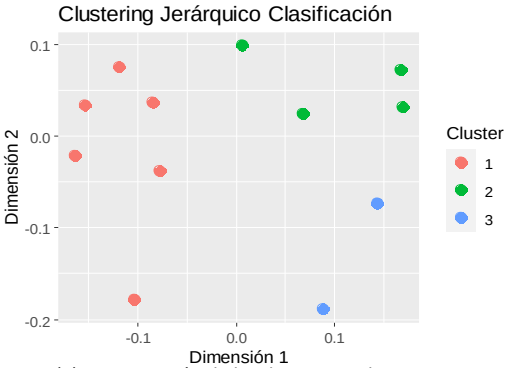


(a) IGT de la sensibilidad en el modelo *Random Forest*



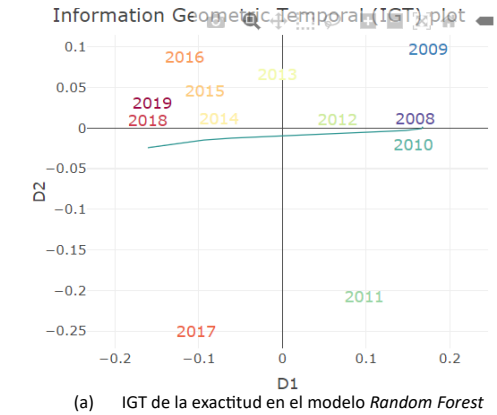
(b) Dendrograma del agrupamiento de la sensibilidad en el modelo *Random Forest*

(c) Agrupación de los datos en 2 *clusters*

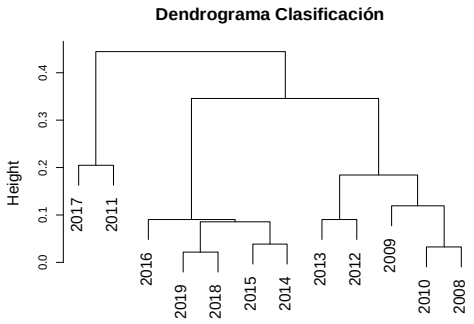


(e) Agrupación de los datos en 3 *clusters*

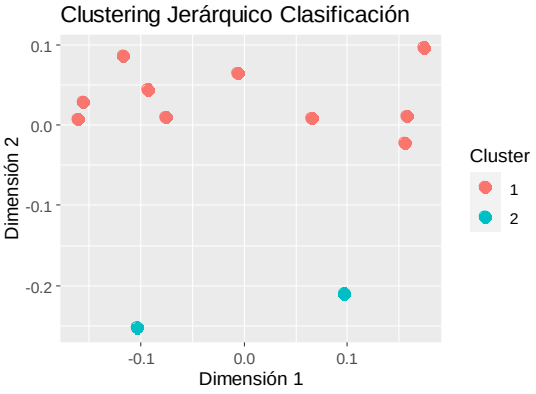
Figura 5.28: Resultados del agrupamiento de la métrica sensibilidad en el modelo *Random Forest*



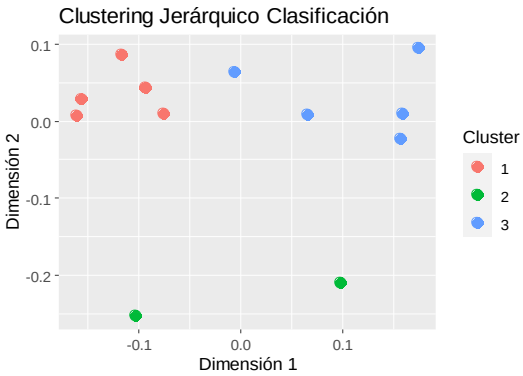
(a) IGT de la exactitud en el modelo *Random Forest*



(b) Dendrograma del agrupamiento de la exactitud en el modelo *Random Forest*

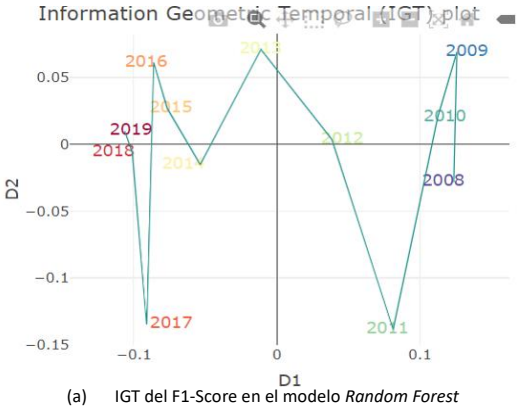


(c) Agrupación de los datos en 2 *clusters*

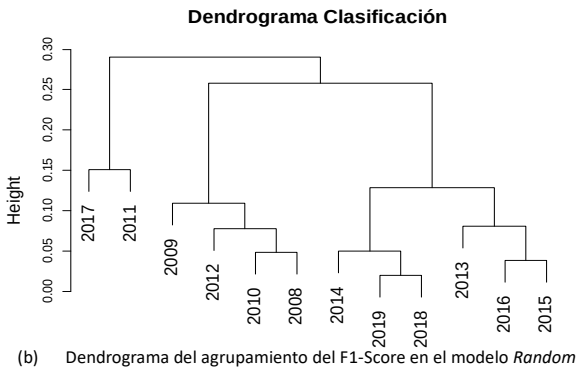


(d) Agrupación de los datos en 3 *clusters*

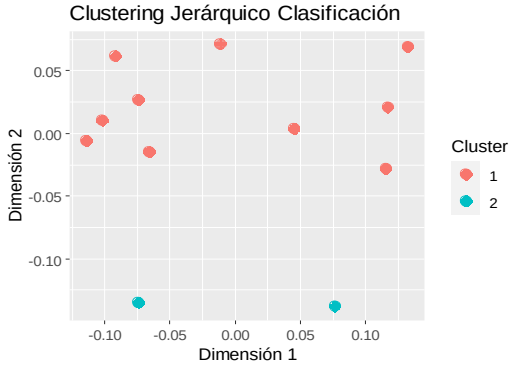
Figura 5.29: Resultados del agrupamiento de la métrica exactitud en el modelo *Random Forest*



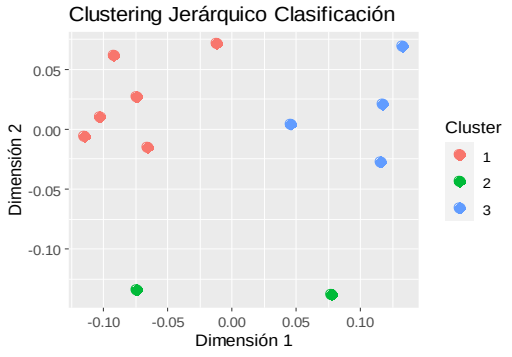
(a) IGT del F1-Score en el modelo *Random Forest*



(b) Dendrograma del agrupamiento del F1-Score en el modelo *Random Forest*



(c) Agrupación de los datos en 2 *clusters*



(d) Agrupación de los datos en 3 *clusters*

Figura 5.30: Resultados del agrupamiento de la métrica F1-Score en el modelo *Random Forest*

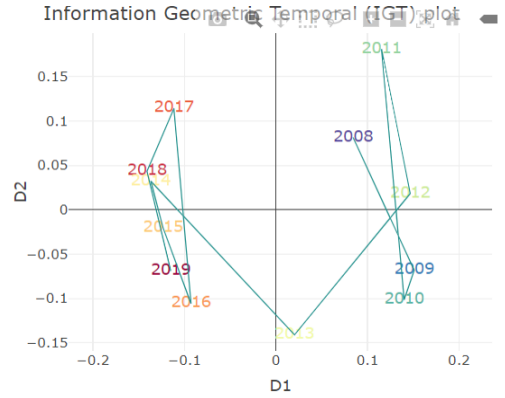
Tabla 5.3: P-valor tras aplicar un test chi-cuadrado para probar la hipótesis de independencia de ambos agrupamientos comparando el agrupamiento jerárquico de la IGT obtenida de las métricas del modelo *Random Forest* y del resultado de la unión de las dos primeras dimensiones del MCA de cada clase del *dataset*

Métrica	**p-valor para Nº clusters = 2**	**p-valor para Nº clusters = 3**
Área bajo la curva ROC	0.003892	0.01735
Precisión	0.01917	0.025
Sensibilidad	0.003892	0.004701
Exactitud	1	0.03015
F1-Score	1	0.08857

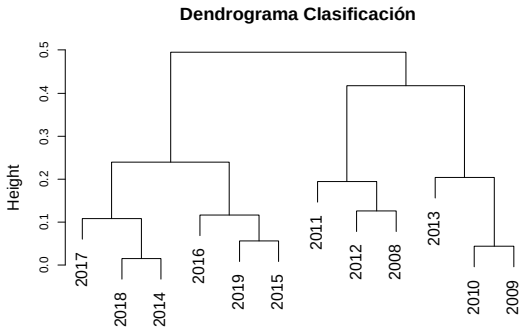
Tras obtener los resultados de todos los agrupamientos de las métricas del modelo *Random Forest* en 2 y 3 agrupamientos tal y como se había justificado previamente y realizar un test chi-cuadrado con el resultado del agrupamiento del análisis temporal (figura 5.25), para las métricas área bajo la curva ROC, precisión, sensibilidad y exactitud, aunque en este último caso solo para un número de agrupamientos igual a 3, se ha obtenido un p-valor inferior al límite de 0.05 que se había establecido, por lo que se puede concluir con suficiente relevancia estadística que se rechaza la hipótesis de independencia de los dos agrupamientos comparados, quedando así demostrada la relación entre el análisis de variabilidad temporal y los resultados del modelo de *Random Forest*. Para las métricas F1-Score y exactitud, en este último caso con el número de clusters igual a 2, el p-valor es superior a 0.05, por lo que no se puede afirmar con suficiente relevancia estadística que se rechaza la hipótesis nula y se asume la independencia entre los grupos comparados.

5.4.2 Gradient Boosting

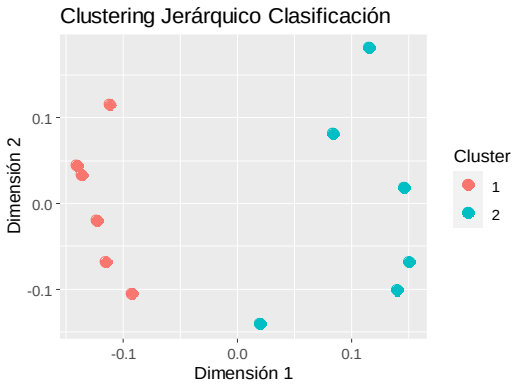
Las figuras 5.31, 5.32, 5.33, 5.34 y 5.35 muestran la representación de la IGT de las métricas de los modelos de *Gradient Boosting* y su agrupación en 2 y 3 *clusters*.



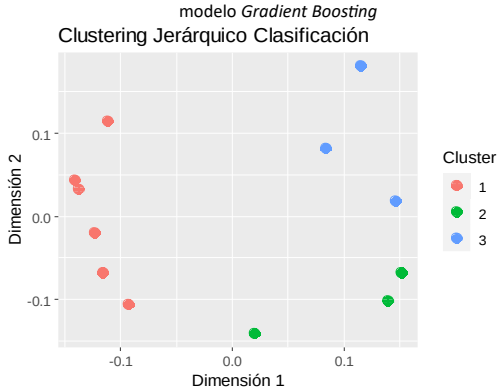
(a) IGT de la sensibilidad en el modelo *Gradient Boosting*



(b) Dendrograma del agrupamiento de la sensibilidad en el modelo *Gradient Boosting*

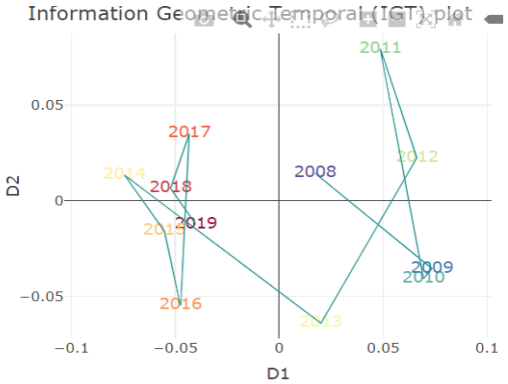


(c) Agrupación de los datos en 2 *clusters*

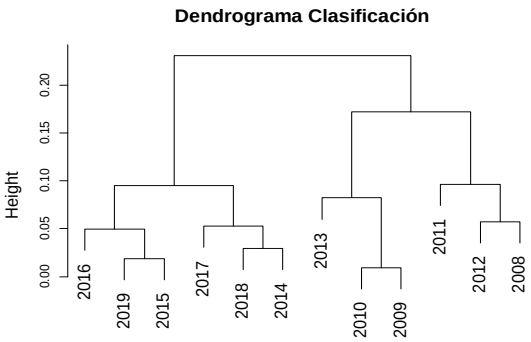


(d) Agrupación de los datos en 3 *clusters*

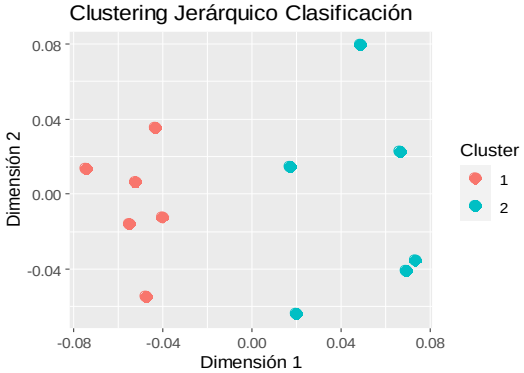
Figura 5.31: Resultados del agrupamiento de la métrica área bajo la curva ROC en el modelo *Gradient Boosting*



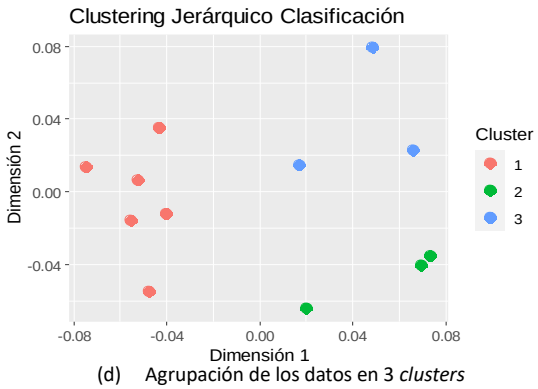
(a) IGT de la sensibilidad en el modelo *Gradient Boosting*



(b) Dendrograma del agrupamiento de la sensibilidad en el modelo *Gradient Boosting*

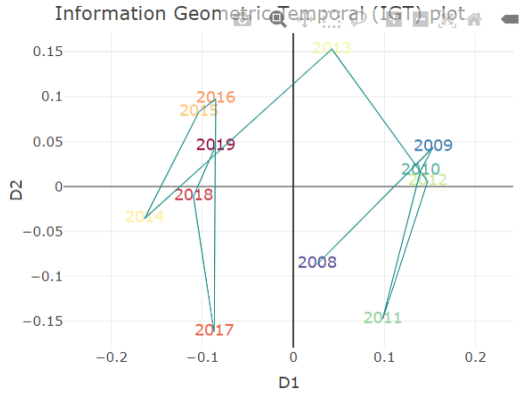


(c) Agrupación de los datos en 2 *clusters*

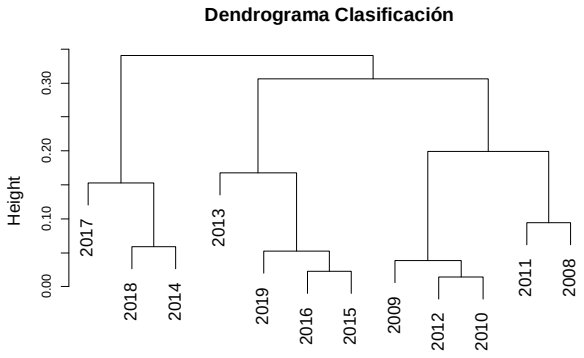


(d) Agrupación de los datos en 3 *clusters*

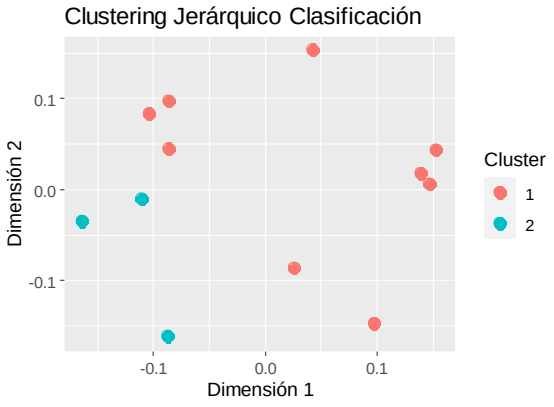
Figura 5.32: Resultados del agrupamiento de la métrica precisión en el modelo *Gradient Boosting*



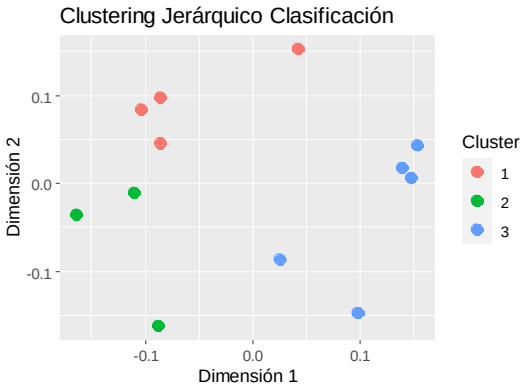
(a) IGT de la sensibilidad en el modelo *Gradient Boosting*



(b) Dendrograma del agrupamiento de la sensibilidad en el modelo *Gradient Boosting*

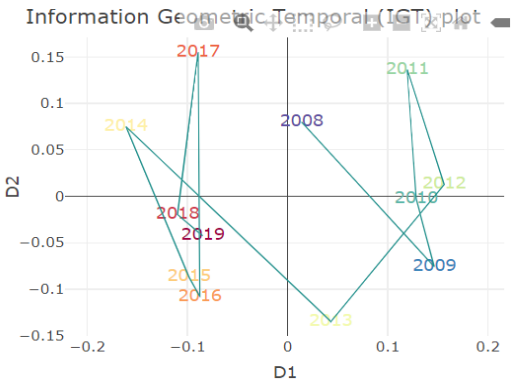


(c) Agrupación de los datos en 2 *clusters*

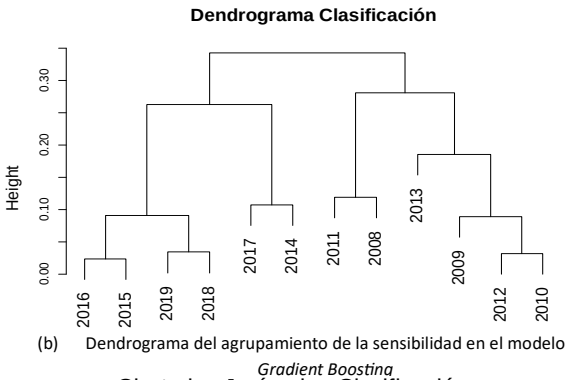


(d) Agrupación de los datos en 3 *clusters*

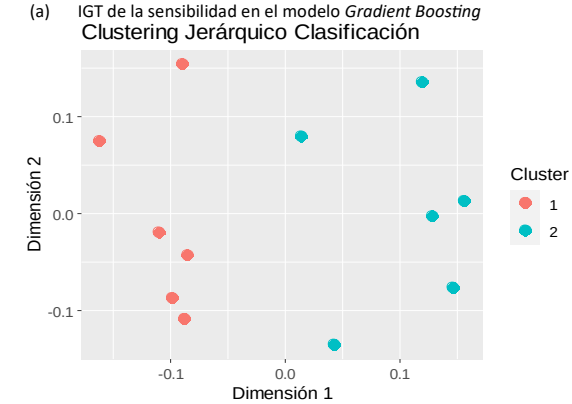
Figura 5.33: Resultados del agrupamiento de la métrica sensibilidad en el modelo *Gradient Boosting*



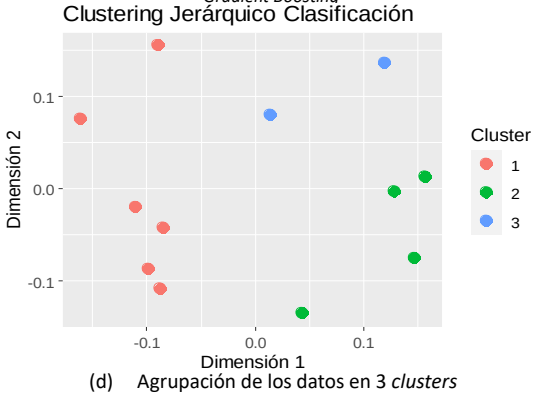
(a) IGT de la sensibilidad en el modelo *Gradient Boosting*



(b) Dendrograma del agrupamiento de la sensibilidad en el modelo *Gradient Boosting*

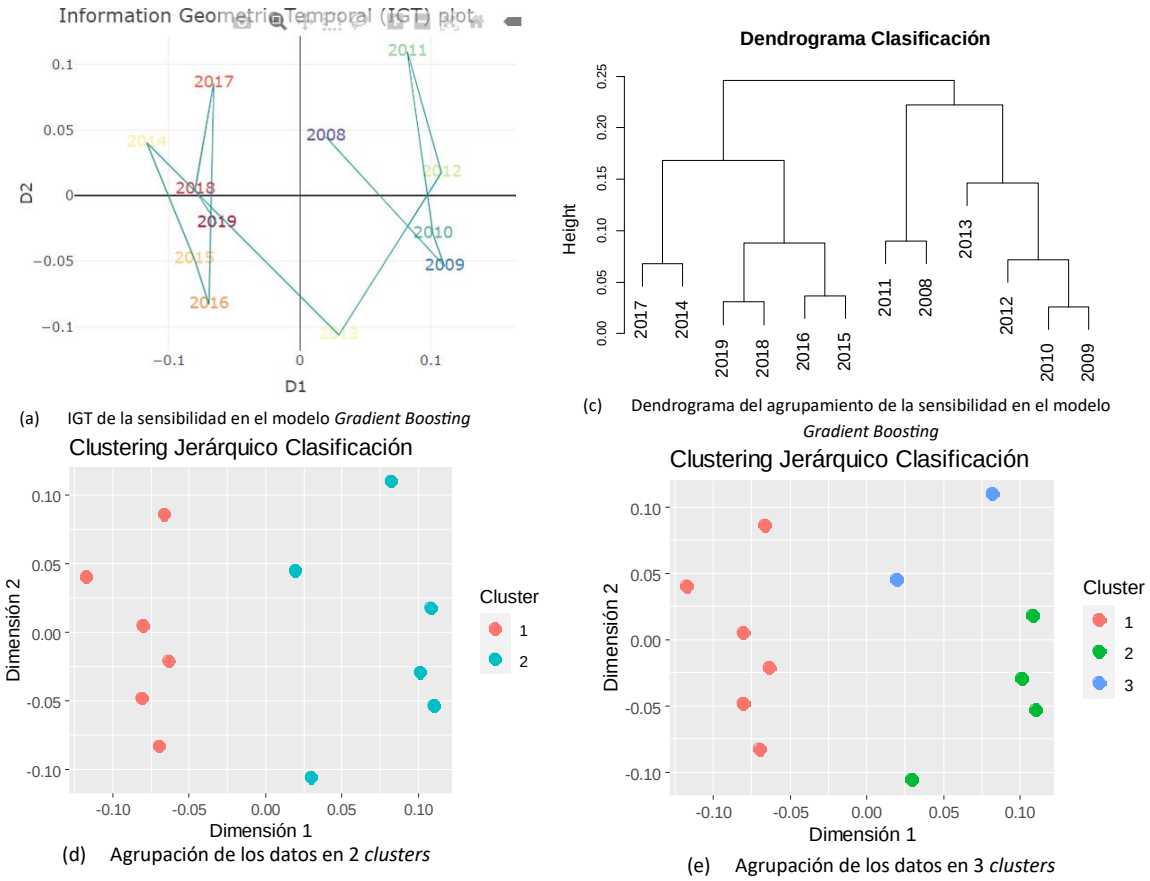


(c) Agrupación de los datos en 2 *clusters*



(d) Agrupación de los datos en 3 *clusters*

Figura 5.34: Resultados del agrupamiento de la métrica exactidud en el modelo *Gradient Boosting*



(a) IGT de la sensibilidad en el modelo *Gradient Boosting*

(c) Dendrograma del agrupamiento de la sensibilidad en el modelo *Gradient Boosting*

(d) Agrupación de los datos en 2 *clusters*

(e) Agrupación de los datos en 3 *clusters*

Figura 5.35: Resultados del agrupamiento de la métrica exactitud en el modelo *Gradient Boosting*

Tabla 5.4: P-valor del agrupamiento jerárquico de la IGT obtenida de las métricas del modelo *Gradient Boosting* y del resultado de la unión de las dos primeras dimensiones del MCA de cada clase del *dataset* tras aplicar un test chi-cuadrado sobre para probar la hipótesis de independencia de ambos agrupamientos

Métrica	**Nº clusters = 2**	**Nº clusters = 3**
Área bajo la curva ROC	0.003892	0.01735
Precisión	0.00389	0.01735
Sensibilidad	0.1824	0.02891
Exactitud	0.003892	0.004701
F1-Score	0.003892	0.004701

Tras obtener los resultados de todos los agrupamientos de las métricas del modelo *Gradient Boosting* y realizar un test chi-cuadrado con el resultado del agrupamiento del análisis temporal (figura 5.25), en todos los casos a excepción de la métrica de la sensibilidad para un número de *clusters* igual a 2, se ha obtenido un p-valor inferior al límite de 0.05 que se había establecido, por lo que se puede afirmar con

suficiente relevancia estadística que se rechaza la hipótesis de independencia de los dos agrupamientos comparados, quedando así demostrada la relación entre el análisis de variabilidad temporal y los resultados del modelo de *Gradient Boosting*. En el caso de la métrica de la sensibilidad para un número de agrupamientos igual a 2 el p-valor es superior a 0.05, por lo que se cumple la hipótesis nula de independencia de los dos conjuntos comparados y no se puede afirmar que los resultados estén relacionados.

Capítulo 6

Discusión

En el presente capítulo se discute la relevancia de los resultados obtenidos, así como las limitaciones de este y las posibles líneas de investigación futuras.

6.1 Relevancia

En el presente estudio, se ha demostrado la existencia de la relación entre los resultados del análisis de variabilidad temporal y los modelos predictivos desarrollados, quedando así probado el objetivo principal del trabajo. Ha quedado reflejado cómo las variaciones en las distribuciones de probabilidad descubiertas a lo largo de los lotes temporales de los datos debidas a, entre otros posibles factores, el cambio en el protocolo de registro de los diagnósticos y procedimientos pasando del uso de códigos ICD-9 a ICD-10 afecta al rendimiento de las predicciones, habiendo un salto en los resultados de los modelos cuando se utilizan datos posteriores a la implantación de los códigos ICD-10 con modelos previos a estos y vicecersa.

Asimismo, los resultados obtenidos de los análisis temporal multivariante y univariante proporcionan una gran información acerca de la variabilidad en la distribución de probabilidad del dataset MIMIC IV en el tiempo, permitiendo identificar cambios, patrones, anomalías o sesgos. Por un lado, los estudios de la distribución de probabilidad de los datos año a año (figuras 5.5, 5.6 y 5.7) muestran saltos bastante notables entre las distribuciones de los datos entre años consecutivos. Este hecho puede estar debido al preprocesado aplicado sobre los datos con el objetivo de llevar a cabo su desidentificación y obtención de fecha real, condicionando de esta forma todos los resultados obtenidos en el trabajo. Sin embargo, se ha optado por aplicar un método de desidentifiación para obtener los datos año a año en lugar de usar los datos agrupados en conjuntos de 3 años por dos razones principales. En primer lugar, por el respaldo de la existencia de artículos científicos ya publicados con el uso de este preprocesado, como el caso de Yao, H. et al., (2023); y en segundo lugar, la obtención de los datos año a año permite llevar a cabo un análisis más exhasutivo sobre la presencia de variabilidad temporal y los factores que puedan provocarla.

En los resultados mencionados en el párrafo anterior se observa claramente un aumento en la dispersión de la distribución de probabilidad de los datos a partir de la implantación de los códigos ICD-10, resultado esperable por otro lado debido al aumento de posibilidades de clasificación de los diagnósticos y procedimientos, ya que los códigos ICD-9 se siguieron usando y se implantó a la vez un nuevo protocolo con un mayor número de códigos que el anterior, por lo que la probabilidad se distribuyó en un mayor dominio.

Por otro lado, tal y como se ha mencionado previamente en la redacción de resultados, la variabilidad encontrada en los datos ha sido asociada tanto con la presencia de *covariate shifts* como de *concept shifts*, y no quedando demostrada la existencia de *prior probability shift* en los estudios realizados.

También cabe destacar que los resultados del análisis temporal están influenciados por la gran diferencia en la prevalencia de las dos clases. Esto ha provocado que la influencia de las variables sobre la variación de la clase negativa sea mucho más homogénea que la de la clase positiva debido al mayor número de observaciones de cada variable. En el caso de la clase positiva, las variables con mayor prevalencia son las que presentan una mayor influencia en las variaciones de la distribución de probabilidad de esta clase.

Con respecto a los resultados obtenidos de los modelos predictivos, no se puede concluir que haya uno mejor que otro, ya que ambos presentan resultados muy similares en todas las métricas obtenidas. Sin embargo, lo que sí que cabe mencionar es que para este estudio, en todas las métricas obtenidas a excepción de la sensibilidad el resultado del p-valor obtenido del test chi-cuadrado de los agrupamientos de las métricas y el análisis temporal ha sido menor para los resultados del modelo *Gradient Boosting*, pudiendo asumir también en este caso la relación entre ambos agrupamientos con un mayor número de métricas, ya que a excepción de la sensibilidad para un número de *clusters* igual a 2, en el resto de agrupamientos se pude asumir dependencia ente ellos. Para el modelo de *Random Forest,* las métricas para las que pasa esto eran menos, presentando independencia solamente en el área bajo la curva ROC, precisión, sensibilidad, y exactitud para 3 *clusters.* Para los casos de exactitud y F1-Score, el p-valor obtenido para un número de *clusters* igual a 2 ha sido de 1, resultado el cual es extraño y que se explica debido al mal funcionamiento del método de agrupamiento utilizado para estos casos debido a la presencia de 2 lotes temporales los cuales actúan como *outliers* y que provocan que la partición de los datos en 2 conjuntos no se realice de forma correcta. Es por ello, que sería conveniente también en un futuro implementar un nuevo método de agrupamiento más robusto ante la presencia de *outliers* como en estos casos. Es por ello que estos resultados no son concluyentes para decir que la relación del análisis temporal es mayor con el modelo *Gradient Boosting* que con el *Random Forest*, pero sería conveniente la realización de este estudio en nuevos *datasets* para observar si se repite este patrón o es un caso puntual para los datos presentes en este trabajo.

El análisis desarrollado en el presente trabajo es especialmente importante en el ámbito médico, no solo porque está en juego la vida de pacientes y los resultados tienen que ser lo más precisos posibles, sino también porque el entorno en el que se recopilan estos datos es dinámico y está sujeto a diversas influencias, como cambios en las prácticas clínicas, avances tecnológicos, variaciones en la población o alteraciones en los estándares de atención médica. Todos estos factores pueden dar lugar a cambios en las distribuciones de los datos.

Además, otro aspecto importante de este tipo de estudios es evaluar la estabilidad y la generalización de los modelos de IA a lo largo del tiempo. La comprensión de las variaciones temporales permite identificar si los modelos de *machine learning* son capaces de mantener su rendimiento óptimo en diferentes momentos y situaciones, y en el caso de identificar cambios, ajustar los modelos en consecuencia adaptando el algoritmo, las variables utilizadas o los parámetros para que se mantenga la robustez a lo largo del tiempo.

Por último, este trabajo ha permitido llevar una revisión del estado del arte de los métodos de análisis de la variabilidad temporal de los repositorios de datos biomédicos y establecer un método para relacionar esta variabilidad con el resultado de los modelos de inteligencia artificial desarrollados. Gracias al uso de un dataset tan ampliamente utilizado en la comunidad científica los resultados obtenidos han podido ser comparados con otras publicaciones científicas. Al igual que en artículo de Yao, H. et al., (2023), se ha detectado la presencia de variabilidad temporal, pero en el presente trabajo sí que se han categorizado los *dataset shifts* encontrados, o el ejemplo del artículo Ji, C. X. et al., (2023), donde también se muestra la influencia de los *dataset shifts* sobre la creación de modelos predictivos, pero diferenciándose del presente trabajo en que en este se ha propuesto un método para demostrar esta influencia de forma sencilla utilizando la herramienta de la IGT.

6.2 Limitaciones

La principal limitación que se ha tenido a la hora desarrollar este estudio ha sido la búsqueda de un dataset que reuniese las características deseadas descritas en el apartado 3.1, limitación la cual fue aliviada en gran medida al encontrar y tener la posibilidad de usar tras la aprobación pertinente la base de datos MIMIC-IV. Sin embargo, el uso de esta base de datos supone un ligero contratiempo, y este es el hecho de haber tenido que llevar a cabo un proceso de desidentificación de los datos. A pesar de haber utilizado el método desarrollado por Yao, H. et al., (2023) para obtener las fechas de una forma aproximada tal y como se ha comentado en el apartado 4.1. Esté método no completamente preciso, y el alcance de esta imprecisión sobre los resultados del trabajo está por determinar. Si bien, tras realizar el análisis utilizando los datos agrupados en grupos de 3 años ofrece resultados equivalentes.

6.3 Trabajo futuro

A partir de la realización de este trabajo se han abierto varias líneas de investigación futuras. La primera de todas es la aplicación de un método de reducción de la dimensionalidad no lineal, como T-sne o autoencoders, los cuales permitan explicar en mayor medida los agrupamientos debidos a la variabilidad de los datos que el MCA utilizado en este caso.

En segundo lugar, evaluación d ellos resultados de modelado mediante redes neuronales con el mismo objetivo que los modelos de *machine learning* desarrollados en este trabajo con el fin de estudiar los resultados y la robustez ante la variabilidad temporal de los datos de entrenamiento de este nuevo modelo.

Por último, generalizar los resultados de este estudio sobre otros repositorios biomédicos distintos, como el UCI ML Drug Review dataset y el COVID-19 effect on Liver Cancer Prediction Dataset.

Capítulo 7

Conclusión

En el presente capítulo, se exponen las conclusiones finales obtenidas a partir de la realización del trabajo.

7.1 Conclusiones

Ante el creciente interés de los sistemas basados en inteligencia artificial en medicina, se ha identificado el problema del uso de datos de baja calidad asociado a la variabilidad de los datos a lo largo del tiempo, dando lugar a modelos poco robustos o efectivos. Para ello, se ha llevado a cabo un estudio de la variabilidad de los mismos a lo largo del tiempo con el objetivo de identificar cómo afecta a los resultados obtenidos de los modelos y qué procedimientos se pueden aplicar para intentar reducir el efecto de la variabilidad. De este modo, a la vista de los resultados obtenidos se puede llegar a las siguientes conclusiones, separándolas en conclusiones principales (CP) relacionadas con el objetivo principal y conclusiones secundarias (CSX) relacionadas con los objetivos secundarios, donde X es el objetivo al que hacen referencia:

CP. Se ha conseguido demostrar estadísticamente que los resultados del análisis multivariante de la variabilidad temporal y de los modelos de *machine learning* están relacionados.

CS1. Tras los resultados obtenidos de los análisis univariantes y multivariantes, se ha confirmado la presencia de variabilidad en la distribución de probabilidad a lo largo del tiempo en los datos del repositorio MIMIC-IV. Dentro de esta variabilidad se han identificado *covariate shifts* y *concept shifts*, y la causa principal de esta es el cambio de protocolo de registro de los diagnósticos y procedimientos aplicados a los pacientes que se produjo en el año 2015 con la aparición de los códigos ICD 10.

CS1. Se ha identificado cuáles son las variables que producen la variabilidad temporal sobre la base de datos completa, y cómo afecta cada variable a las variaciones en la distribución de probabilidad de cada clase.

CS1. Se ha demostrado que el rendimiento de los modelos predictivos desarrollados está directamente relacionado con los datos utilizados para entrenamiento y validación, presentando mejores resultados los modelos entrenados y validados con datos de lotes temporales pre y post implantación del uso de códigos ICD-10 por separado que mezclándolos entre ellos, siendo los modelos que presentan mejores resultados los posteriores a 2015, demostrando de esta forma la importancia de tener en cuenta los *dataset shifts* antes de desarrollar cualquier herramienta de IA.

CS2. Gracias a los modelos de *Random Forest* y *Gradient Boosting* desarrollados, se puede afirmar que es posible llevar a cabo una clasificación de los datos con el fin de intentar predecir la muerte de un paciente durante su estancia en el hospital de manera relativamente satisfactoria.

CS3. Se ha conseguido desarrollar un algoritmo que permita comparar los resultados de los modelos de clasificación con los resultados del análisis de variabilidad temporal por medio de

la herramienta de la IGT, consiguiendo demostrar con suficiente significancia estadística la dependencia o independencia de ambos resultados.

Para finalizar, ha quedado más que demostrado que es necesario crear predictores que puedan adaptarse a las nuevas condiciones a las que puede estar sometido el ser humano, para poder garantizar su seguridad en todo momento y tener la habilidad de generalización para todos los casos posibles. El diseño de predictores y clasificadores robustos supone una gran mejora para la prevención y diagnóstico en salud.

Bibliografía

Abdi, H., & Valentin, Dominique. (2007). Multiple Correspondence Analysis. https://personal.utdallas.edu/~Herve/Abdi-MCA2007-pretty.pdf

Alaiz-Rodríguez, R., & Japkowicz, N. (2008). Assessing the impact of changing environments on classifier performance. Proceedings of the Canadian Society for computational studies of intelligence, 21st conference on Advances in artificial intelligence, 13-24.

Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., & Tuytelaars, T. (2018). Memory aware synapses: Learning what (Not) to forget. arXiv. https://doi.org/10.48550/arXiv.1711.09601

Aljundi, R., Kelchtermans, K., & Tuytelaars, T. (2019). Task-free continual learning. 11254-11263. https://openaccess.thecvf.com/content_CVPR_2019/html/Aljundi_Task-Free_Continual_Learning_CVPR_2019_paper.html

Becker, C., Mayfield, W. D., Murphy, S. Y., Wang, B., Gobbert, M. K., & Barajas, C. (2019). An approach to tuning hyperparameters in parallel: A performance study using climate data cybertraining: big data + high-performance computing + atmospheric sciences. https://doi.org/10.13016/m2dxhb-r86g

Brabec, J., Komárek, T., Franc, V., & Machlica, L. (2020). On model evaluation under non-constant class imbalance. En V. V. Krzhizhanovskaya, G. Závodszky, M. H. Lees, J. J. Dongarra, P. M. A. Sloot, S. Brissos, & J. Teixeira (Eds.), Computational Science – ICCS 2020 (pp. 74-87). Springer International Publishing. https://doi.org/10.1007/978-3-030-50423-6_6

Breiman, L. (2017). Classification and regression trees. Routledge. https://doi.org/10.1201/9781315139470

Cha, G.-W., Moon, H.-J., & Kim, Y.-C. (2021). Comparison of random forest and gradient boosting machine models for predicting demolition waste based on small datasets and categorical variables. International Journal of Environmental Research and Public Health, 18(16), 8530. https://doi.org/10.3390/ijerph18168530

Costa, P. S., Santos, N. C., Cunha, P., Cotter, J., & Sousa, N. (2013). The use of multiple correspondence analysis to explore associations between categories of qualitative variables in healthy ageing. Journal of Aging Research, 2013, 302163. https://doi.org/10.1155/2013/302163

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. En C. Zhang & Y. Ma (Eds.), Ensemble Machine Learning: Methods and Applications (pp. 157-175). Springer. https://doi.org/10.1007/978-1-4419-9326-7_5

Dockès, J., Varoquaux, G., & Poline, J.-B. (2021). Preventing dataset shift from breaking machine-learning biomarkers. arXiv. https://doi.org/10.48550/arXiv.2107.09947

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), 44:1-44:37. https://doi.org/10.1145/2523813

Ghimire, B., Rogan, J., Galiano, V. R., Panday, P., & Neeti, N. (2012). An evaluation of bagging, boosting, and random forests for land-cover classification in cape cod, massachusetts, usa. GIScience & Remote Sensing, 49(5), 623-643. https://doi.org/10.2747/1548-1603.49.5.623

Gifi, A. (1991). Nonlinear multivariate analysis.

Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the roc curve for multiple class classification problems. Machine Learning, 45(2), 171-186. https://doi.org/10.1023/A:1010920819831

Ho, T. K. (1995). Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition, 1, 278-282 vol.1. https://doi.org/10.1109/ICDAR.1995.598994

Institute of Medicine (US) Committee on Data Standards for Patient Safety. (2004). Patient safety: Achieving a new standard for care (P. Aspden, J. M. Corrigan, J. Wolcott, & S. M. Erickson, Eds.). National Academies Press (US). http://www.ncbi.nlm.nih.gov/books/NBK216086/

Ji, C. X., Alaa, A. M., & Sontag, D. (2023). Large-scale study of temporal shift in health insurance claims. arXiv. https://doi.org/10.48550/arXiv.2305.05087

Kelly, M. G., Hand, D. J., & Adams, N. M. (1999). The impact of changing populations on classifier performance. Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 367-371. https://doi.org/10.1145/312129.312285

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114(13), 3521-3526. https://doi.org/10.1073/pnas.1611835114

Levman, J., & Takahashi, E. (2015). Multivariate analyses applied to healthy neurodevelopment in fetal, neonatal, and pediatric mri. Frontiers in Neuroanatomy, 9, 163. https://doi.org/10.3389/fnana.2015.00163

Liaw, S. T., Rahimi, A., Ray, P., Taggart, J., Dennis, S., de Lusignan, S., Jalaludin, B., Yeo, A. E. T., & Talaei-Khoei, A. (2013). Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature. International Journal of Medical Informatics, 82(1), 10-24. https://doi.org/10.1016/j.ijmedinf.2012.10.001

Maclin, R., & Opitz, D. (1999). Popular ensemble methods: An empirical study. Journal of Artificial Intelligence Research, 11, 169-198. https://doi.org/10.1613/jair.614

Matthew , S. (2019). Towards Data Science. Understanding Dataset Shift. https://towardsdatascience.com/understanding-dataset-shift-f2a5a262a766

Medicine, I. of, Services, B. on H. C., & Safety, C. on D. S. for P. (2003). Patient safety: Achieving a new standard for care. National Academies Press.

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., & Herrera, F. (2012). A unifying view on dataset shift in classification. Pattern Recognition, 45(1), 521-530. https://doi.org/10.1016/j.patcog.2011.06.019

Nestor, B., McDermott, M. B. A., Chauhan, G., Naumann, T., Hughes, M. C., Goldenberg, A., & Ghassemi, M. (2018). Rethinking clinical prediction: Why machine learning must consider year of care and feature aggregation. arXiv. https://doi.org/10.48550/arXiv.1811.12583

Nonlinear multivariate analysis | wiley. (s. f.). Wiley.Com. Recuperado 27 de junio de 2023, de https://www.wiley.com/en-us/Nonlinear+Multivariate+Analysis-p-9780471926207

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. Neural Networks, 113, 54-71. https://doi.org/10.1016/j.neunet.2019.01.012

Polikar, R. (2012). Ensemble learning. En C. Zhang & Y. Ma (Eds.), Ensemble Machine Learning: Methods and Applications (pp. 1-34). Springer. https://doi.org/10.1007/978-1-4419-9326-7_1

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2008). Dataset shift in machine learning. http://www.acad.bg/ebook/ml/The.MIT.Press.Dataset.Shift.in.Machine.Learning.Feb.2009.eBook-DDU.pdf

Rebuffi, S.-A., Kolesnikov, A., Sperl, G., & Lampert, C. H. (2017). Icarl: Incremental classifier and representation learning. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5533-5542. https://doi.org/10.1109/CVPR.2017.587

Sáez, C., Gutiérrez-Sacristán, A., Kohane, I., García-Gómez, J. M., & Avillach, P. (2020). EHRtemporalVariability: Delineating temporal data-set shifts in electronic health records. GigaScience, 9(8), giaa079. https://doi.org/10.1093/gigascience/giaa079

Sáez, C., Martínez-Miranda, J., Robles, M., & García-Gómez, J. M. (2012). Organizing data quality assessment of shifting biomedical data. Studies in Health Technology and Informatics, 180, 721-725.

Sáez, C., Robles, M., & García-Gómez, J. M. (2013). Comparative study of probability distribution distances to define a metric for the stability of multi-source biomedical research data. 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 3226-3229. https://doi.org/10.1109/EMBC.2013.6610228

Sáez, C., Rodrigues, P. P., Gama, J., Robles, M., & García-Gómez, J. M. (2015). Probabilistic change detection and visualization methods for the assessment of temporal stability in biomedical data quality. Data Mining and Knowledge Discovery, 29(4), 950-975. https://doi.org/10.1007/s10618-014-0378-6

Sáez, C., Romero, N., Conejero, J. A., & García-Gómez, J. M. (2021). Potential limitations in COVID-19 machine learning due to data source variability: A case study in the nCov2019 dataset. Journal of the American Medical Informatics Association: JAMIA, 28(2), 360-364. https://doi.org/10.1093/jamia/ocaa258

Sáez, C., Zurriaga, O., Pérez-Panadés, J., Melchor, I., Robles, M., & García-Gómez, J. M. (2016). Applying probabilistic temporal and multisite data quality control methods to a public health mortality registry in Spain: A systematic approach to quality control of repositories. Journal of the American Medical Informatics Association: JAMIA, 23(6), 1085-1095. https://doi.org/10.1093/jamia/ocw010

Sáez Silvestre, C. (2016). Probabilistic methods for multi-source and temporal biomedical data quality assessment [Tesis doctoral, Editorial Universitat Politècnica de València]. https://doi.org/10.4995/Thesis/10251/62188

Sáez Silvestre, C., Robles Viejo, M., & García Gómez, J. M. (2014). Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances. Statistical Methods in Medical Research, 1-25. https://doi.org/10.1177/0962280214545122

Strauss, T., & Maltitz, M. J. von. (2017). Generalising ward's method for use with manhattan distances. PLOS ONE, 12(1), e0168288. https://doi.org/10.1371/journal.pone.0168288

T, S., & Mj, von M. (2017). Generalising ward's method for use with manhattan distances. PloS One, 12(1). https://doi.org/10.1371/journal.pone.0168288

Tercan, H., Deibert, P., & Meisen, T. (2022). Continual learning of neural networks for quality prediction in production using memory aware synapses and weight transfer. Journal of Intelligent Manufacturing, 33(1), 283-292. https://doi.org/10.1007/s10845-021-01793-0

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 12(4), 5-33. https://doi.org/10.1080/07421222.1996.11518099

Widmer, G., & Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. Machine Learning, 23(1), 69-101. https://doi.org/10.1007/BF00116900

Yao, H., Choi, C., Cao, B., Lee, Y., Koh, P. W., & Finn, C. (2023). Wild-time: A benchmark of in-the-wild distribution shift over time. arXiv. https://doi.org/10.48550/arXiv.2211.14238

https://dexur.com/pcs9/ accedido en 04/03/2023

https://www.icd10data.com/Convert accedido en 04/03/2023

Parte II

Presupuesto

Capítulo 1

Presupuesto

En esta sección, se presenta el presupuesto que ha sido necesario para desarrollar este trabajo de fin de grado.

1.1 Presupuesto desglosado

1.1.1 Costes de Hardware

Unidades	Descripción	Detalles	Proveedor	Cantidad	Precio de la unidad (€)	Precio total (€)
u	HP Envy 13	Intel® Core™ i7 processor (2.7 GHz), 8 GB LPDDR SDRAM, 256 GB SSD storage, Windows 11	HP	1	899	899
Total:						899

1.1.2 Costes de Software

Unidades	Descripción	Proveedor	Cantidad	Duración (años)	Precio de la unidad (€)	Precio total (€)
Licencia	Microsoft office profesional 2021	Microsoft	1	1	299	299
Licencia	Microsoft Windows 11 professional	Microsoft	1	1	145	145
Licencia	PyCharm profesional	Jetbrains	1	1	301.29	301.29
Total:						745.29

1.1.3 Costes de personal

Descripción	Tareas	Rango	Cantidad (h)	Precio de la unidad (€)	Gasto de la seguridad social (€)	Salario (€)	Coste total (€)
Ingeniero biomédico	Llevar a cabo el proyecto	Junior	600	15	2124	6876	9000
Científico de datos	Supervisar el proyecto	Senior	120	30	828	2772	3600
Ingeniero biomédico	Supervisar el proyecto	Senior	96	30	662.4	2217.6	2880
Total:							15480

1.2 Presupuesto total

Descripción	**Coste (€)**
Costes de Hardware	899
Costes de Software	745.29
Costes de personal	15480
Total:	17124.29