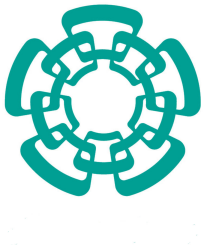

Centro de Investigación y de Estudios Avanzados
del Instituto Politécnico Nacional

Unidad Cinvestav Tamaulipas

Método de predicción de episodios hipotensivos basado en una codificación de series de tiempo de presión arterial media y cadenas de Markov

Tesis que presenta:

Jaime Edwin Arciniegas García

Para obtener el grado de:

Maestro en Ciencias en Ingeniería y Tecnologías Computacionales

Director de la Tesis:
Dr. Hiram Galeana Zapién

Cd. Victoria, Tamaulipas, México. Noviembre, 2019

La tesis presentada por Jaime Edwin Arciniegas García fue aprobada por:

_____

Dr. Edwyn Javier Aldana Bobadilla

Dr. Iván López Arévalo

Dr. Hiram Galeana Zapién, Director

Cd. Victoria, Tamaulipas, México., 14 de Noviembre de 2019

Esta investigación fue parcialmente financiada por el proyecto SALUD-2014-C01-233836 del Fondo Sectorial en Investigación en Salud y Seguridad Social (FOSISS)

Agradecimientos

- Le agradezco a mi familia por el apoyo incondicional. En especial a mi padre Jaime Arciniegas Rojas, mi madre María Isabel García Velez y mis hermanos.
- Le agradezco también a mi director de tesis el Dr. Hiram Galeana Zaién por brindarme la posibilidad de trabajar tan importante tema de investigacón, por su dirección y por su paciencia durante el desarrollo de esta tesis de posgrado.
- A mis revisores, el Dr. Edwyn Javier Aldana Bobadilla y el Dr. Iván López Arévalo por su valiosa retroalimentación a esta tesis. En especial al Dr. Edwyn Javier Aldana Bobadilla por su contribución durante el desarrollo del método de agrupación de códigos.
- A todos los investigadores del Cinvestav Unidad Tamaulipas por su orientación durante mi proceso de formación.
- Al Conacyt por el apoyo económico brindado durante los dos años de duración de la maestría.
- Al Cinvestav por darme la oportunidad de realizar un posgrado de calidad en tan prestigiosa institución.

Índice General

Índice de Figuras

Índice de Tablas

Índice de Algoritmos

Resumen

Método de predicción de episodios hipotensivos basado en una codificación de series de tiempo de presión arterial media y cadenas de Markov

por

Jaime Edwin Arciniegas García
Unidad Cinvestav Tamaulipas
Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, 2019
Dr. Hiram Galeana Zapién, Director

La recolección y almacenamiento de información clínica y fisiológica en el ámbito hospitalario ha motivado el desarrollo de sistemas de apoyo a la decisión clínica (CDSS, por sus siglas en inglés). Estos CDSS están encaminados a procesar la información recabada para que mediante el uso de métodos computacionales sea posible anticipar la ocurrencia de un episodio adverso (anormalidad en los signos vitales), lo que permitiría mejorar el tratamiento y diagnóstico del paciente. En particular, uno de los episodios adversos de mayor interés son los episodios agudos hipotensivos (AHE, por sus siglas en inglés) que ocurren cuando en una ventana de 30 minutos al menos el 90 % de las muestras de presión arterial media (MAP, por sus siglas en inglés) se encuentran por debajo del umbral de 60 mmHg. La identificación oportuna de AHEs es de gran interés en la práctica clínica ya que se sabe que es precursor de otros eventos más severos que deterioran la salud del paciente. En este contexto, la presente tesis estudia el problema de predecir la ocurrencia de AHEs a partir del análisis y procesamiento de series de tiempo de MAP. De manera particular, se propone un método que consiste en tres elementos principales: a) una codificación que permite cambiar a un nuevo espacio de códigos las series de tiempo de MAP; b) la aplicación de un algoritmo de agrupación de datos categóricos que permite encontrar en el conjunto de datos codificado los grupos correspondientes a los estados normotensivos (No-AHEs) e hipotensivos (AHEs); c) la predicción de AHEs mediante el uso de una cadena de Markov de dos o más estados para modelar la ocurrencia de AHEs en la matriz estocástica obtenida con los estados identificados en el paso previo. Para validar el método propuesto

se usó un conjunto de series de tiempo de MAP de la base de datos MIMIC-II (Monitoreo Inteligente Multiparámetro en Cuidados Intensivos II), el cual fue preparado mediante la aplicación de las tareas de pre-procesamiento típicas. A cada serie de tiempo se identificaron los episodios hipotensivos contenidos en ésta (marcas de clase) a partir de la definición AHEs ampliamente utilizada. De acuerdo a los resultados obtenidos, el método propuesto permite predecir AHEs en ventanas de tiempo de hasta 7 minutos con una precisión aceptable, alcanzando resultados similares a los reportados en la literatura.

Abstract

Method of prediction of hypotensive episodes based on a time series coding of mean arterial pressure and Markov chains

by

Jaime Edwin Arciniegas García
Cinvestav Tamaulipas
Center for Research and Advanced Studies of the National Polytechnic Institute, 2019
Dr. Hiram Galeana Zapién, Advisor

The acquisition and storage of clinical and physiological information in the hospital scope have motivated the development of clinical decision support systems (CDSS). This CDSS is aimed at processing the information collected so that through the use of computational methods it is possible to anticipate the occurrence of an adverse episode (or abnormality in vital signs), which would improve the treatment and diagnosis of patients. In particular, one of the most interesting adverse events is acute hypotensive episodes (AHE) that occur when in a 30 minutes window at least 90 % of the mean arterial pressure samples (MAP) are below a threshold of 60 mmHg. The timely identification of AHEs is of great interest in clinical practice since it is known to be a precursor to other more severe events that deteriorate the patient's health. In this context, the present thesis studies the problem of predicting the occurrence of AHEs from the analysis and processing of time series of MAP. In particular, the proposed method consists of three main elements: a) an encoding approach that allows changing the MAP time series to a new code space; b) the application of a categorical data grouping algorithm that allows to find in the coded data set the groups corresponding to the normotensive (Non-AHEs) and hypotensive (AHEs) states; c) the prediction of AHEs by using a Markov chain of two or more states to model the occurrence of AHEs in the stochastic matrix obtained from the states identified in the previous step. To validate the proposed method, a set of MAP time series from the MIMIC-II database (Multiparameter Intelligent Monitoring in Intensive Care II) was used, which was prepared to apply the typical pre-processing tasks. Also, in each time series, the hypotensive episodes

contained in it (class marks) were identified from the widely used AHEs definition. According to the results obtained, the proposed method allows predicting AHEs in time windows of up to 7 minutes with acceptable accuracy, achieving results similar to those reported in the literature.

Nomenclatura

ABP	Presión arterial
ACC	Exactitud
AG	Algoritmo genético
AHE	Episodio agudo hipotensivo
ANN	Redes neuronales artificiales
AUC	Área bajo la curva
CDSS	Sistemas de apoyo a la decisión clínica
DBP	Presión arterial diastólica
ECG	Electrocardiograma
FN	Falso negativo
FP	Falso positivo
ICU	Unidad de cuidados intensivos
k-**NN**	k vecinos más cercanos
MAP	Presión arterial media
MIMIC	Monitoreo Inteligente Multiparamétrico en Cuidados Intensivos
SaO2	Saturación de oxígeno en la sangre
SBP	Presión arterial sistólica
SEN	Sensibilidad
SPE	Especificidad
SpO2	Saturación de oxígeno capilar periférica
SVM	Máquina de vectores de soporte
TN	Verdadero negativo
TP	Verdadero positivo

1

Introducción

En el contexto de sistemas de apoyo a la decisión clínica (CDSS, por sus siglas en inglés), en este capítulo se describen los antecedentes y motivaciones de abordar el problema de predicción de episodios agudos hipotensivos (AHEs, por sus siglas en inglés). Asimismo, se presentan los objetivos y la metodología para la realización de la presente investigación.

1.1 Antecedentes y motivación

La aplicación de tecnologías de información y de comunicación (TICs) en el ámbito médico ha permitido desarrollar sistemas computacionales que forman parte de las herramientas de apoyo que los profesionales de la salud [1] disponen para el seguimiento del estado de salud de pacientes. Para ello es necesario que tales sistemas usados en unidades hospitalarias permitan recolectar y almacenar dos tipos de información de los pacientes [53]: 1) información clínica, la cual se refiere a todos aquellos

[1] Personas capacitadas que brindan atención médica a pacientes en hospitales, los cuales pueden ser personal de enfermería, médicos de atención primaria, médicos especialistas, etc.

aspectos relativos al expediente de un paciente como es el caso de datos personales, resultados de laboratorio, medicamentos suministrados, bitácoras de enfermería, etc.; y 2) información fisiológica consistente en series de tiempo de signos vitales que los profesionales de la salud visualizan en monitores ubicados en la cabecera del paciente para conocer tendencias estadísticas básicas de parámetros fisiológicos del paciente.

En este contexto, la creciente disponibilidad de ese tipo de información en la forma de bases de datos especializadas ha motivado cambios de paradigma como es el caso de la denominada medicina predictiva, la cual está enfocada en detectar de forma oportuna la ocurrencia de una enfermedad o episodio adverso [73]. Para hacer eso posible, uno de los principales pilares de la medicina predictiva consiste en explotar la información histórica recopilada para extraer el conocimiento oculto en aras de modelar episodios adversos específicos y emplearlos para predecir su ocurrencia. En este sentido, se prevé que los futuros CDSS incorporen técnicas computacionales para realizar las tareas de predicción del estado de salud de un paciente en términos de la ocurrencia de un episodio adverso, el cual se define como una alteración a un signo vital respecto a una condición de salud normal del paciente [60]. Dicha normalidad se establece en función de valores de referencia (umbrales) definidos para cada signo vital.

En la Figura 1.1 se ilustra el modelo conceptual de un CDSS genérico, el cual está compuesto por un conjunto de etapas funcionales que van desde la extracción de datos fisiológicos o clínicos, hasta las tareas de análisis e inferencia de datos para el pronóstico o diagnóstico de la evolución temporal [43]. Cada una de las etapas ilustradas se resume a continuación:

1. **Adquisición de datos.** Corresponde a la información clínica o fisiológica de un paciente, la cual puede estar registrada en un sistema de bases de datos existente. En el caso de la información fisiológica, las mediciones se obtienen de monitores de signos vitales y son representadas como series de tiempo de signos vitales a diferente granularidad en función de la frecuencia de muestreo considerada en esta etapa.

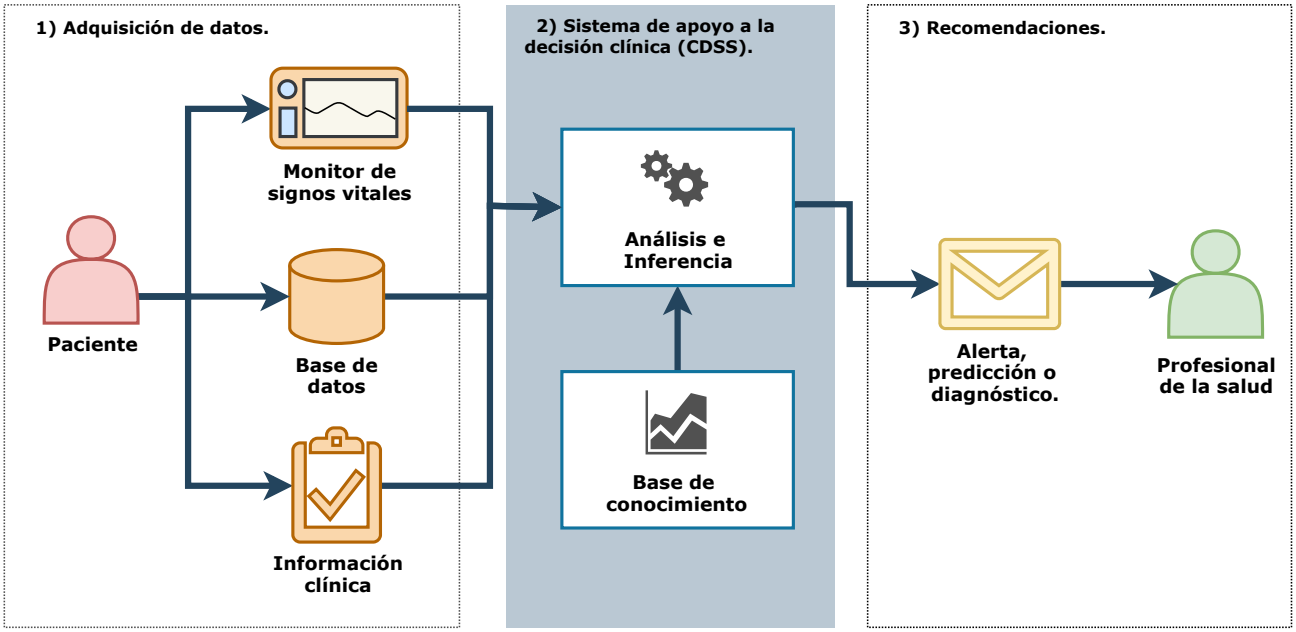

Figura 1.1: Diagrama genérico de un sistema de apoyo a la decisión clínica (CDSS).

2. **Sistema de apoyo a la decisión clínica.** Un CDSS está compuesto generalmente por dos componentes básicos: una base de conocimiento y un proceso de análisis e inferencia. La base de conocimiento es una estructura de conocimiento o un modelo de datos que contiene la información de las relaciones existentes entre las variables. Dicha información contenida en una base de conocimiento es usada para caracterizar anormalidades, síntomas o enfermedades de un paciente. Por otra parte, el proceso de análisis e inferencia puede ser potencialmente realizado por técnicas de aprendizaje máquina o reconocimiento de patrones que permitan explotar los datos contenidos en la base de conocimiento con el objetivo de extraer información relevante de los datos disponibles [9]. Con esto en mente, un CDSS puede emplear diversas técnicas de análisis e inferencia de datos como regresión logística, red neuronal artificial, máquina de vectores de soporte, árboles de decisión, algoritmos genéticos y Naive Bayes [45, 54, 56]. Estas técnicas han sido empleadas en diferentes contextos o dominios de aplicación como es el caso de oncología, reconocimiento de objetos, pacientes con Alzeheimer, diabetes, hepatitis, radiología, entre otros [6, 64]. En el contexto de la tesis es de interés el desarrollo de un enfoque

que permita el procesamiento de una base de conocimiento para la caracterización de un tipo de episodios adversos de forma que sea posible su integración en un enfoque de predicción de los mismos.

3. **Recomendaciones.** Por último, el CDSS debe incluir un medio de visualización de los resultados de la etapa de procesamiento en términos, por ejemplo, de una ventana de predicción de valores de un signo vital en observación. Este tipo de resultados se presentan como recomendaciones (alertas, predicciones o diagnósticos) que pueden apoyar al profesional de la salud para identificar de forma temprana la ocurrencia de un episodio adverso, lo cual permitiría iniciar de manera oportuna un tratamiento. Cabe aclarar que la información que proporcione un CDSS está enfocado a que el profesional de la salud disponga de indicios de posibles alteraciones o episodios adversos que el sistema prevé puedan ocurrir. Sin embargo, la valoración y decisión final del tratamiento y seguimiento recae en el profesional de la salud con base a su criterio y experiencia.

De acuerdo a las características descritas de los CDSSs, éstos son de gran relevancia en el contexto de las unidades de cuidados intensivos (ICU, por sus siglas en inglés) en las cuales es preciso disponer de herramientas de vigilancia médica que permitan un diagnóstico oportuno para la predicción de episodios adversos [11, 39, 49, 63]. En este sentido, en la ICU uno de los eventos adversos de mayor interés son los AHE [71], que se definen como una condición de baja presión sanguínea que puede presentar un paciente durante un intervalo de tiempo. La identificación de este tipo de episodios es de gran importancia porque se sabe que es precursor de otros episodios de mayor gravedad, los cuales incluso pueden conducir a la muerte del paciente [35, 68, 71, 76]. Por tanto, en pacientes hospitalizados en la ICU es necesario anticipar la ocurrencia de un AHE para mejorar el diagnóstico y tratamiento mediante el uso de un CDSS que incorpore técnicas de procesamiento para la predicción de signos vitales [2, 3, 27, 33, 41, 58, 67].

Para la identificación de un AHE se requiere conocer las tres posibles medidas de presión arterial

(ABP, por sus siglas en inglés): 1) la presión sistólica (SBP, por sus siglas en inglés), la cual es la máxima presión medible cuando el corazón se contrae y la sangre comienza a ser bombeada; 2) la presión diastólica (DBP, por sus siglas en inglés) que es la presión mínima que ocurre cuando late el corazón, y 3) la presión arterial media (MAP, por sus siglas en inglés), que es una combinación de las anteriores. De acuerdo a recomendaciones en la práctica clínica, la MAP es una fuente de información más fiable en comparación con la SBP y la DBP para detectar la hipotensión [14, 78]. Por tanto, las tareas de procesamiento de AHEs comúnmente se basan en el análisis de las series de tiempo de la MAP a fin de identificar, modelar y predecir su ocurrencia.

En la presente tesis un AHE se define como un episodio que ocurre cuando al menos el 90 % de las muestras de MAP, en cualquier periodo de 30 min de una serie de tiempo correspondiente, se encuentran por debajo de un umbral de 60 mmHg [52]. A manera de ejemplo, en la Figura 1.2 se ilustra un segmento de una serie de tiempo de la MAP de una duración total de 150 minutos, considerando una frecuencia de muestreo regular de un minuto. En esta figura se ilustra gráficamente un AHE en el intervalo de tiempo de 74 minutos a 110 minutos, mientras que en el resto de la serie de tiempo no se presenta ningún episodio adverso en la MAP. Por tanto, tomando de referencia la definición del AHE, en una determinada serie de tiempo de MAP se pueden identificar las dos posibles condiciones en ésta en términos de AHE y NoAHE.

1.2 Planteamiento del problema

El desarrollo de modelos para la predicción de AHEs ha sido objeto de estudio en la literatura para su posible integración a CDSSs. El diseño y validación de dichos modelos de predicción requiere el uso de conjuntos de datos con series de tiempo de MAP. En este sentido, los esfuerzos de investigación se han centrado en identificar patrones en las series de tiempo de MAP y particularmente modelar los AHEs considerando el componente de la frecuencia y en algunos casos el componente del tiempo. Sin embargo, una de las principales desventajas de los enfoques existentes de predicción de AHEs es

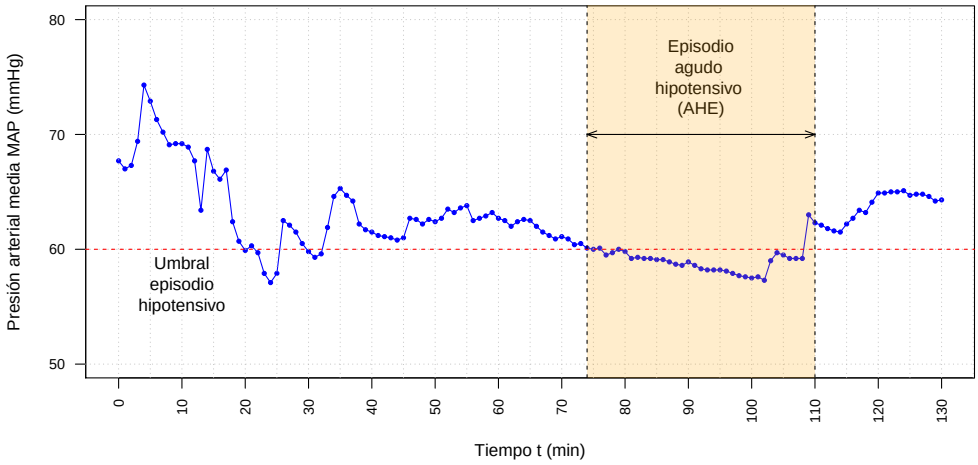

Figura 1.2: Segmento de serie de tiempo de MAP que presenta un AHE.

que no incluyen en el modelado las propiedades estocásticas de las series de tiempo de la MAP [16], lo cual puede afectar la predicción. Estos tres factores (componente de la frecuencia, componente del tiempo y propiedades estocásticas) son importantes porque se estima que su análisis en conjunto permitiría obtener información más útil que describa con mayor detalle los AHEs. En el estado del arte se encuentran algunas soluciones que desconocen el componente de tiempo y centran sus esfuerzos en el análisis del componente de la frecuencia [7, 25, 35, 46, 62]. En algunos casos se hace uso de transformada *wavelet* [45] (desarrollada inicialmente para el análisis de señales deterministas de energía finita) para preservar la información de la componente de tiempo. Sin embargo, presenta una limitante porque dicha técnica no fue desarrollada para análisis de series de tiempo con propiedades estocásticas. Por lo tanto, no se observa una solución que contemple los tres factores lo cual abre una brecha de oportunidad para el desarrollo de esta propuesta de tesis.

En este contexto, el problema de investigación abordado en la presente tesis consiste en modelar la predicción de un de AHEs mediante el uso de un algoritmo basado en cadenas de Markov debido a que esta técnica está desarrollada para modelar procesos estocásticos. Por otra parte, los otros dos factores asociados a los componentes de frecuencia y componentes de tiempo son abordados por medio de una representación que se compone de una transformación binaria, codificación y

clasificación de estados para mejorar la predicción de AHEs; el desarrollo de esta representación es importante porque se sabe que es la clave para obtener soluciones eficientes y efectivas [40].

1.3 Hipótesis

A partir de un conjunto de series de tiempo de MAP es posible establecer una codificación que permita caracterizar los AHEs contenidos en estás con fines de predicción.

1.4 Objetivos general y específicos

El objetivo general planteado para la presente tesis es el siguiente.

- Proponer un método de codificación que permita caracterizar la ausencia o presencia de los AHEs contenidos en un conjunto de series de MAP, con propósitos de generalización y predicción de dichos episodios hipotensivos.

A partir de la definición del objetivo general, se han establecido los siguientes objetivos específicos de la investigación.

- Transformar las observaciones de MAP en cadenas binarias a partir de las cuales sea posible determinar la presencia o ausencia de un AHE.

- Definir un modelo de segmentación de códigos que permita determinar cuando una cadena binaria es o no un AHE.

- Definir un modelo de generalización basado en los códigos segmentados con el propósito de predecir con un margen de antelación pre-establecido la posible ocurrencia de un AHE.

1.5 Metodología

Para alcanzar los objetivos planteados en la presente tesis, se propone la metodología de investigación que se ilustra en la Figura 1.3. A continuación se detallan cada una de las etapas consideradas en dicha metodología.

1. **Descripción de conjunto de datos.** En esta etapa se analiza el conjunto de datos de series de tiempo de MAP a utilizar para el diseño y validación del enfoque de predicción de AHEs. En particular, tomando como referencia el trabajo realizado en [28], se han seleccionado un total de 626 series de tiempo MAP del conjunto de datos.

2. **Estudio del tema de investigación.** En esta etapa se estudian los conceptos propios del dominio del contexto de la investigación y enfoques existentes que han sido propuestos en la literatura para la predicción o pronóstico de AHEs.

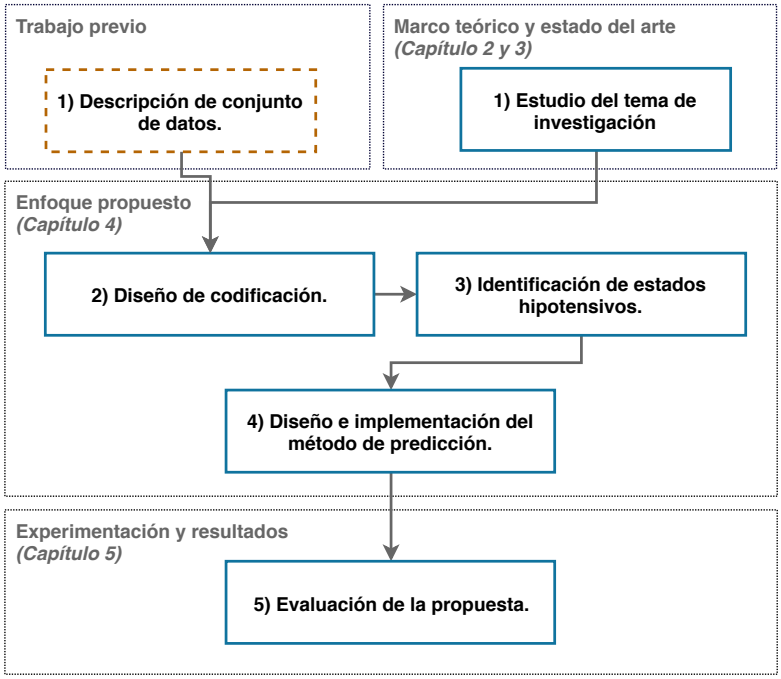

Figura 1.3: Metodología para el diseño, desarrollo y validación del método propuesto.

3. **Diseño de codificación.** El diseño de codificación compone una de las tareas de preprocesamiento. Esta tarea tiene la característica particular de que la codificación generada es una variable categórica (o cualitativa nominal) que contiene la información relevante del componente de la frecuencia y del componente de tiempo de la MAP, lo cual se prevé sea de utilidad para inferir y predecir AHEs.

4. **Identificación de estados hipotensivos.** A partir de las series de tiempo de MAP codificadas, en esta etapa se seleccionó un método para agrupar los datos categóricos resultantes. Esta agrupación de datos categóricos es necesaria para agrupar secuencias de códigos similares que puedan estar asociados a estados que representan la ocurrencia o no de un AHE.

5. **Diseño e implementación del método de predicción.** En esta etapa se plantea la integración de cadenas de Markov para la predicción de estados a futuro usando la representación de estados generados a partir de la codificación de los AHEs.

6. **Evaluación de la propuesta.** Finalmente, en esta etapa se diseñan casos de experimentación y se evalúan los resultados obtenidos durante la predicción. Esta evaluación se realizó usando métricas aceptadas y ampliamente usadas en el estado del arte para la predicción de AHE.

1.6 Organización de la tesis

La estructura de la presente tesis es la siguiente. En el Capítulo 2 se presenta el marco teórico de la investigación, el cual se enfoca en describir los conceptos fundamentales relacionados con el análisis de series de tiempo así como técnicas comúnmente usadas en el estado del arte para la predicción de AHEs. Posteriormente, en el Capítulo 3 se analiza el trabajo relacionado al problema planteado en la presente tesis. En el Capítulo 4 se detalla el diseño y desarrollo del enfoque propuesto para la predicción de AHEs en series de tiempo de MAP. En el Capítulo 5 se presentan los resultados

experimentales obtenidos para evaluar la eficiencia del método de predicción de AHEs propuesto. En el Capítulo 6 se presentan las conclusiones de la investigación y se identifican posibles líneas de trabajo a futuro de la investigación realizada. Adicionalmente, se incluyen en dos anexos información complementaria sobre el diseño de la solución propuesta.

2

Marco teórico

En este capítulo se presentan los principales conceptos fundamentales relacionados con el contexto de la tesis y el problema de investigación. Asimismo, se describen las técnicas computacionales que comúnmente se han usado en el estado del arte para la predicción de episodios agudos hipotensivos (AHEs, por sus siglas en inglés).

2.1 Monitorización de signos vitales

La monitorización de signos vitales es un procedimiento habitual que utilizan los profesionales de la salud en hospitales para conocer de forma objetiva y continua el estado de salud de los pacientes. A su vez, tal como se describió en el Capítulo 1, se prevé que la información fisiológica monitorizada pueda ser explotada por los sistemas de apoyo a la decisión clínica (CDSS, por sus siglas en inglés). A continuación se presenta una breve descripción de los principales signos vitales que son usualmente recolectados por los monitores de signos vitales en zona hospitalaria y en las unidades de cuidados intensivos (ICU, por sus siglas en inglés).

- **Presión arterial.** La presión arterial (ABP, por sus siglas en inglés) es una variable fisiológica que se define como la fuerza que ejerce la sangre contra las paredes de las arterias; dicha variable cambia en el tiempo en función de diferentes factores de comportamiento como es el caso de la frecuencia cardiaca. A partir de la ABP se puede calcular la presión arterial media (MAP, por sus siglas en inglés) tomando como referencia los valores máximos y mínimos, lo cual se le conoce como presión arterial sistólica (SBP, por sus siglas en inglés) y presión arterial diastólica (DBP, por sus siglas en inglés), respectivamente. Tal como se explica en el Capítulo 1, la combinación de la SBP y la DBP permite la definición de la presión arterial media (MAP, por sus siglas en inglés). La MAP es la variable que comúnmente se usa para la detección o predicción de AHEs, la MAP se calcula a partir de la siguiente ecuación [7, 35, 45, 46, 62]:

$$MAP = \frac{SBP + 2 * DBP}{3} \tag{2.1}$$

- **Frecuencia cardíaca.** La frecuencia cardíaca es el número de veces que el corazón realiza el ciclo cardíaco completo (SBP y DBP) y se mide en contracciones por minuto, ya que cuando nos tomamos el pulso lo que notamos es la contracción del corazón (SBP). Es decir, cuando el corazón alcanza el pico de contracción mientras expulsa la sangre hacia el resto del cuerpo [66]. La frecuencia cardíaca es también una señal no estacionaria, donde su variación puede contener indicadores de enfermedades actuales o advertencias sobre enfermedades cardíacas inminentes. Los indicadores pueden estar presentes en todo momento o pueden ocurrir de forma estocástica, durante ciertos intervalos.

- **Electrocardiograma.** Un electrocardiograma (ECG) es una herramienta de diagnóstico que mide y registra la actividad eléctrica del corazón en función del tiempo [17]. Un ECG también se le denomina a la información contenida en una tira de papel o en una línea en una pantalla que produce un dispositivo de ECG. Esta información suele ser representada como una serie de tiempo que contiene un conjunto de información de la medición de la actividad eléctrica

del corazón en un tiempo determinado. Usualmente, esta información es interpretada por un médico quien determina si hay alguna actividad anormal o inusual analizando el trazo electrocardiográfico, los picos y las caídas de la serie de tiempo resultante.

- **Saturación de oxígeno en la sangre.** La saturación de oxigeno en la sangre (SaO2, por sus siglas en inglés) es una unidad de medida de la cantidad de oxígeno disponible en el torrente sanguíneo. Esta unidad de medida registra la cantidad de moléculas de oxigeno presente en los glóbulos rojos. La sangre que contiene glóbulos rojos se oxigena en los pulmones, donde las moléculas de oxígeno viajan desde el aire hacia la sangre y se combina con la hemoglobina presente en los glóbulos rojos con el fin de ser llevado al resto del cuerpo cuando el corazón bombea la sangre. El porcentaje de glóbulos rojos o eritrocitos que están completamente saturados con oxígeno se conoce como saturación arterial de oxígeno o nivel de oxígeno en sangre (AaO2) [12]. El valor de SaO2 puede ser aproximada de forma no invasiva, rápida y de forma continua a través de un oxímetro de pulso [4, 74], en este caso se le conoce como saturación de oxígeno capilar periférica (SpO2, por sus siglas en inglés).

Es importante clarificar que cada signo vital cuenta con valores de referencia recomendados como normales, los cuales son utilizados por los profesionales de la salud para diagnosticar el posible deterioro de salud de un paciente o la ocurrencia de un episodio adverso. En este sentido, en aras de integrar una funcionalidad de predicción en un CDSS, es necesario disponer de un modelo que caracterice el comportamiento normal de un signo vital en particular. En el caso de interés para la presente tesis, el umbral (valor de referencia) de la presión arterial media MAP que comúnmente se usa para determinar la ocurrencia de un AHE se establece en 60 mmHg.

2.2 Series de tiempo de signos vitales

Los signos vitales de un paciente se registran de forma continua durante el periodo de estancia en un hospital, para lo cual se puede obtener una serie de tiempo por cada signo vital. Una serie

de tiempo es un conjunto de muestras obtenidas en la monitorización, las cuales se encuentran ordenadas de forma sucesiva. Formalmente, una serie de tiempo X es un conjunto de variables X_t donde cada una ocurre de forma aleatoria en un tiempo específico t [13]. Cada una de estas variables X_t puede ser representada básicamente como las observaciones de un signo vital en el tiempo t. Las series temporales que son objeto de estudio en la presente tesis son aquellas en las que el número de muestras de las observaciones es discreto [1]. Tal es el caso de la MAP, la cual es una serie de observaciones en tiempo discreto que se registra en bases de datos o que se recolecta en tiempo real de un monitor de signos vitales.

Las series de tiempo se utilizan en estadística, procesamiento de señales, predicción del clima, comunicaciones y, en gran medida, en cualquier área de la ciencia aplicada e ingeniería donde se requieran mediciones temporales. Esto es posible al analizar las series de tiempo que resultan de la observación temporal de un evento. El análisis de series de tiempo es una disciplina que estudia los métodos que permiten describir la evolución temporal de los datos con el fin de extraer información como patrones o valores estadísticos que detallen su comportamiento.

El análisis de series de tiempo no es una tarea fácil porque éstas tienen un orden temporal y natural. Esto hace que el análisis de series de tiempo sea distinto de otros estudios en los que no existe un orden natural de las observaciones. Este caso se observa, por ejemplo, durante el análisis de datos espaciales donde las observaciones generalmente se relacionan con ubicaciones geográficas [37]. En particular, el análisis de series de tiempo se puede aplicar a datos continuos de valores reales, datos numéricos discretos o datos categóricos discretos. Los signos vitales representados como series de tiempo tienen las siguientes propiedades:

- **Tiempo discreto.** Las muestras de las series de tiempo se encuentran en un tiempo discreto y los registros son realizados a intervalos regulares $t = \{1, 2, \ldots, n\}$, expresados en unidades de tiempo como segundos, minutos, etc.

[1] En la presente tesis usa de forma indistinta del concepto de series de tiempo para hacer referencia a las series de tiempo discreto.

- **Orden natural.** Esta propiedad define que las series de tiempo están ordenadas de una forma natural cronológicamente. Esta propiedad es muy importante ya que determina el tipo de relaciones posibles entre una muestra y otra en una ventana de tiempo determinada.

A manera de ejemplo, en la Figura 2.1 se muestra una serie de tiempo en la cual el eje de las abscisas contiene la evolución de la muestra en el tiempo t y en el eje de las ordenadas se muestra el intervalo y los valores que toman las muestras correspondientes. En particular, se ilustra una serie de tiempo de MAP cuantificada en unidades de milímetros de mercurio (mmHg) y registrada a intervalos regulares de un minuto.

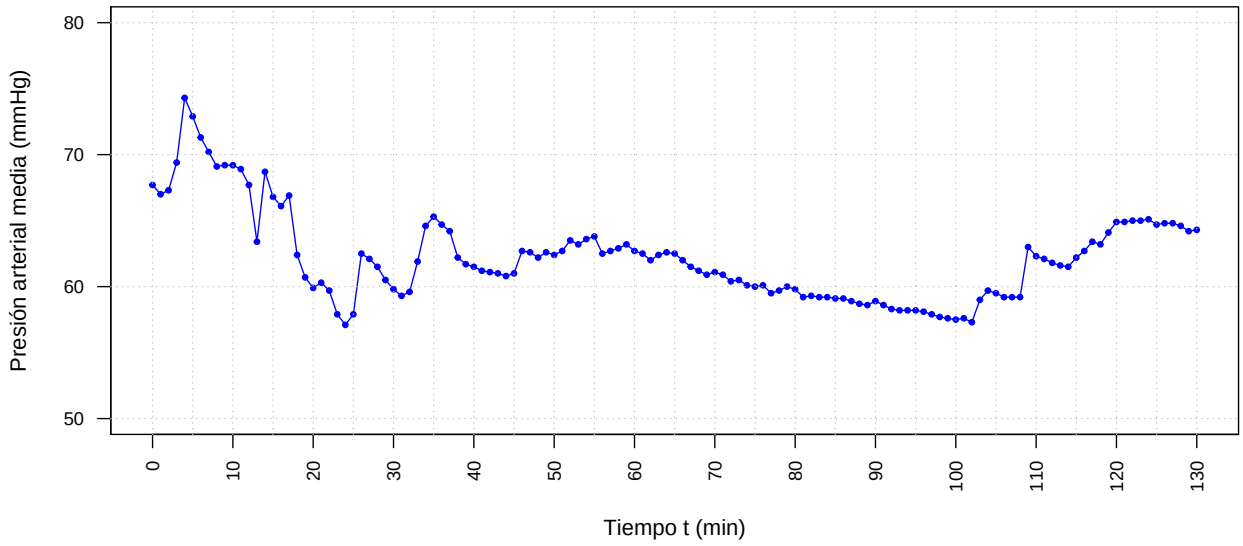

Figura 2.1: Ejemplo de una serie de tiempo de MAP.

En una serie de tiempo de signos vitales pueden presentarse cuatro diferentes componentes de variabilidad, las cuales se ilustran en la Figura 2.2 y se detallan a continuación [75].

- **Tendencia secular.** Una tendencia secular es una variable relacionada con un patrón consistente dentro de un periodo de tiempo determinado. Es una tendencia estadística que no

presenta efecto estacional ni cíclico. Esta tendencia puede identificarse a lo largo del tiempo, lo cual permite desarrollar pronósticos y predicciones de comportamiento [2].

- **Variación estacional.** Es un elemento variable en el análisis de las series de tiempo, el cual se refiere al fenómeno en el que una muestra cambia siguiendo cierta tendencia estacional según las características de la misma. Dichas variaciones pueden presentarse en intervalos regulares específicos. La variación estacional puede ser causada por varios factores externos y consiste en patrones periódicos, generalmente regulares y predecibles en los niveles de una serie de tiempo.

- **Variación cíclica.** Es un elemento variable en el análisis de las series de tiempo, el cual se refiere al fenómeno en el que una muestra cambia sin seguir una tendencia estacional. Existe un patrón cíclico cuando los datos muestran subidas y caídas que no son de un período fijo. Usualmente se desconoce de antemano la duración del ciclo actual. La variación cíclica es diferente de la variación estacional porque si las fluctuaciones no son de periodo fijo entonces son cíclicas, por otro lado, si el periodo no cambia y está asociado con algún aspecto del calendario, entonces el patrón es estacional.

- **Variación irregular.** Como su nombre lo indica, estas variaciones son de patrón irregular e indefinido. En general, se mezclan con variaciones estacionales y cíclicas ocasionadas por factores estocásticos o aleatorios. Estas variaciones también se denominan erráticas, accidentales o aleatorias. Incluyen todos los tipos de variaciones en una serie de tiempo que no son atribuibles a las fluctuaciones de tendencia secular, variación estacional y variación cíclica.

Si bien el análisis de series de tiempo consiste principalmente en extraer y evaluar las componentes descritas, se deben considerar las propiedades estocásticas de las series de tiempo de signos vitales. Para ello se requiere de un análisis de variación irregular, el cual es más complejo y es diferente al análisis de tendencia secular, variación estacional y variación cíclica [57]. Por otra parte, el análisis

[2] En otros ámbitos como el financiero, las tendencias seculares se pueden identificar en variables como el producto interno bruto, la inflación, los precios de las acciones, los índices de acciones y otros indicadores evaluados por los analistas para desarrollar un pronóstico que permita a los inversionistas tomar una decisión.

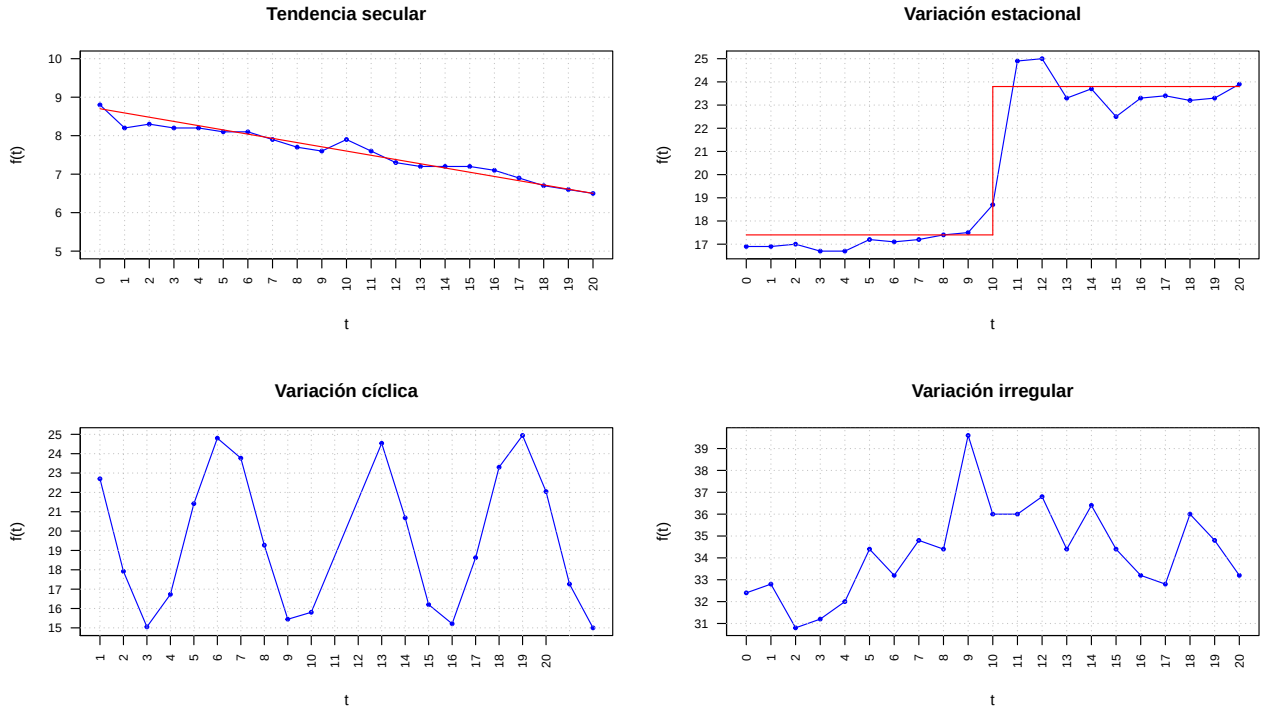

Figura 2.2: Componentes de variabilidad en una serie de tiempo.

de la variación irregular aporta información del componente de la frecuencia pero no del componente del tiempo. Por tal motivo, el análisis de series de tiempo de signos vitales requiere el uso de métodos que permitan describir las propiedades estocásticas, el componente de la frecuencia y el componente del tiempo.

2.3 Bases de datos especializadas

El diseño e implementación de métodos computacionales para caracterizar y predecir un AHE requiere de datos históricos de signos vitales de pacientes. Por ello, varios grupos de investigación se han dado a la tarea de recolectar datos clínicos y fisiológicos generando conjuntos de bases de datos especializadas. Generalmente, estas bases de datos son de acceso público y se encuentran disponibles para su descarga, como es el caso de las bases de datos de signos vitales: 1) CapnoBase [38], 2)

BrainIT [59], y 3) MIMIC-II [44].

En el contexto de predicción de AHEs, la base de datos de signos vitales ampliamente usada en la literatura es la denominada Monitoreo Inteligente Multiparámetro en Cuidados Intensivos (MIMIC, por sus siglas en inglés) en su versión II [3]. Esto se debe a que la MIMIC-II dispone de series de tiempo de SBP, DBP y MAP. Dicha base de datos es un conjunto de datos abiertos disponible recolectado por el Laboratorio para Fisiología Computacional en el Instituto Tecnológico de Massachusetts. La base de datos MIMIC-II contiene datos clínicos de la cabecera del paciente y signos vitales capturados de monitores de pacientes. Los datos en MIMIC-II fueron recolectados entre 2001 y 2008 de una variedad de ICU (médicas, quirúrgicas, atención coronaria y neonatal) en un hospital universitario. La información contenida en MIMIC-II se puede clasificar en:

- **Datos clínicos.** Son datos demográficos, resultados de pruebas de laboratorio, medicamentos suministrados a los pacientes, alergias, historial clínico familiar, etc.

- **Signos vitales.** Son recolectados a partir de monitores de signos vitales ubicados en la cabecera del paciente que registran variables fisiológicas como presión arterial, temperatura corporal, pulso, frecuencia respiratoria, etc.

En particular, los datos fisiológicos contenidos en la base de datos de MIMIC-II son: electrocardiograma, frecuencia cardíaca, MAP y saturación de oxígeno [15, 44]. En esta tesis se usa un conjunto de datos de la base de datos MIMIC-II consistente en 626 series de tiempo de MAP, las cuales fueron preprocesadas en [28]; una mayor información de este conjunto de datos se puede consultar en el Capítulo 5.

[3] Adicionalmente, existe la versión III de la MIMIC, enfocada principalmente a proveer una interfaz web para la manipulación y descarga de series de tiempo de signos vitales [36].

2.4 Principales técnicas usadas en la predicción de AHEs

A continuación se enuncian y describen las principales técnicas utilizadas en el estado del arte para la predicción de AHEs.

2.4.1 Aprendizaje supervisado

El objetivo del aprendizaje supervisado es construir un modelo de la distribución de etiquetas de clase usando el conocimiento *a priori* (o características predictoras). El modelo resultante se usa para asignar etiquetas de clase a las instancias desconocidas donde se conocen los valores de las características del predictor, pero se desconoce el valor de la etiqueta de clase.

2.4.1.1. Regresión lineal

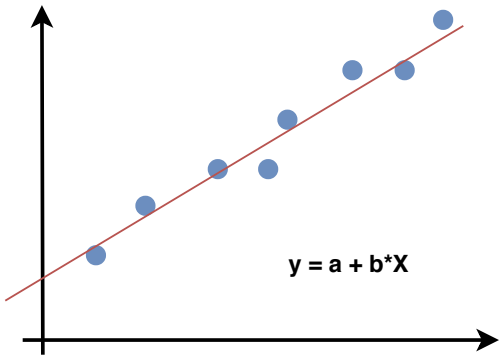

Figura 2.3: Representación de una regresión lineal.

La regresión lineal es una técnica estadística de aproximación de dos variables. Formalmente, la regresión lineal es una aproximación usando una recta que se representa como una ecuación lineal $y = a + b * X$ donde los coeficientes a y b definen la recta. El coeficiente b es la pendiente de recta y el coeficiente a es el punto de la recta con el que corta el eje de la ordenada. El análisis de regresión lineal consiste en encontrar los coeficientes a y b que mejor se ajusten a la relación entre

las dos variables. Una representación gráfica se muestra en la Figura 2.3 donde se aprecia una línea que aproxima a la relación existente entre dos variables [50].

2.4.1.2. k vecinos más cercanos

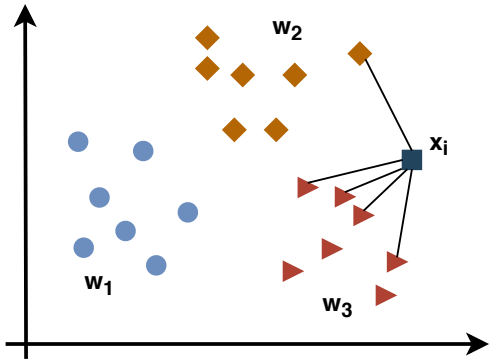

Figura 2.4: Ejemplo de agrupación mediante $k-$NN ($k = 3$).

El enfoque conocido como k vecinos más cercanos ($k-$NN, por sus siglas en inglés) es un algoritmo de clasificación que asigna una etiqueta a un nuevo elemento basado en una función de similitud. A fin de explicar el principio de operación de este algoritmo considere la Figura 2.4, en la que se ilustran diversos elementos agrupados en tres clases $\{w_1, w_2, w_3\}$ y un nuevo elemento X_i. Cada elemento perteneciente a una clase así como el nuevo elemento están identificados con formas diferentes (circulo, rombo, triángulo y cuadrado, respectivamente). En este caso el nuevo elemento se asocia a un grupo usando una medida de similitud de distancia euclideana con los $k = 5$ vecinos más cercanos y por voto mayoritario se asigna a la clase w_3 debido a que el nuevo elemento X_i se encuentra más cerca de cuatro elementos de dicha clase.

2.4.1.3. Máquina de vectores de soporte

Una máquina de vectores de soporte (SVM, por sus siglas en inglés) es un método que construye una solución donde el hiperplano óptimo es aquel con la máxima distancia $z = z^+ + z-$. Donde z^+ y z^- son las márgenes que separa al hiperplano de los datos más cercanos. A dichos márgenes se le

denominan vectores de soporte. En la Figura 2.5 se ilustra el hiperplano óptimo en dos dimensiones representados por una recta, también se observa las márgenes z^+ y z^- que separan los vectores de soporte.

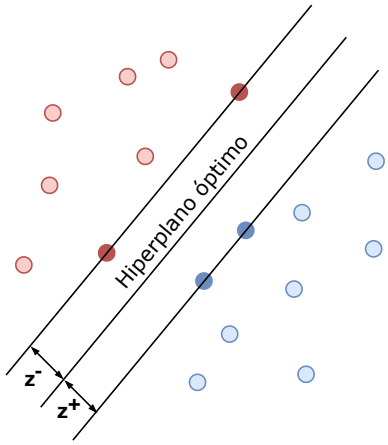

Figura 2.5: Hiperplano de separación óptima usando SVM.

2.4.1.4. Redes neuronales artificiales

Las redes neuronales artificiales (ANNs, por sus siglas en inglés) constan de diversos procesos simples y conectados llamados neuronas, cada uno de los cuales produce una secuencia de activaciones. Tal como se ilustra en la Figura 2.6, las neuronas en una ANN están organizadas en capas donde se observa su estructura general y las interacciones de las neuronas representadas como flechas que conectan a las neuronas. Las neuronas de la capa de entrada se activan a través de estímulos representados como valores reales, mientras que otras neuronas se activan a través de conexiones ponderadas de neuronas previamente activadas como es el caso de la capa oculta o de la capa de salida. El aprendizaje o la asignación de un valor real consiste en encontrar pesos que hagan que la ANN muestre el comportamiento deseado. Dependiendo del problema y de cómo estén conectadas las neuronas, tal comportamiento puede requerir largas cadenas causales de etapas computacionales conocidas como capas, donde cada capa transforma (a menudo de forma no lineal) la activación agregada de la red. En este contexto, el aprendizaje profundo consiste en asignar con

precisión los valores reales o pesos en muchas de estas neuronas en las diferentes capas [65].

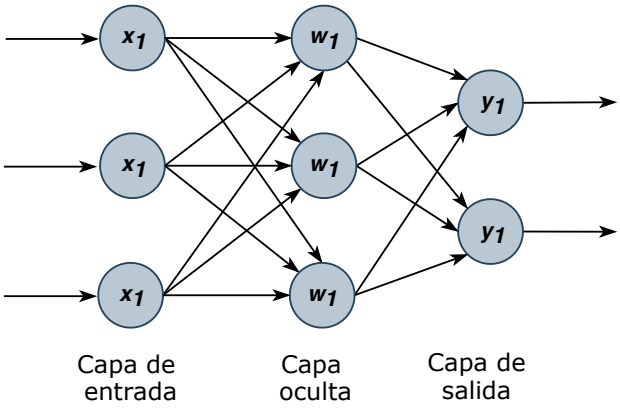

Figura 2.6: Estructura general de una red neuronal artificial.

2.4.1.5. Redes Bayesianas

Es un modelo probabilístico de variables aleatorias que puede ser representado tal como se ilustra en la Figura 2.7 usando un grafo acíclico dirigido donde los nodos son las variables aleatorias y cada una de las aristas o arcos son las relaciones probabilísticas entre las variables. Esta dependencia refleja de forma numérica la relación de causa y efecto entre dos variables. A partir de una observación o de valores hipotéticos se puede entrenar una red bayesiana que consiste en el proceso de plasmar las variables aleatorias (nodos) y sus dependencias (aristas) en una red que represente el sistema a modelar. Esta red puede ser usada finalmente para inferir nuevos valores a partir del conocimiento adquirido.

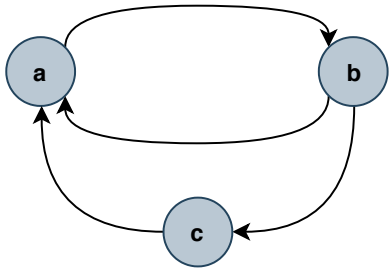

Figura 2.7: Grafo acíclico dirigido de una red bayesiana.

2.4.2 Aprendizaje no supervisado

EL objetivo del aprendizaje no supervisado es la creación de un modelo con base en las observaciones sin conocimiento *a priori*, es decir sin la información previa de las etiquetas de clases de las observaciones. Como no se tiene conocimiento *a priori* el aprendizaje no supervisado es usado principalmente para describir la estructura de los datos de entrada encontrando algún tipo de organización que apoye el análisis de los mismos. La construcción del modelo se realiza encontrando similitudes entre las características que describen diferentes puntos de datos.

2.4.2.1. Agrupamiento mediante $k-$means

El algoritmo $k-$means es un método de agrupamiento que dado un conjunto de n observaciones se divide en k grupos usando una medida de distancia para asociar un punto de datos. $k-$means requiere que se indique *a priori* un número de grupos. Considere la Figura 2.8 donde se ilustran dos grupos ($k = 2$) y su respectivo centroide que los agrupan. Una vez determinado se siguen los siguientes pasos:

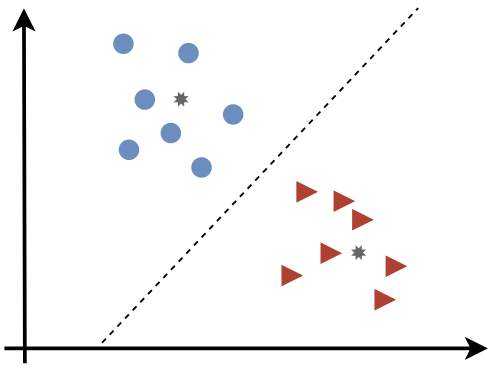

Figura 2.8: Ejemplo de agrupación mediante $k-$means ($k = 2$).

1. Dado un número de grupos k se inicializan aleatoriamente sus respectivos puntos centroides. Los centroides son vectores de la misma longitud que cada vector de punto de datos, es decir que las dimisiones son iguales.

2. Cada punto de datos se clasifica calculando la distancia entre éste punto y cada centro de grupo, y luego se clasifica el punto de datos calculando la distancia al centro más cercano.

3. Con base a estos puntos clasificados, se recalcula el centro del grupo tomando la media de todos los vectores en el grupo.

4. Finalmente, se repiten estos pasos para un número determinado de iteraciones o hasta que los centros de grupo no cambien mucho entre iteraciones.

2.4.2.2. Agrupamiento jerárquico

Los algoritmos de agrupamiento jerárquico se dividen en dos categorías: algoritmos de división o algoritmos aglomerativos. Los algoritmos de agrupación jerárquica usualmente no requieren que se especifique la cantidad de agrupaciones. Los aglomerativos tratan cada punto de datos p como un solo grupo al principio y luego combinan sucesivamente en grupos pares hasta que todos los grupos se hayan fusionado en un solo grupo que contiene todos los puntos de datos. En los algoritmos de división todos los puntos de datos p son inicialmente un grupo y luego se realiza sucesivamente divisiones mientras se desciende en jerarquía. En ambos casos la agrupación de los puntos de datos p se da por una medida de distancia o similitud.

En la Figura 2.9, se ilustra un ejemplo de una comparativa de la aplicación de los algoritmos de división y aglomerativo. En este caso en la región izquierda se observan los puntos de datos $p = \{p1, p2, p3, p4\}$ y la agrupación realizada en cada uno de los tres pasos, en la región derecha se observa un árbol o dendograma que muestra la agrupación de los puntos de datos en los diferentes niveles para ambos algoritmos (aglomerativos y de división). Usando el algoritmo aglomerativo obtenemos que los elementos se agrupan en el primer nivel de la siguiente forma $\{p1, p2\}$ y $\{p3, p4\}$, mientras que con el algoritmo de división se obtienen los grupos $\{p1, p2, p3\}$ y $\{p4\}$. Finalmente, la decisión de usar uno u otro algoritmo depende del dominio del problema.

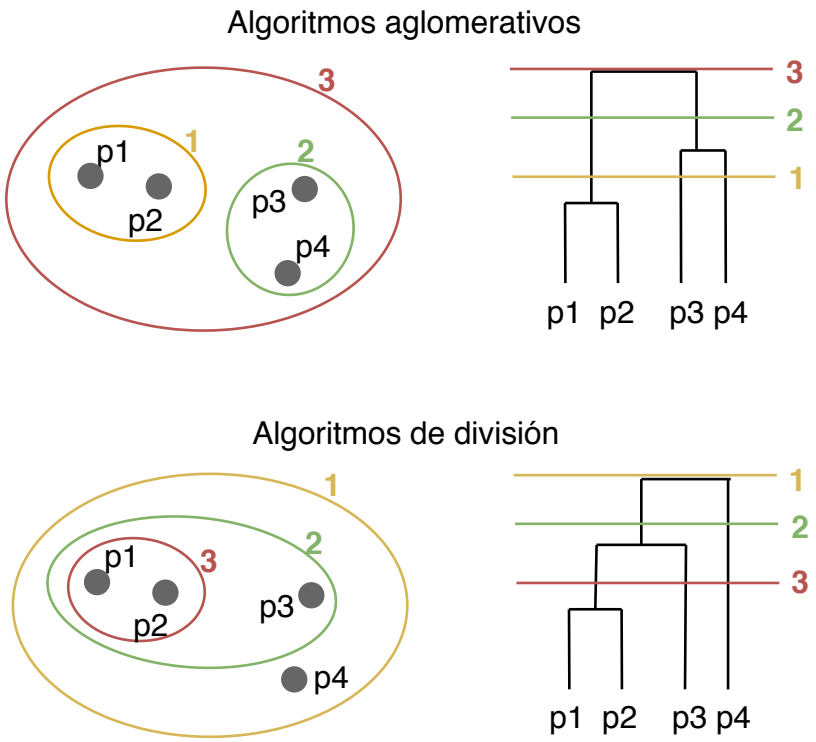

Figura 2.9: Algoritmos de agrupamiento jerárquico.

2.4.3 Cadenas de Markov

Formalmente, una cadena de Markov es un proceso en tiempo discreto en el que una variable aleatoria X_n cambian su valor con el paso del tiempo. Esta variable aleatoria puede, por ejemplo, representar un signo vital dado una serie de observaciones. Una cadena de Markov tiene la propiedad de que $X_n = j$ solo depende del estado inmediatamente anterior X_{n-1}. Esta propiedad de la cadena no depende del tiempo n en que se considere, esto es, la probabilidad son las mismas en cada paso como se expresa en la Ecuación 2.2.

$$P(X_n = j | X_{n-1} = i) \tag{2.2}$$

En una cadena homogénea finita X con m posibles estados S_m tal que $S_n \in X$ se puede introducir la notación,

$$p_{pj} = P(X_n = j | X_{n-1} = i),$$

Donde $i, j = 1, 2, \ldots, m$. Si $p_{ij} > 0$ entonces se dice que el estado S_i tiene una relación de causa y efecto con el estado S_j. La relación puede ser mutua si también existe una probabilidad de transición del estado $p_{ji} > 0$. Para cada i fijo, la serie de valores $\{p_{ij}\}$ es una distribución de probabilidad, ya que en cualquier paso puede ocurrir alguno de los sucesos $S = \{s1, s2, \ldots, s_m\}$ y son mutuamente excluyentes. Los valores p_{ij} se denominan probabilidades de transición si satisfacen la condición

$$p_{ij} > 0, \sum_{j=1}^{m} p_{ij} = 1, \tag{2.3}$$

Estas probabilidades de transición p y los estados S_m pueden ser representados usando un grafo dirigido como se ilustra en la Figura 2.10, donde los nodos representan un estado y las aristas o flechas representan una probabilidad de transitar de un estado en el tiempo i a un estado el tiempo j para cada $i, j = \{1, 2, \ldots, m\}$.

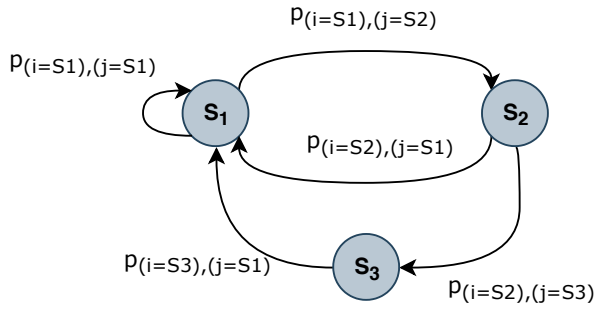

Figura 2.10: Grafo dirigido de cadenas de Markov.

Usando una representación de una matriz de filas y columnas, estos valores se combinan formando una matriz de transición A de tamaño m x m [47], donde

$$A = [p_{ij}] = \begin{bmatrix} p_{11} & p_{12} & \dots & p_{1m} \\ p_{21} & p_{22} & \dots & p_{2m} \\ \dots & \dots & \dots & \dots \\ p_{m1} & p_{m2} & \dots & p_{mm} \end{bmatrix} \tag{2.4}$$

Se puede observar que cada fila de la matriz es una distribución de probabilidad, es decir, $\sum_{j=1}^{m} p_{ij} = 1$.

Probabilidad a n pasos

Una probabilidad de bastante interés es la probabilidad de llegar a S_j después de n pasos, dada una distribución de probabilidad $\{p_i^{(0)}\}$. Se observa que $\{p_i^{(0)}\}$ es la probabilidad de que el sistema ocupe inicialmente el estado S_i, de modo que

$$\sum_{i=j}^{m} p_i^{p_{ij}}$$

Esto se puede expresar de forma vectorial: sea $P^{(0)}$ y $P^{(1)}$ los vectores de probabilidad dados por

$$P^{(0)} = (p_1^{(0)}, \dots, p_m^{(0)})$$

y

$$P^{(1)} = (p_1^{(1)}, \dots, p_m^{(1)}),$$

donde $P^{(0)}$ es la distribución de probabilidad inicial y $P^{(1)}$ es la probabilidad de que alcance cada uno de los estados $S_1, \dots, S_m$ después de un paso. Con esta notación se puede expresar

$$P^{(1)} = [p_{ij}^{(1)}] = [\sum_{i=1}^{m} p_i(0) p_{ij}] = P^{(0)} A,$$

donde A es la matriz de transición.

Del mismo modo,

$$P^{(2)} = P^{(1)}A = P^{(0)}A^2$$

y en n pasos,

$$P^{(n)} = P^{(n-1)}A = P^{(0)}A^n \tag{2.5}$$

Donde $P^{(n)} = P^{(0)}A^n$ es la ecuación para calcular la probabilidad de transitar desde un estado inicial $P^{(0)}$ dado una matriz de transición A^n en n pasos. [47]

2.5 Métricas de clasificación y predicción

En esta sección se presentan las métricas disponibles para determinar el desempeño de agrupación, clasificación y predicción en el contexto de problemas de análisis de datos. En particular, se presentan la matriz de confusión como una herramienta para el cálculo de las métricas de coeficiente de correlación de Matthews, exactitud, tasa de error, sensibilidad y especificidad.

Matriz de confusión

Una matriz de confusión es una herramienta que permite medir el desempeño de un algoritmo cuando se desea evaluar el valor predicho sobre el valor real. Esta medición se representa en una matriz cuadrada que cuantifica los aciertos o desaciertos obtenidos. Cada columna de la matriz representa el número de clasificaciones o predicciones para cada clase, mientras que cada fila representa los valores reales o etiquetas *a priori*. A continuación se presenta la matriz de confusión (B) para una clasificación binaria donde una clase es etiquetada como positiva y otra clase como negativa.

$$B = \begin{pmatrix} TP & FN \\ FP & TN \end{pmatrix}$$

- **Verdaderos positivos (TP).** Es el número de aciertos para la clase positiva.
- **Verdaderos negativos (TN).** Es el número de aciertos para la clase negativa.
- **Falsos positivos (FP).** Es el número de errores resultado de asignar la clase negativa a la clase positiva.
- **Falso negativos (FN).** Es el número de errores resultado de asignar la clase positiva a la clase negativa.

Con base en la matriz de confusión es posible calcular las siguientes métricas:

Coeficiente de correlación de Matthews

La métrica de coeficiente de correlación de Matthews (MCC, por sus siglas en inglés) fue introducida por primera vez por B.W. Matthews para evaluar el rendimiento de la predicción de la estructura secundaria de proteínas [48]. El MCC determina si existe una relación entre dos variables dicotómicas es decir, variables para las cuales sólo es posible considerar dos clases. El MCC es una unidad de medida ideal para la medición de clases desbalanceadas, esto ocurre en nuestro caso porque el tamaño de las muestras AHE y NoAHE se distribuyen de forma desigual lo cual es común en muchas aplicaciones bio-informáticas [18, 69, 72]. El MCC puede interpretarse como una discretización de la correlación de Pearson para variables binarias [10]. A continuación se presenta la ecuación para el cálculo del MCC:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \tag{2.6}$$

Exactitud

La exactitud (ACC, por sus siglas en inglés) es la tasa de aciertos globales de un clasificador. Esta unidad de medida cuantifica la proporción de verdaderos positivos y verdaderos negativos sobre el total de aciertos o desaciertos de la clase positiva y negativa. A continuación se presenta la función para el cálculo de la exactitud:

$$ACC = \frac{TP + TN}{TOTAL} \tag{2.7}$$

Tasa de error

La tasa de error (ERRORRATE) es la tasa de desaciertos globales de un clasificador. Esta unidad de medida cuantifica la proporción de falsos positivos y falsos negativos sobre el total de aciertos o desaciertos de la clase positiva y negativa.

$$ERRORRATE = \frac{FP + FN}{TOTAL} \tag{2.8}$$

Sensibilidad

La sensibilidad (SEN, por sus siglas en inglés) mide la proporción de positivos reales que se identifican correctamente como tales sobre el total de positivos reales identificados a *priori*. Esto permite medir la capacidad del clasificador de acertar los casos positivos. A continuación se presenta la función para el cálculo de la sensibilidad:

$$SEN = \frac{TP}{TP + FN} \tag{2.9}$$

Especificidad

La especificidad (SPE, por sus siglas en inglés) mide la proporción de negativos reales que se identifican correctamente como tales sobre el total de negativos reales identificados a *priori*. Esto permite medir la capacidad del clasificador de acertar los casos negativos. A continuación se presenta la función para el cálculo de especificidad:

$$SPE = \frac{TN}{TN + FP} \tag{2.10}$$

2.6 Resumen

En este capítulo se han presentado las características y propiedades de los diferentes signos vitales que son usualmente recolectados por los monitores de signos vitales en las ICU. También se detallaron las propiedades de las series de tiempo de signos vitales y los diferentes componentes de variabilidad. Seguido se presentaron las bases de datos especializadas con registros de signos vitales de pacientes de la ICU. A continuación se describieron algunas técnicas abordadas por diferentes autores en el estado del arte para la predicción de AHEs . Por último, se presentaron las métricas de clasificación y predicción más usadas para la evaluación de los resultados obtenidos en el dominio del tema.

3

Estado del Arte

En este capítulo se presenta una revisión del estado del arte relacionado con el problema de investigación planteado en la presente tesis. En particular, se analizan los diferentes enfoques que han sido propuestos para predecir los eventos agudos hipotensivos (AHEs, por sus siglas en inglés).

3.1 Introducción

La resolución del problema de predicción de AHEs generalmente considera que la serie de tiempo de MAP sea segmentada en diferentes ventanas sucesivas. Una ventana de serie de tiempo es un intervalo de tiempo discreto predefinido con algún propósito. En este sentido, existen dos tipos de enfoques de segmentación que han sido utilizados:

- **Sin intervalo entre ventanas.** Una serie de tiempo se divide en dos segmentos: observación y predicción. El tamaño de cada ventana se da en minutos para la ventana de observación O y para la ventana de predicción P.

- **Con intervalo entre ventanas.** Una serie de tiempo se divide como se mencionó anteriormente, pero hay una brecha o intervalo entre ventanas de G minutos entre las ventanas de observación y predicción. La introducción de este intervalo entre ventanas es una predicción más desafiante porque se realiza en un futuro próximo y no inmediato, con el objetivo de predecir un evento adverso con al menos G tiempo antes de que ocurra.

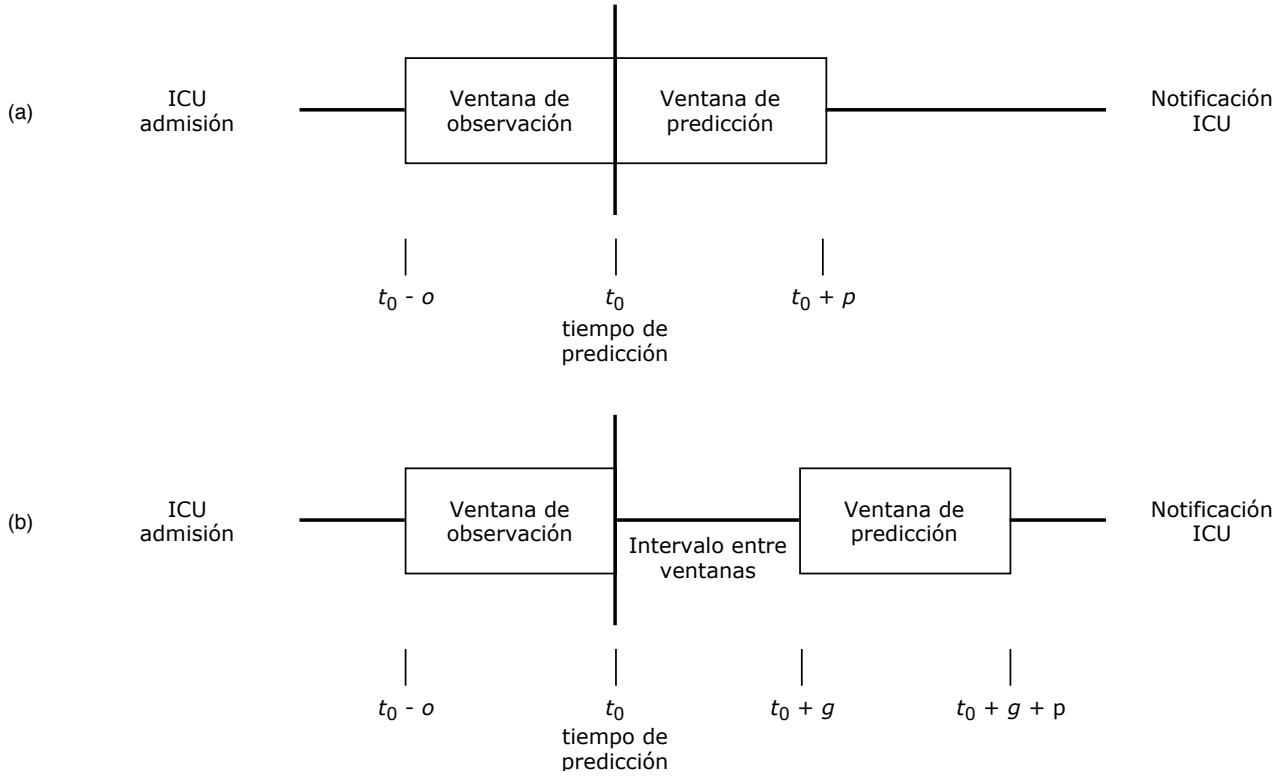

Figura 3.1: Enfoques de segmentación de series de tiempo MAP: sin y con intervalo entre las ventanas de observación y predicción.

La Figura 3.1 representa la diferencia entre los enfoques sin intervalo entre ventanas (a) y con intervalo entre ventanas (b). Los modelos predictivos generalmente se desarrollan siguiendo uno de estos enfoques. En ambos casos, la idea es explotar los datos disponibles dentro de la ventana de observación para predecir la ocurrencia de un evento adverso en la ventana de predicción. La ventana de observación comprende diferentes tipos de datos recopilados durante o minutos, incluidos datos fisiológicos y, si están disponibles, también datos clínicos. En el caso de predicción temprana de AHEs,

la ventana de observación se compone comúnmente de mediciones de MAP registradas durante o minutos. Los datos recopilados dentro de la ventana de observación se utilizan como entrada para un modelo predictivo, mientras que los datos dentro de la ventana de predicción se usa normalmente para calcular una métrica de precisión que denota el rendimiento del modelo predictivo.

Tanto en los enfoques sin intervalo entre ventana y con intervalo entre ventanas, un evento adverso podría estar presente durante la ventana de predicción. Sin embargo, el tiempo de inicio (t) de dicho evento adverso es generalmente desconocido. Considere el enfoque de intervalo entre ventanas, donde podría ocurrir un evento adverso en cualquier momento durante la ventana de predicción, por ejemplo, en el medio de la ventana de predicción ($t = t_0 + p/2$) o al comienzo de la ventana misma ($t = t_0$). Considerando el último caso, la predicción se haría justo después de que finalice la ventana de observación como se observa en la Figura 3.1 (a). Esto podría no ser útil en entornos prácticos porque no hay tiempo suficiente para tomar una acción por adelantado. En otras palabras, el valor pronóstico de un modelo predictivo basado en un enfoque sin intervalo entre ventanas podría depender del tiempo de inicio t del evento adverso. Para aliviar este problema, es posible extender el enfoque anterior de segmentación sin intervalo entre ventanas al incluir una brecha entre las ventanas de observación y predicción. De esta manera, un modelo predictivo basado en un enfoque con intervalo entre ventanas podría ser más útil en entornos prácticos al identificar un evento adverso con g minutos de anticipación como se muestra en la Figura 3.1 (b). Por lo tanto, el valor pronóstico de un modelo predictivo con base en un intervalo entre ventanas depende del tamaño del intervalo, independientemente de si el tiempo de inicio de un evento adverso es $t = t_0$.

Es intuitivo pensar que el uso explícito de un enfoque con intervalo entre ventanas puede conducir a predicciones anteriores, pero no necesariamente más precisas. Es decir, si comparamos un modelo predictivo desarrollado utilizando ambos enfoques de segmentación, podríamos esperar que se puedan obtener mejores resultados (en términos de precisión) utilizando un enfoque de segmentación Sin intervalo entre ventanas. Como el tiempo es un recurso valioso para la intervención temprana en la UCI, es de suma importancia desarrollar enfoques de segmentación basados en intervalo entre

ventanas capaces de proporcionar predicciones más tempranas de AHEs con alto valor pronóstico. Por lo tanto, es importante poder predecir AHEs antes de que ocurran de manera oportuna y aumentar las oportunidades de mejora del diagnóstico y tratamiento de pacientes [14].

Esto plantea un interrogante sobre el tamaño de la ventana o intervalo que se debe considerar entre la observación y predicción. Son varios los estudios que sugieren la ventaja de una predicción más temprana [14, 20, 55] pero son pocos los que abordan este problema. Donald, R. *et al.* [20] detalla que su método basado en una red bayesiana crea grupos estrechamente relacionados lo que dificulta la predicción de AHEs en un intervalo menor a 15 minutos. Por otra parte, en [27] se describe que un aumento en el número de muestras de signos vitales recolectados a intervalos más regulares permite que el algoritmo capture patrones que caracterizan más fluctuaciones de la presión arterial mejorando la calidad de las predicciones. En el mismo estudio se sugiere que otros patrones como número de símbolos y longitud de la serie de tiempo afectan la predicción de un AHE. Dicho esto, no existe un consenso en un tamaño o intervalo predeterminado entre la observación y predicción lo que deriva en una variedad de estudios que implementan diferentes tamaños de ventanas o intervalos entre la observación y predicción.

3.2 Variedad de enfoques de predicción de AHEs

La predicción, pronóstico o inferencia de hechos futuros es de gran relevancia en la medicina predictiva dado que tiene como objetivo identificar los pacientes en riesgo de desarrollar una enfermedad, permitiendo así la prevención o el tratamiento temprano de esa enfermedad [30]. Con base en las postulaciones de la medicina predictiva, una gran variedad de enfoques o métodos se han desarrollado para su aplicación en el ámbito médico. A fin de analizar los enfoques de predicción de AHEs que han sido propuestos en la literatura, a continuación se definen cuatro aspectos o factores que se utilizarán para guiar el análisis del trabajo relacionado.

1. **Conjunto de datos.** En este factor se describe los conjuntos de datos usados (de signos

vitales y datos clínicos) como fuente de información para la predicción de AHEs. También se evalúo de forma teórica y empírica las ventajas y desventajas de la selección de uno o más signos vitales.

2. **Pre-procesamiento y representación.** En este factor se presentan las técnicas de pre-procesamiento, limpieza de datos, reducción, transformación e integración usada para la preparación de los datos. También se presentan la representación usada de las series de tiempo y en particular de los componentes de frecuencia y componentes de tiempo presentes en las series de tiempo de signos vitales.

3. **Construcción de modelo.** Es de interés analizar este aspecto para identificar las diferentes técnicas usadas durante el proceso de predicción de AHEs. De forma particular, es necesario conocer las ventajas y desventajas, así como las propiedades estocásticas de las series de tiempo.

4. **Presentación y evaluación.** En este último aspecto se desea conocer las consideraciones usadas para la definición de un AHE, en términos del tamaño de la ventana de observación y el porcentaje de muestras de MAP por debajo de un umbral no deseado. También se evalúa los estados considerados durante la predicción. Por último, se indican los resultados de las métricas de clasificación y predicción alcanzados por diferentes propuestas así como los enfoques de segmentación (sin intervalo entre ventanas y con intervalo entre ventanas).

3.2.1 Conjunto de datos

Un conjunto de datos es una tabla o una matriz que puede contener información de datos clínicos o fisiológicos de un paciente. Esta información puede ser explotada por un algoritmo para obtener conocimiento oculto, por ejemplo, predicción de AHEs. Para la construcción del modelo de predicción de AHEs es necesario disponer de un conjuntos de datos que contenga información fisiológica (signos vitales) o en algunos casos información clínica. El conjunto de datos mayormente usado en el estado

del arte corresponde a la base de datos MIMIC-II [1, 7, 8, 19, 21, 22, 26, 26, 27, 28, 31, 32, 33, 34, 35, 78]. Otros conjuntos de datos abiertos que se usan son: MIMIC III en [51] y BrainIT en [20]. También se observa el uso de fuentes de datos privados, en [31] usan una muestra del hospital local para la validación externa que consta de un conjunto de series de tiempo de MAP.

De estos conjuntos de datos se usa principalmente el signo vital correspondiente a la MAP [7, 26, 28, 32, 33, 55], el cual es de interés para la predicción de AHE, en otros casos se observa el uso de la ABP [2, 19, 21, 78] sobre el cual calculan directamente la MAP. Otros signos vitales usados durante la predicción suelen ser ritmo cardíaco, SDP, DBP, MAP, ritmo respiratorio, SpO2 [22]. Otros autores proponen el uso de diferentes signos vitales, en [45] usan el ritmo cardíaco, presión de pulso y gasto cardíaco relativo. También se usan otros signos vitales menos explorados como el uso sólo de ECG [57] para la predicción aunque los resultados finales no destacan en el estado del arte.

3.2.2 Pre-procesamiento y representación

En [27] se propone una descomposición de la señal en segmentos iguales para extraer información en el dominio del tiempo usando transformada *wavelet* para cada segmento, lo cual puede suponer un inconveniente al asumir segmentos iguales dado que las series de tiempo de origen fisiológico son aleatorias [16]. Los autores en [35] describen un método para la extracción de cinco características globales (pico, moda, oblicuidad, *kurtosis* y entropía de *Shannon*) representados en cinco vectores independientes que contienen información en el dominio de la frecuencia.

Por su parte, en [23] se describe un estudio de dos nuevas características basadas en la transformada base de *Hilbert*. Dichas características al igual que la transformada *wavelet* tienen una representación en el dominio de frecuencia. En este sentido, en [61] también propone otro método basado en descomposición de la señal usando la transformada *wavelet* para extraer las características en el dominio de la frecuencia y tiempo. Los autores en [19] utilizan un algoritmo que procesa y extrae parámetros Gaussianos de una fuente de datos de series de tiempo de presión arterial (ABP,

por sus siglas en inglés). Posteriormente aplica una técnica de regresión logística para construir un vector que indica la probabilidad de que se presente o no un AHE.

Otros autores en [55] realizan una división en rangos de gravedad sobre características en el dominio de la frecuencia. Por otro lado otros autores [19, 45, 45] centran su propuestas en el pre-procesamiento de los datos y el uso de trasformada *wavelet* para la generación de nuevas características a partir de series de tiempos de datos fisiológicos. Asimismo, en [32, 33] se propone dos métodos que se basan en la extracción de las características en el dominio de la frecuencia usando la modulación y la amplitud de la frecuencia.

3.2.3 Construcción de modelo

En [62] se propone una red neuronal multi-modelo que a partir de un procedimiento de análisis de correlación de una serie de tiempo MAP actual, y un grupo de plantillas MAP representativas, se identifican y muestrean las plantillas más similares. En una siguiente etapa, el multi-modelo correspondiente previamente es entrenado usando las plantillas históricas con el objetivo de predecir la evolución futura de la señal de entrada de MAP actual. Algo similar proponen en [25] que simplifica el enfoque usando ventanas de *parzen* para construir un modelo de normalidad en vez del modelo de correlación propuesto en [62]. En [7] se presenta un nuevo algoritmo escalable y robusto para encontrar los mejores parámetros *wavelet* usando procesos gaussianos [1] e integra características estadísticas para mejorar la precisión del algoritmo propuesto.

En [46] se presenta un algoritmo de búsqueda y coincidencia de patrones basado en la similitud que identifica datos de series de tiempo. En este caso, los segmentos de series de tiempo son representados por vectores de características que reflejan los patrones dinámicos de series de tiempo fisiológicas. Los vectores de características son usados para generar el modelo de mezclas Gaussianas (GMM, por sus siglas en inglés) que calcula la similitud entre segmentos usando la medida de distancia de Mahalanobis. Finalmente, la clasificación es realizada usando el algoritmo de 1-NN con series de

[1] Los procesos gaussianos son muestras de valores que cambian aleatoriamente en el tiempo.

tiempo fisiológicas sintéticas y reales de una variedad de fuentes. Por otra parte, en [42] se compara la red bayesiana dinámica y $k-$NN para la predicción de la ocurrencia de AHE y cómo afecta en la predicción el desbalance de las clases. También sugieren de forma empírica que la red bayesiana obtiene mejor rendimiento que el algoritmo de $k-$NN.

Otros autores, en [35] aplican un pre-procesamiento a partir de un algoritmo genético y una SVM con el objetivo de optimizar el conjunto de características, a continuación realiza el aprendizaje, construye el modelo, y clasifica usando una SVM. También, en [55] se propone usar una SVM para cuantificar los datos de presión arterial en rangos de gravedad clínicamente aceptados lo que mejora el tiempo máximo de predicción a dos horas respecto a procesar la información en crudo. En [21] presentan un algoritmo llamado GP-RF que combinan árboles aleatorios (RF, por sus siglas en inglés) y programación genética (GP, por sus siglas en inglés). Los RF se usan como un modelo de clasificación binario, y la GP se utiliza para envolver a los árboles de funciones en RF debido a su fuerte capacidad de búsqueda global. En [45] implementan una ANN para predice la ocurrencia de AHEs. En [31] implementa un algoritmo denominado FlocTrac que extraen características de alteraciones tempranas que describen el debilitamiento de los mecanismos compensatorios cardiovasculares que afectan el ciclo cardíaco.

3.2.4 Presentación y evaluación

El uso de estados normotensivos e hipotensivos son recurrentes en el estado del arte [45], debido a que el objeto de interés es la predicción de AHE y el segundo estado de interés corresponde a los momentos de tiempo cuando no se presenta esta condición (NoAHE).

Las métricas usadas para la evaluación del desempeño de agrupación, clasificación y predicción en el estado del arte son variadas. Principalmente, la SEN y SPE son los más recurrentes [20, 31, 51, 62], algunos incluyen el uso de la ACC [8, 32, 34, 35, 56]. Otros autores, en [31] evalúa el éxito del algoritmo aplicando índices de SEN, SPE y área bajo la curva.

En la Tabla 3.1, se describen los signos vital usados para la predicción de un AHE así como el

número de muestras y el umbral sobre el cual se considera que se presenta un AHE dado una tamaño fijo de una ventana.

Tabla 3.1: Configuración de parámetros para definición de AHEs.

Fuente	Muestras	Signo vital	Umbral	Tamaño ventana
En [23] [24] [61] [62] [70] [78] [27] [35] [34] [55] [8]	90 %	MAP	$\leq$ 60 mmHg	$\geq$ 30min
En [67] [26] [19]	90 %	MAP	$<$ 65 mmHg	$\geq$ 30min
En [25]	80 %	ABP	$<$ 60 mmHg	$\geq$ 30min
En [57]	80 %	SDP o MAP	$\leq$ 90 mmHg o $\leq$ 70 mmHg	$\geq$ 5min

Por último, una revisión a las discusiones, resultados y conclusiones señalan que el estado fisiológico de un paciente varía con el tiempo durante la misma estancia en la UCI. Por tanto, los datos de diferentes momentos con registros iguales o similares deben contener patrones informativos [45]. En [31] se evalúa el rendimiento del algoritmo aplicando las métricas de sensibilidad, especificidad y área bajo la curva y estiman que los descriptores del componente de la frecuencia no aportan suficiente información.

3.3 Discusión

En la Tabla 3.2 se muestra un resumen de los principales enfoques para la predicción de AHEs. En dicha tabla se indican el autor, el año, el método o técnica, el conjunto de datos DS, el tamaño de la ventana de observación O_w, el tiempo de antelación G_n antes de la predicción, la exactitud (ACC), sensibilidad (SEN), especificidad (SPE) y tres factores. Estos tres factores son considerados durante el planteamiento del problema y son importantes porque se estima que su análisis en conjunto permitiría obtener información más útil. A continuación se describen los tres factores.

- **Factor 1 (F1).** Este factor hace referencia a la componente de la frecuencia. En particular, al uso de descriptores o características que reflejan los cambios de frecuencia en una serie de

tiempo.

- **Factor 2 (F2).** Este factor hace referencia a la componente del tiempo y denota el uso de descriptores que representa diferentes segmentos de una observación de una serie de tiempo o un descriptor o característica sensible a los cambios en el tiempo.

- **Factor 3 (F3).** Este factor indica el uso o consideración de las propiedades estocásticas de la series de tiempo, lo cual puede requerir el uso de métodos o técnicas que modelen procesos estocásticos (aleatorios).

Tabla 3.2: Enfoques propuestos para predicción de AHEs.

Autor	Año	Método o técnica	DS	O_w (min)	G_n (min)	ACC	SEN	SPE	F1	F2	F3
Feras Hatib [31]	2018	Usan regresión logística e integran nueve características	MIMIC II	30	5	NA	0.92	0.92	✓	✓	✗
				30	10	NA	0.89	0.90	✓	✓	✗
				30	15	NA	0.88	0.87	✓	✓	✗
Dazhi Jiang [35]	2017	Algoritmo genético y Máquina de vectores de soporte (AG-SVM)	MIMIC II	60	0	0.89	0.92	0.88	✓	✓	✗
				60	0	0.81	0.78	0.82	✓	✓	✗
Sakyajit Bhatt. [7]	2014	Algoritmo basado en SVM	MIMIC II	45	120	0.94	0.94	0.95	✓	✗	✗
				50	5	0.92	0.97	0.89	✓	✗	✗
				20	115	0.92	0.84	0.96	✓	✗	✗
Teresa Rocha [62]	2011	Red neuronal GRNN	MIMIC II	140	0	NA	0.82	0.78	✓	✗	✗
Joon Lee [45]	2010	Transformada *wavelet* y Red neuronal Artificial	MIMIC II	30	60	0.86	0.81	0.86	✓	✗	✗
				60	60	0.86	0.83	0.86	✓	✗	✗
				30	120	0.83	0.79	0.83	✓	✗	✗
				60	120	0.83	0.79	0.83	✓	✗	✗

La propuesta presentada en [31] obtiene los índices de clasificación más altos (sensibilidad 0.92 y especificidad 0.92) usando un conjunto de nueve descriptores que aportan información útil al algoritmo de regresión logística que realiza la predicción. También es de resaltar los resultados en [35] que usa un descriptor de entropía de *Shannon* que aporta información en el componente del

tiempo y a su vez otros conjunto de cuatro descriptores (pico, *kurtosis*, moda y una medida de probabilidad de distribución o *Skewness*) que aporta información del componente de la frecuencia. Por otro lado otros autores en [7, 45, 62] centran su propuesta en el uso de descriptores en el dominio de la frecuencia lo cual compromete en cierta medida los índices de ACC, SEN, SPE.

3.4 Resumen

En este capítulo se han presentado los enfoques de segmentación de ventanas de series de tiempo, en particular, los beneficios de la segmentación con intervalo entre la ventana de observación y la ventana de predicción. Asimismo, a fin de analizar los enfoques de predicción de AHEs que han sido propuestos en la literatura se definen y presentan cuatro aspectos para el análisis de los trabajos relacionados. Por último, se realiza una discusión sobre los tres factores identificados en el planteamiento del problema.

4

Diseño y desarrollo del método propuesto

En este capítulo se detalla el diseño y desarrollo del método propuesto para la predicción de episodios agudos hipotensivos (AHEs, por sus siglas en inglés). En particular, se propone una codificación de las series de tiempo de presión arterial media (MAP, por sus siglas en inglés), se emplea una estrategia de agrupación de los datos categóricos resultantes para la identificación de estados hipotensivos y se realiza la predicción de AHE con base en la codificación.

4.1 Descripción general

Con base en la metodología planteada en el Capítulo 1, para el enfoque propuesto se describe de forma general las etapas establecidas que agrupan el diseño de codificación, la identificación de estados hipotensión y el diseño e implementación del método de predicción en dos etapas descritas a continuación.

1. **Codificación e identificación de estados hipotensivos.** A partir de la disponibilidad de

un conjunto de series de tiempo MAP pre-procesadas, esta etapa consiste en convertir cada una de las series de tiempo de MAP a un nuevo espacio de códigos mediante la definición de una codificación de los AHEs. Para ello, tomando en cuenta la definición de AHE descrita en el Capítulo 1, cada serie de tiempo de MAP se transforma a un espacio binario y, mediante el enfoque descrito en la Sección 4.2, se obtiene una codificación de los AHEs existentes en las series de tiempo MAP consideradas. Finalmente, se hace uso de un algoritmo de agrupamiento de datos categóricos para encontrar los k grupos en el conjunto de datos codificado que se obtuvo en el paso previo. En la Figura 4.1 se muestran los tres pasos a seguir para la realización de esta etapa.

2. **Predicción de AHEs mediante cadenas de Markov.** Asumiendo que en una ventana de observación una muestra de la serie MAP puede formar parte de un episodio hipotensivo (AHE) o un episodio normotensivo (NoAHE) [1], en esta etapa se hace uso de una cadena de Markov de dos estados para modelar la ocurrencia de AHEs en una serie de tiempo de MAP. En particular, la matriz de transición de dichos estados se obtiene de la etapa previa en la cual se identificaron los k grupos resultantes en el conjunto de datos codificado.

4.2 Codificación e identificación de estados hipotensivos

En esta sección se detalla cada una de los pasos del diseño del método de codificación y generación de estados binarios. Como se ilustra en la parte superior de la Figura 4.1 el insumo del enfoque propuesto consiste en un conjunto de datos de MAP pre-procesados. En este caso como se explicará en el Capítulo 5, se ha utilizado un conjunto de datos extraídos de MIMIC-II.

[1] Es decir, si la muestra no es etiquetada como un AHE dado la definición descrita en el Capítulo 1 para un AHE.

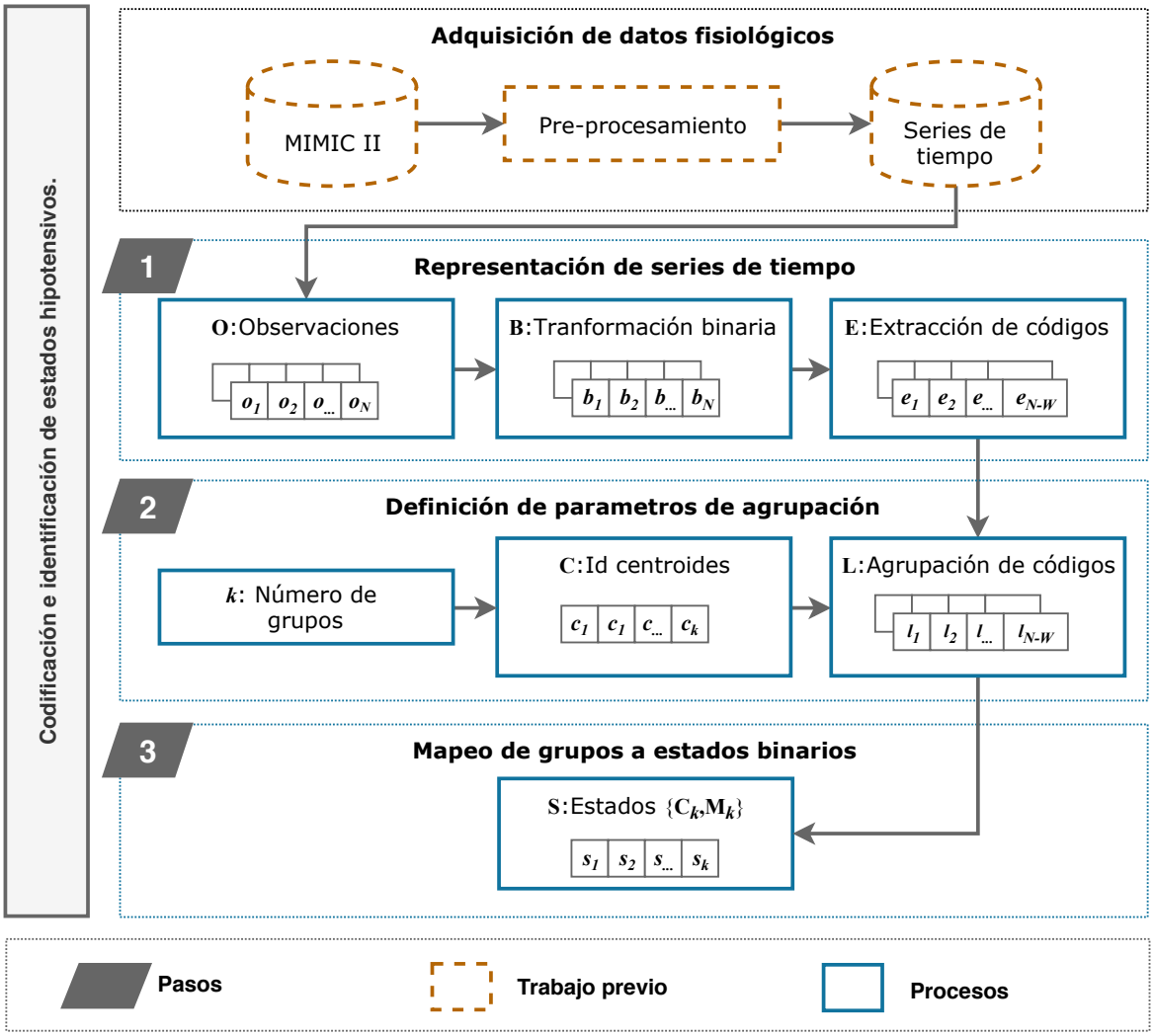

Figura 4.1: Flujo de procesos de la etapa de codificación y generación de estados binarios.

4.2.1 Representación de series de tiempo

Una representación de una serie de tiempo implica que su descripción contenga las características principales asociadas a los componentes de la frecuencia y el tiempo. En este contexto, un tipo de dato cuantitativo que describe un segmento de una serie de tiempo de MAP puede representarse como un AHE. Esta representación es posible al asignar una etiqueta a un segmento de la MAP a partir de algunas características observadas. Representar un segmento de una serie de tiempo de la MAP como un AHE no es una tarea fácil. Esto requiere el diseño de una representación que preserve

las características de una serie de tiempo como es el caso de la MAP que contiene información en el componente del tiempo y la frecuencia.

4.2.1.1. Observación

El conjunto de observaciones a considerar corresponde a series de tiempo de la MAP, cada una de las observaciones (O) cuentan con diferentes longitudes de tiempo dado un número n de determinadas muestras tomadas a intervalos regulares de un minuto. La observación se representada como un vector O de números decimales en un rango de $0 \leq O \leq 160$ medida en mmHg. Estas observaciones (O) tienen una componente estocástica que es característica de una señal biomédica o serie de tiempo [16].

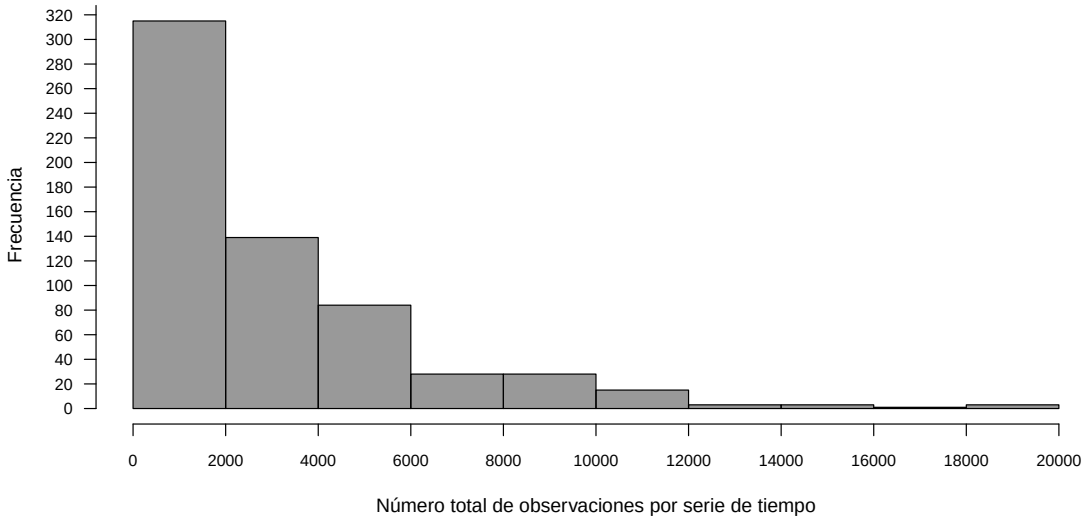

Figura 4.2: Frecuencia número de muestras por serie de tiempo.

En la Figura 4.2 se ilustra la frecuencia del número de muestras contenidas en las 626 series de tiempo de MAP consideradas en la presente tesis. Como se muestra en la gráfica, para 310 series de tiempo el número de muestras se encuentra entre 0 y 2,000. También se observa que 140 series de tiempo contienen un total de 2,000 a 4,000 muestras. Asimismo, se observa el número de total de muestra por serie de tiempo agrupados en segmentos que incrementan en un rango de 2,000

muestras.

4.2.1.2. Transformación binaria

En esta etapa se plantea una transformación binaria (B) a partir de una serie de observaciones $O = \{O_1, O_2, \ldots, O_n\}$, lo cual pueda reducir y abstraer ciertas características de interés al momento de identificar un AHE. Por tal motivo, con base en la definición de un AHE, se plantea la aplicación de una transformación binaria de la señal usando la Ecuación 4.1 y un umbral hipotensivo de 60 $mmHg$.

$$B(i) = \begin{cases} 1 & \text{en caso de } i < 60 \\ 0 & \text{en caso contrario} \end{cases} \tag{4.1}$$

En la Figura 4.3, se ilustran dos señales. La señal en la parte superior corresponde a 90 muestras de la MAP O_{90}. A dicha señal se le aplica la transformación binaria usando la Ecuación 4.1, dando como resultado la señal discreta B_{90} que contiene los valores binarios ($[0, 1]$).

4.2.1.3. Extracción de códigos

A partir de la transformación binaria B_N de la serie de tiempo de MAP se aplica un mecanismo de ventana deslizante que permite la extracción de códigos. Para esto se define una ventana E de tamaño $w = 30$, lo cual equivale a 30 muestras tomadas a intervalos regulares de un minuto. Esta ventana E se desliza en el tiempo $t = 1, 2, \ldots, N - w$, mientras se extraen un conjunto de vectores cada uno formado de 30 ceros o unos ($[0, 1]$) que representan cada uno de los diferentes códigos en el tiempo. Estos códigos tienen una característica importante en términos de memoria, debido a que cada nuevo código en el tiempo $t + 1$ contiene un total de 29 caracteres del código anterior t más un nuevo carácter. Esto se realiza pensando en proveer propiedades de la componente de tiempo en cada uno de los códigos.

En la Tabla 4.1 se observa el mecanismo de ventana deslizante y los códigos E en el tiempo

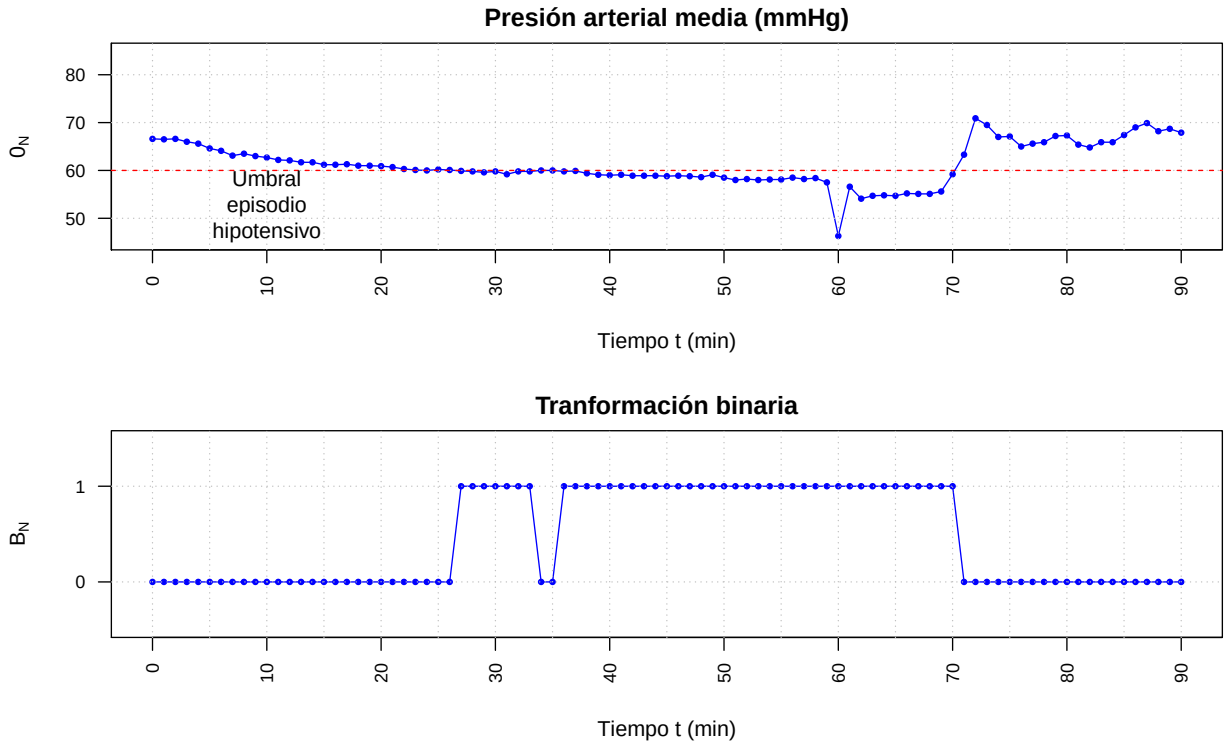

Figura 4.3: Ejemplo de una transformación binaria de la MAP.

Tabla 4.1: Vista de una extracción de códigos realizada a un segmento de una serie de tiempo de MAP.

Tiempo (t)	E (BIN)																												E (HEX)	E (DEC)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30		
t = 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	00000043	67
t = 2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	1	00000087	135
t = 3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	1	1	0000010F	271
t = 4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	1	1	1	0000021F	543
t = 5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	1	1	1	1	0000043F	1087
t = 6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	1	1	1	1	1	0000087F	2175
t = 7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	1	1	1	1	1	000010FF	4351
t = 8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	1	1	1	1	1	1	000021FF	8703
t = 9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	000043FF	17407
t = 10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	000087FF	34815
t = 11	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	00010FFE	69630
t = 12	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	00021FFC	139260
t = 13	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	00043FF8	278520
t = 14	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	00087FF0	557040
t = 15	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0010FFE0	1114080
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
t = N-W	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	3FFFFFFF	1073741823

$t = \{1, 2, 3, \ldots, N - w$ y respectivamente cada uno de los códigos en notación binaria (BIN), hexadecimal (HEX) y decimal (DEC). Básicamente los códigos generados por el mecanismo de

ventana deslizante son una representación de los estados que contienen propiedades del dominio del tiempo dado la capacidad de memoria y del dominio de la frecuencia en forma discreta de ceros y unos.

Si bien esta representación por medio de un código contiene información en los dominios de la frecuencia y tiempo, el total de códigos o estados diferentes que se pueden generar es de 2^w. Esto supone un problema debido a que para un total de $w = 30$ se pueden generar 1,073,741,824 códigos diferentes. Por tal motivo, se presenta a continuación una forma de agrupar códigos similares en grupos que representen diferentes estados en un modelo de predicción.

4.2.2 Definición de parámetros de agrupación

Debido a la problemática del alto número de códigos obtenidos del conjunto de datos de series de tiempo de MAP, se evaluó inicialmente de forma teórica y empírica las ventajas y desventajas de los métodos de agrupamiento y clasificación.

- Un método de clasificación es una forma de aprendizaje automático que a partir de un conocimiento previo (representado como patrones y etiquetas) se clasifica un nuevo patrón asignado una etiqueta o marca de clase.

- Un método de agrupación es una forma de aprendizaje automático que agrupa diferentes elementos con base a un criterio en k subconjuntos. Estos subconjuntos pueden variar en tamaño para $k \geq 2$ hasta n elementos a agrupar. Diferentes criterios de agrupación permiten generar grupos que compartan ciertas características.

La aplicación de estos métodos de clasificación y agrupación para nuestra problemática planteada presenta las siguientes ventajas y desventajas que se pueden observar en la Tabla 4.2. Por tal motivo, dado el problema del alto número de códigos la opción más viable es optar por el uso de un método de agrupamiento. Un método de agrupamiento es también una secuencia de pasos o instrucciones

que potencialmente pueden ser escritas como un algoritmo. Un algoritmo de agrupamiento implica básicamente dar respuesta a dos interrogantes: 1) ¿Cuál es el criterio de agrupación? y 2) ¿Cuál es el número de grupos a generar? estas interrogantes son básicamente los parámetros a definir para nuestro algoritmo de agrupamiento.

Tabla 4.2: Ventajas y desventajas de la clasificación y la agrupación de estados.

Método	Ventajas	Desventajas
Clasificación	Se acota el modelo a una salida de clases binarias (AHE/NoAHE).	Se estima que pueden existir más estados representativos.
Agrupación	Permite ajustar el número de estados (k).	Se requiere definir *a priori* el número de estados (k)

Existen diferentes criterios de agrupación definidos en la literatura, los cuales pueden ser clasificados en distancia (por ejemplo, distancia euclidiana), similitud (por ejemplo, matriz de correlación) y verosimilitud (por ejemplo, estimar una función verosimilitud para distribución normal). Por otra parte, es posible generar $k \geq 2$ grupos hasta n número de elementos a agrupar. Por tal motivo, la selección de un criterio de agrupación y número determinado de grupos no es una tarea fácil. Para esto se definió un orden para dar respuesta a las dos interrogantes planteadas anteriormente, de forma que primero se responde a la interrogante sobre el criterio de agrupación y segundo se responde a la interrogante sobre el número de grupos a generar. Este orden da mayor importancia a la selección teórica de un criterio de agrupación antes de evaluar el número de grupos que puedan representar de una mejor forma los estados. A continuación se presentan las consideraciones para la selección del criterio de agrupación.

4.2.2.1. Criterio de agrupación

Durante el proceso de desarrollo de esta fase se evaluaron diferentes criterios de agrupación. Inicialmente se implementó el algoritmo de $k-$means y la medida de distancia euclidiana, lo cual fue descartado; sin embargo, este primer experimento permitió identificar algunas propiedades de los estados a agrupar. Una de dichas propiedades es que los códigos con mayor número de frecuencia

de valores en uno son los correspondientes a los AHEs, esto es así debido a que la representación usada para los códigos define el valor de uno para un umbral menor de 60 mmHg. Otra propiedad es que los códigos deben ser procesados como una cadena de datos categóricos y no como un único valor binario o decimales, para la cual se observa en la literatura algunos algoritmos de agrupación categórica que potencialmente pueden presentar rendimientos superiores al momento de evaluar las agrupaciones realizadas. En este contexto se decide entonces probar dos algoritmos modificados para agrupación categórica:

- k-**modas.** Un algoritmo usado regularmente en la literatura para agrupar datos categóricos, tiene como particularidad que los centros son generados calculando la moda de los datos, esto implica que es un método basado en la frecuencia de los datos para actualizar las modas.

- $k-$**means con distancia de Chebyshev.** Una propuesta de algoritmo que usa una medida de distancia de Chebyshev para calcular los centros a partir de la media y trata los estados como valores categóricos.

Finalmente, se observó que la propuesta del algoritmo de $k-$means y el uso de la distancia de Chebyshev representa de forma más eficaz los estados representativos. Los resultados de la agrupación categórica usando k-medias y distancia Chebyshev pueden ser consultados en el Capítulo 5.

4.2.2.2. Número de grupos a generar

Una primera aproximación de un número de grupo a generar se puede observar en la Figura 4.4 que muestra la densidad por medio de una línea suavizada que ondula resaltando por medio de picos posibles grupos. Esta gráfica fue generada a partir de todos los códigos de las muestras de MAP disponibles, en el eje de las abscisas se tiene el rango de códigos con límites de 0 hasta 2^{30} y en el eje de las ordenadas se aprecia la densidad de los códigos. La representación de los códigos en una gráfica de densidad es posible transformando el valor o código en binario a un valor decimal.

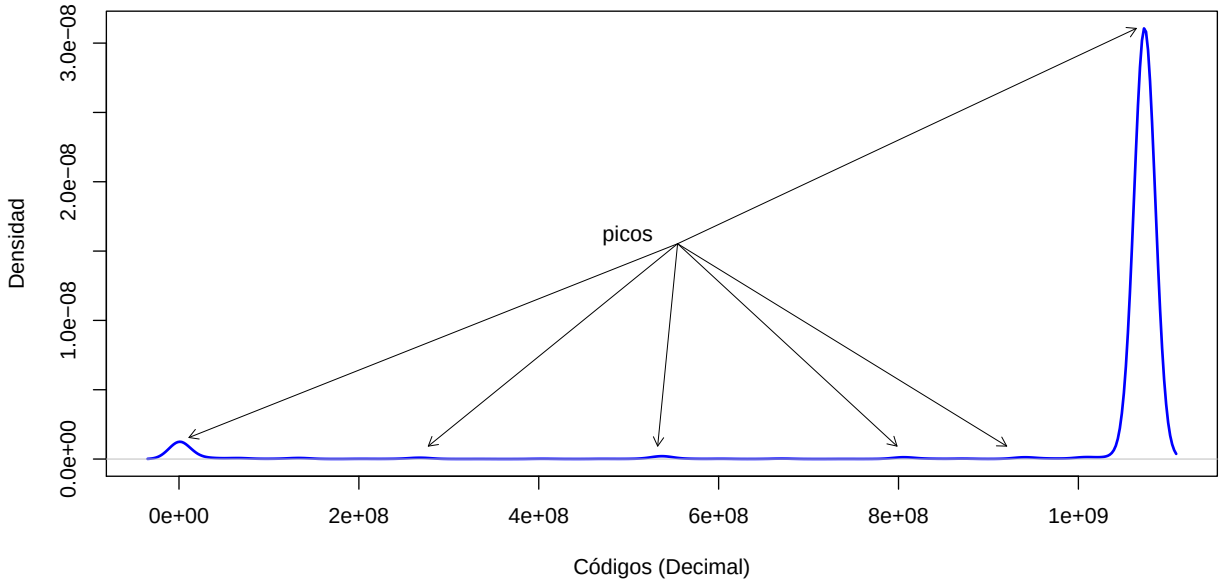

Figura 4.4: Densidad de códigos totales del conjunto de 626 series de tiempo de MAP.

En la Figura 4.4, se ilustran dos picos en los extremos que pueden representar dos grupos y también se observan unas ondulaciones leves en el medio que pueden potencialmente representar otros posibles grupos. La densidad fue calculada usando un *kernel Gaussiano* [29]. A partir de esta observación y la densidad de código se decidió acotar los posibles grupos a generar en $2 \leq k \leq 10$.

4.2.2.3. Algoritmo de agrupación categórica

El algoritmo estándar de agrupación de $k-$means es un algoritmo que agrupa n elementos en k grupos. Un criterio de agrupación permite asociar elementos a un centro y a un identificador o etiqueta de grupo. Para esta propuesta se presenta un algoritmo de $k-$means con medida de distancia de Chebyshev.

La distancia de Chebyshev o distancia de valor máximo calcula la magnitud absoluta de las diferencias entre las coordenadas de un par de puntos (representados por un vector).

$$D(i,j) = \max_{k} |X_{ik} - X_{jk}| \tag{4.2}$$

Procedimiento del algoritmo de agrupación categórica:

- **Paso 1.** Inicializar los centros de forma aleatoria en un espacio de C elementos o categorías.
- **Paso 2.** Agrupar todos los elementos X_i con el centro C_j para cada elemento o categoría, usando la distancia 4.2 donde: $|X_i - X_j| = \min_j |D(i,j)|$
- **Paso 3.** Calcular la media de los elementos para obtener el centro de la función para cada grupo C_J, $C_j = \frac{1}{m_j} \sum_{X_i \in groupj} x_i$, donde m_j es el número de elementos en el grupo.
- **Paso 4.** Repetir el paso 2 y 3 hasta que no cambie cada uno de los grupos asignado C_j.

Un mayor detalle del algoritmo implementado se puede observar en el Anexo B. Este algoritmo recibe como parámetros de entrada los códigos en notación decimal generados en la sección o etapa de la representación de la serie de tiempo y el número de grupos k a generar.

4.2.3 Mapeo de códigos a estados

Se estima que los grupos generados contengan códigos con característica similares dentro de los cuales existe un grupo que contiene los AHEs y el resto de los otros grupos que contienen los códigos etiquetados como NoAHE. Formalmente, dado un conjunto de índices de centroides C con los elementos $\{1, 2, \ldots, k\}$, existe un conjunto unitario que representa los **AHE** y otro conjunto con elementos diferentes etiquetados como **NoAHE**. Esta representación se almacena en un vector M.

En este contexto, el número de combinaciones posibles para M se puede calcular usando combinaciones sin repetición tomando un elemento $n = 1$ de k número de elementos posibles.

$$C_{n,k} = C_{1,k} = \binom{1}{k} = \frac{1!}{k!(1-k)!} = k \tag{4.3}$$

La mejor representación de M corresponde al experimento donde se encuentre el valor máximo del coeficiente de correlación de Matthews al comparar las etiquetas *a priori* (etiquetas asignadas

usando la definición de un AHE) y las etiquetas asignadas por el algoritmo de agrupación categórica para cada uno de los diferentes valores de k grupos.

$$M_i = \begin{cases} 1 & \text{en caso de } \textbf{AHE} \\ 0 & \text{en caso contrario } \textbf{NoAHE} \end{cases} \tag{4.4}$$

Cada uno de estos centros C tiene una etiqueta asignada M_i. A estas etiquetas binarias $[0, 1]$ dado un C se les denomina vector de mapeo a estado binario M el cual es de tamaño k. El mapeo de grupos a estados binarios es una matriz S definida como:

$$S = \begin{bmatrix} C_k \\ M_k \end{bmatrix} \tag{4.5}$$

4.3 Predicción de episodios agudos hipotensivos

En esta etapa propuesta se realizan las dos últimos pasos que se ilustran en la Figura 4.5. En el cuarto paso, se desarrollo un modelo de predicción usando un conjunto de datos de observaciones y se construye la matriz de transición A como se explica en la Sección 4.3.0.1. Por otra parte a partir de un nuevo conjunto de observaciones se calcula las probabilidades iniciales P como se explica en la Sección 4.3.0.2. Finalmente, en el quinto paso se realiza la predicción a n pasos usando la matriz de transición A y las probabilidades iniciales P como se describe en la Sección 4.3.1.

4.3.0.1. Matriz de transición

Una matriz de probabilidades de transición estocástica puede ser construida a partir de una serie de tiempo L_n donde n representa el número de muestras recolectadas a intervalos regulares y t un valor conocido en un instante de tiempo tal que $L_{t-1} = j$ y $L_t = i$ son las probabilidades de transitar respectivamente de un estado a otro A_{ij}. El número posible de estados para i y j es k.

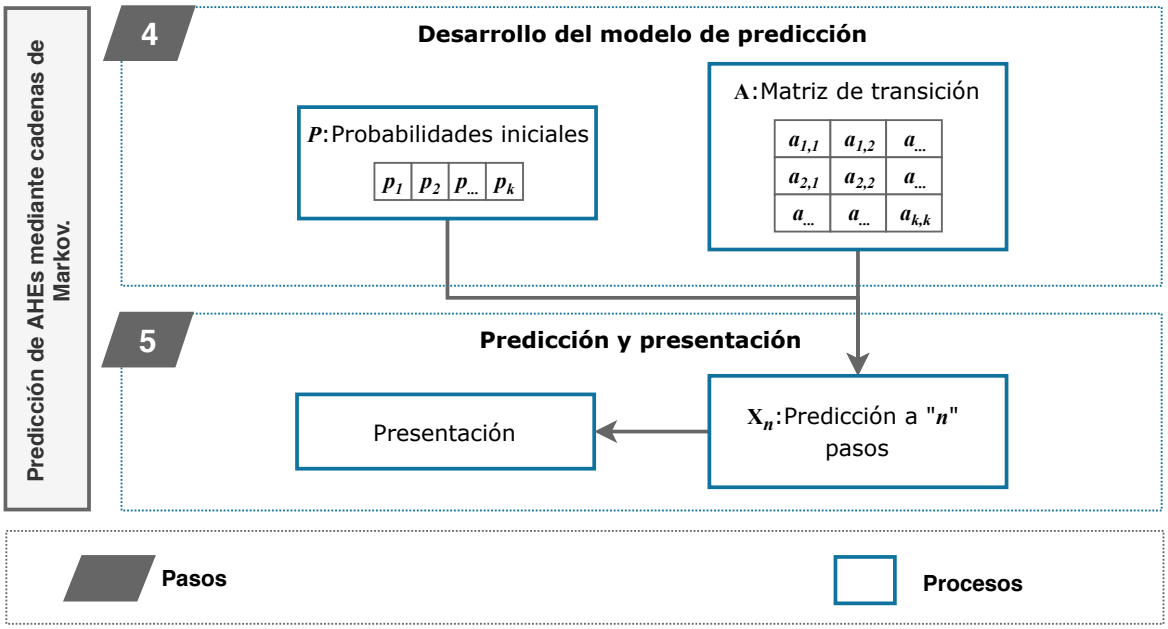

Figura 4.5: Flujo de procesos de la etapa de predicción de episodios agudos hipotensivos.

Sea $L_n \in E|E$: es el conjunto de series de tiempo y n es el número de observaciones de esa serie de tiempo en particular, se construye una matriz de probabilidades de transición A para más de una serie de tiempo siguiendo el método a continuación:

- **Paso 1.** Se suman los cambios de estados para cada serie de tiempo L_n. Con esto se obtiene una matriz con valores enteros en cada uno de sus campos.

$$A_{ij} = \sum_{n=2}^{|L|} A_{L(n-1),L(n)} + 1 \tag{4.6}$$

- **Paso 2.** Se repite el paso 1 para todas las series de tiempo $L \in E$ y se suman todas las matrices resultantes A.

$$a_{ij} = \sum_{n=1}^{|E|} A \tag{4.7}$$

- **Paso 3.** Se calculan las probabilidades de transición. Está matriz cumple con las propiedades

descritas en la Ecuación 2.4.

$$A_{ij} = \frac{a_{ij}}{\sum_{k=1}^{j} a_{ik}} \tag{4.8}$$

4.3.0.2. Probabilidades iniciales

El cálculo de probabilidades iniciales $P^{(0)}$ se realiza a partir de un conjunto de nuevas observación previamente codificadas y agrupadas hasta obtener el vector L en el paso dos del método propuesto. L es un vector que contiene las etiquetas asignadas por el algoritmo de agrupación categórica asociadas a un conjunto de códigos. Formalmente el vector de probabilidades iniciales $P^{(0)}$ es un vector de tamaño k que satisface la condición de la Ecuación 2.3, es decir que la suma del vector de probabilidades iniciales es uno. Asimismo, una única observación L_i genera un vector de probabilidades iniciales $P = p_1 = 0, p_2 = 0, \ldots, p_k = 0$ tal que existe un $P_i = 1$ y uno o más $P_{n \neq i} = 0$ donde $n^{\mathbb{N}}, i^{\mathbb{N}} \in \{2, \ldots, k\}$.

4.3.1 Predicción y presentación

La probabilidad de predicción a n pasos puede ser calculada a partir de las probabilidades iniciales $P^(0)$ y la matriz de transición A usando la Ecuación 2.5. Estas probabilidades son usadas para inferir la ocurrencia de un estado binario AHE o NoAHE. La predicción se realiza a n pasos en este contexto cada paso corresponde a la frecuencia de muestreo de la observación O que equivale a un minuto en todos los casos. En el siguiente capítulo se hará referencia de forma indistinta a la predicción en n minutos haciendo alusión a la predicción a n pasos.

4.4 Resumen

En este capítulo se han presentado el diseño y desarrollo del método propuesto con base en la metodología planteada, la cual consiste en las etapas de codificación e identificación de estados hipotensivos y predicción de AHEs mediante cadenas de Markov. En particular, el aporte presentado

se centra en la etapa uno. Asimismo, ambas etapas se dividen en cinco pasos en total en los que se desarrollan diferentes procesos siguiendo el flujo descrito en la Figura 4.1 y Figura 4.5. Finalmente, nuestro método propuesto inicia con un conjunto de observaciones de series de tiempo de MAP y finaliza con una predicción a n pasos de estados AHE o NoAHE.

5

Experimentación y resultados

En este capítulo se presenta la experimentación realizada para la evaluación del método propuesto. En particular, se describe la infraestructura utilizada y el diseño experimental. Asimismo, se presentan los resultados de validación y evaluación de rendimiento del método propuesto. Finalmente, se presenta una breve comparativa con trabajos similares reportados en el estado del arte.

5.1 Infraestructura requerida

Para el desarrollo de la tesis se utilizó la infraestructura que se resume en la Tabla 5.1, la cual consta de un servidor y una computadora personal. Además, para este trabajo se usaron parcialmente los resultados del desarrollo de la etapa de fuera de línea reportados en [28]. Específicamente, se ha utilizado la tabla *series* definida en el Capítulo 4 en [28], la cual contiene series de tiempo MAP extraídas de la base de datos MIMIC II. Dicha tabla de datos contiene un total 626 series de tiempo de MAP que suman 1,864,761 muestras tomadas a intervalos regulares de 1 minuto. Los campos de dicha tabla que resultan de interés para la presente investigación se detallan a continuación:

Tabla 5.1: Descripción del equipo utilizado en la experimentación.

Equipo de cómputo	Características
Servidor eSalud	Sistema operativo: Ubuntu Linux Server 13.04 LTS Procesador: Inter(R) Xeon(R) CPU E5-2440 1.90Ghz Memoria RAM: 16 GB - Disco Duro: 1 TB
Computadora personal	Sistema operativo: ubuntu 18.10 Procesador: Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz Memoria RAM: 16 GB 133 Mhz DDR3

- **Número de la serie (id_serie).** Contiene un identificador de la serie de tiempo.

- **Tiempo de la muestra (t).** Contiene el tiempo $t \in \mathbb{N}$ en escala de minutos en el que la primera muestra de la serie de tiempo corresponde al tiempo $t = 1$.

- **Valor de la muestra (ft).** Contiene el valor de la muestra de la MAP pre-procesadas.

Las técnicas de pre-procesamiento aplicadas a la muestra (ft) correspondiente a la MAP son:

- **Tratamiento de valores atípicos.** Los valores atípicos son observaciones que son inconsistentes con el resto de las observaciones [5], en este caso los valores atípicos son aquellas muestras f_t de las series de tiempo que tienen valores irreales para un paciente que pueden observarse debido a fallas técnicas en equipos de monitoreo [77]. En este proceso los limites definidos para ft corresponde a $0 \geq ft \geq 160$.

- **Estimación de segmentos faltantes.** Además del tratamiento de valores atípicos, las series de tiempo contienen segmentos faltantes que son procesados con el objetivo de rellenar la información no disponible. En este caso se usó extrapolación en los segmentos ubicados al inicio y al final de la serie de tiempo. Por otra parte en el caso de los segmentos ubicados entre la serie de tiempo se usó interpolación.

5.2 Diseño experimental

La experimentación realizada se ha dividido en dos partes, de las cuales la primera tiene como objetivo validar la etapa codificación e identificación de estados hipotensivos. En la segunda parte de la experimentación se valida la predicción de AHEs mediante cadenas de Markov (ver Capítulo 4). Como parte de la validación y evaluación de resultados se usaron las siguientes métricas de clasificación y predicción: coeficiente de correlación de Mattews (MCC), exactitud (ACC), tasa de error (ERRORRATE), sensibilidad (SEN) y especificidad (SPE), tal como se detalla en las Ecuaciones 2.6, 2.7, 2.8, 2.9 y 2.10, respectivamente.

El método propuesto y su respectivo flujo de datos descritos en el Capítulo 4 así como los algoritmos fue programado en lenguaje R, para el método de agrupamiento se usó los algoritmos descritos en Anexo B.

5.3 Validación de codificación e identificación de estados hipotensivos

En esta sección se validan los pasos propuestos en la etapa de codificación e identificación de estados hipotensivos y se validan los resultados usando la SEN, la ESP y el MCC para diferentes valores de números de grupos k.

En la Figura 5.1 se detalla el proceso de validación que inicia con la extracción de un conjunto de códigos (E). Posteriormente cada código E es etiquetado tomando en cuenta la definición formal de un AHE de lo cual se obtiene Y que denota dicho etiquetado *a priori* en el que se identifican dos estados binarios posibles AHE y NoAHE. Además, al conjunto de códigos E se le aplica un proceso de agrupación categórica (usando el algoritmo descrito en el Anexo B) con el objetivo de agrupar los códigos similares para un número de grupos k determinado obteniendo el vector L. Dicho vector (L) que contiene los k posibles grupos asociados a los códigos es mapeado a un vector de estados

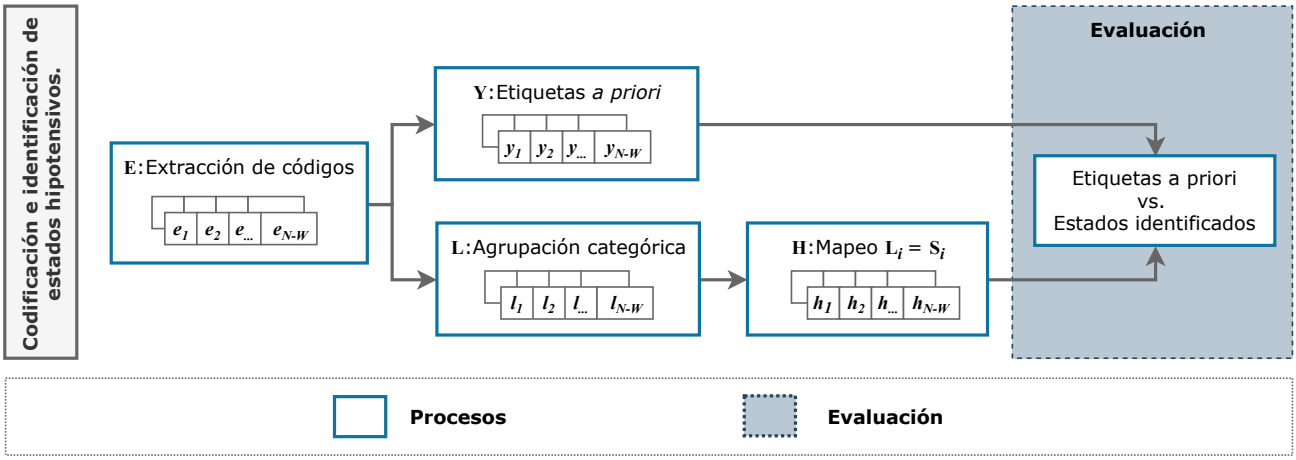

Figura 5.1: Método de validación de la etapa de codificación e identificación de estados hipotensivos.

binarios H usando como referencia la matriz de estados S. Finalmente, se evalúa la aproximación de la codificación e identificación de estados hipotensivos comparando las etiquetas *a priori* Y y el vector de estados mapeados H usando las métricas descritas en el Capítulo 2.

Una primera aproximación de la validación se presenta en la Figura 5.2, en la que se muestra 30,000 códigos iguales a los que se les aplicó el proceso descrito en la Figura 5.1 para $k = 2$ grupos. En este caso en la gráfica de la izquierda se observan los estados 1 y 2. Asimismo, la segunda gráfica muestra los estados 0 y 1 correspondiente a las etiquetas NoAHE y AHE obtenidos al mapear los códigos previamente agrupados. Por otra parte, la tercera imagen a la derecha muestra las etiquetas *a priori*. Finalmente, al comparar visualmente la gráfica del mapeo con la gráfica de las etiquetas *a piori*, se observa que para cada código las etiquetas coinciden en cierta medida.

En la Figura 5.3 se observa también el mismo segmento de la serie de tiempo de 30,000 códigos y las etiquetas asignadas con la diferencia que la agrupación categórica se obtuvo a partir de un valor para $k = 4$ a grupos.

En el experimento siguiente se observa con mayor detalle el comportamiento de un pequeño segmento de códigos que siguen el método de validación que se muestra en la Figura 5.1. Este comportamiento se muestra en la Figura 5.4 por medio de tres gráficas donde se observa en cada una en el eje de las abscisas el código en el tiempo t y el eje de la ordenada la etiqueta asignada

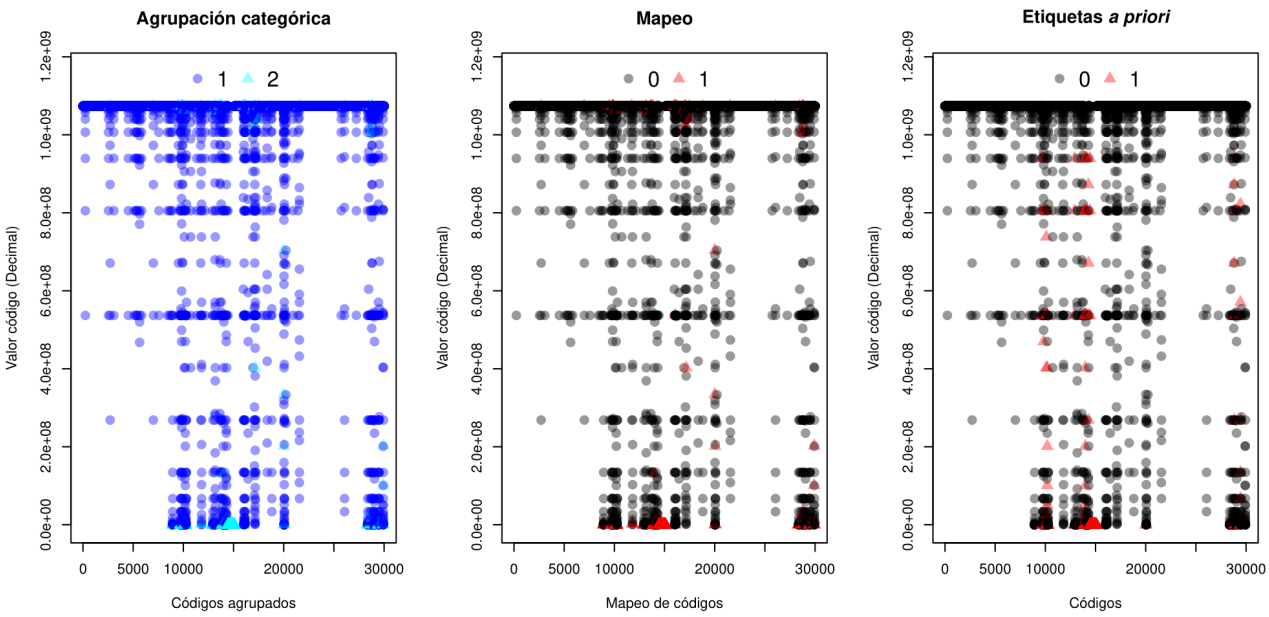

Figura 5.2: Comparativa método propuesto vs. etiquetas *a priori* para $k = 2$.

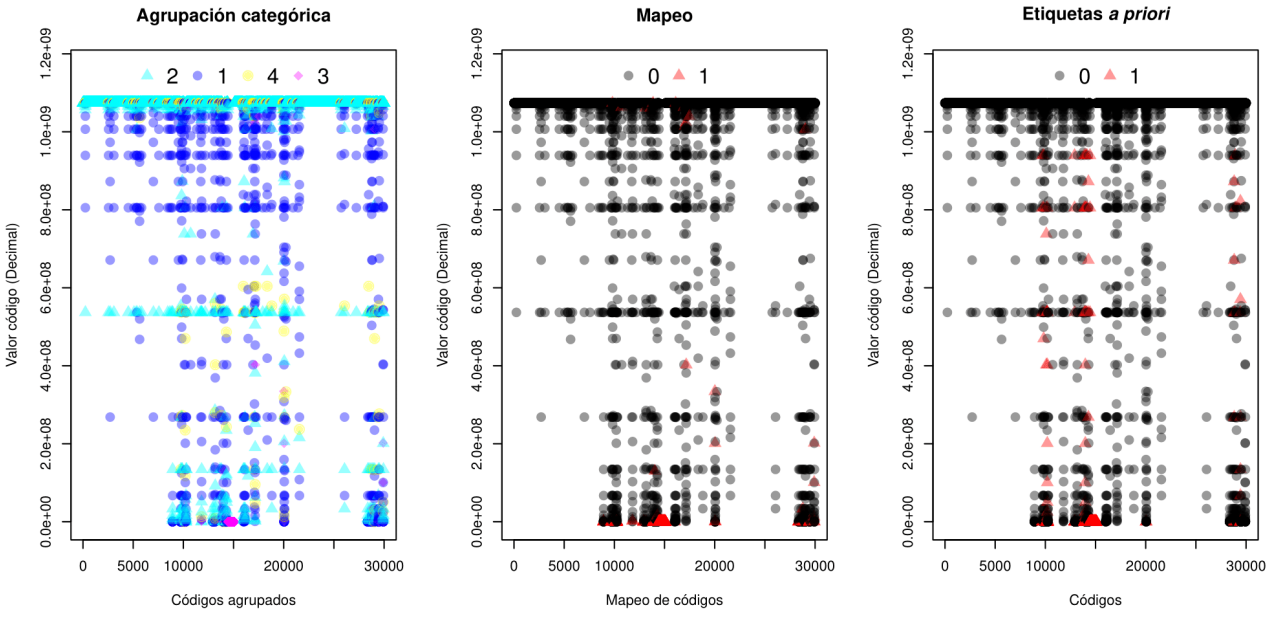

Figura 5.3: Comparativa método propuesto vs. etiquetas *a priori* para $k = 4$.

al código. En este caso la gráfica con la agrupación categórica (L) se realizó para $k = 4$ grupos. Posteriormente, se observa en la gráfica de mapeo de grupos a estados binarios los estados 0 (NoAHE)

y 1 (AHE) resultantes al mapear el vector L. De igual manera, se muestra en la gráfica de etiquetas *a priori* los estados 0 (NoAHE) y 1 (AHE) asignados usando la definición de un AHE. Al comparar la gráfica de mapeo de grupos a estados binarios y las etiquetas *a priori* Y, se destaca que tienen algún grado de aproximación al observar el cambio entre los estados NoAHE y AHE.

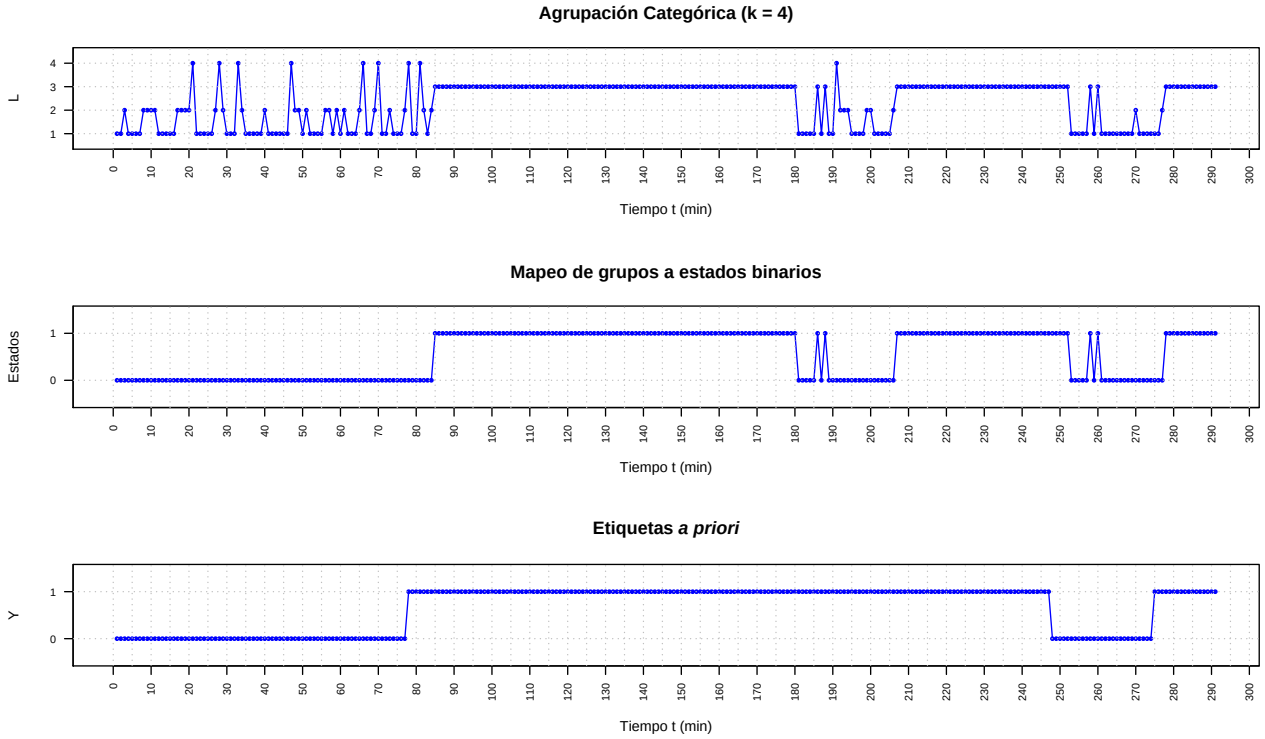

Figura 5.4: Comparativa del mapeo de grupos a estados binarios vs. las etiquetas *a priori*.

En la Figura 5.5, se ilustran la SEN, la SPE y el MCC que se obtienen al comparar las etiquetas *a priori* contra las etiquetas de la agrupación categórica previamente mapeada a los estados binarios. En este experimento se seleccionó el total de 626 series de tiempo las cuales se les aplico el flujo de datos descritos en la etapa uno descrito en el Capítulo 4. Este proceso de comparación se realizó para observar como con diferentes valores de grupo $10 \leq k \leq 2$ se aproxima la agrupación y mapeado de estados binarios a las etiquetas *a priori*. Para esta figura se realizaron 20 experimentos para cada uno de los k grupos, con este número de experimentos se observó que los valores de los centros convergen en un espacio discreto que es igual en algunos casos; este espacio discreto es el resultado de aplicar

el método propuesto descrito en el Capítulo 4 y los algoritmos en el Anexo B. En particular, en la figura se observa que el valor con menor cambio es la ESP el cual se mantiene con variabilidad entre 0.98 y 1. Por otra parte, la sensibilidad tiene el mayor valor en $k = 3$ grupos y después decae para $k \leq 4$ grupos. Por último, se observa el coeficiente de correlación de Mattews, el cual representa una mejor medición dado el desbalance de clases en los códigos, en este caso el mayor valor se obtiene en $k = 4$.

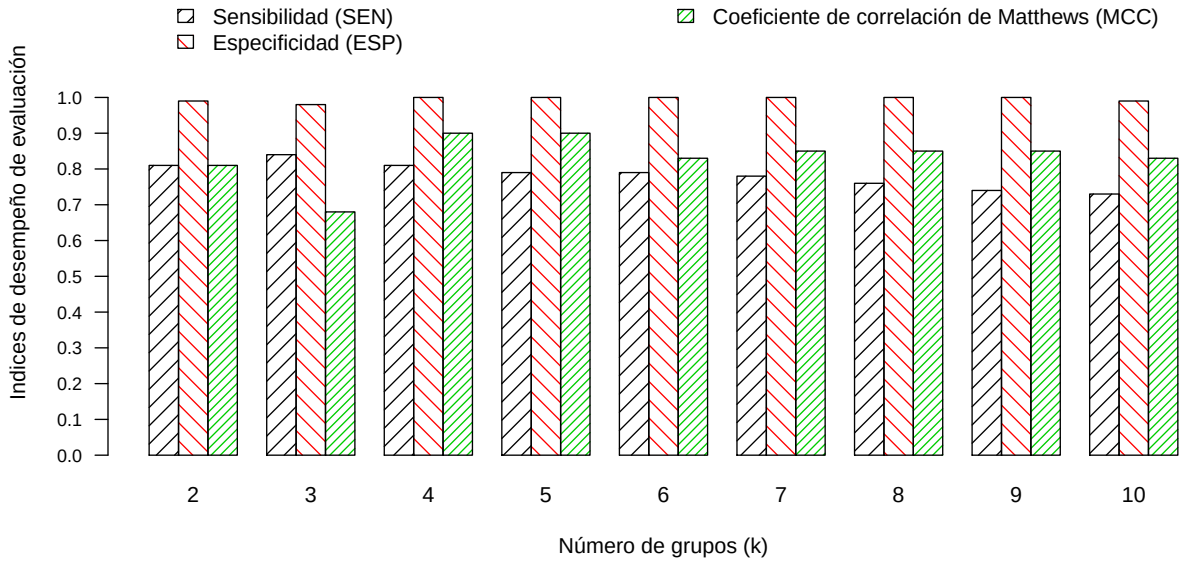

Figura 5.5: Desempeño del algoritmo de agrupación categórica en función del número de grupos k.

Finalmente, el valor de los grupos se ajustó en $k = 4$ con base en los resultados obtenidos en la Figura 5.5. La matriz de estados S y los valores de los centroides asociados a los grupos C_k pueden ser consultados en el Anexo A; estos valores son usados como parte del modelo en la siguiente etapa de predicción de AHEs mediante cadenas de Markov.

5.4 Validación de la predicción de AHEs

Haciendo uso de los resultados obtenidos en la etapa de codificación e identificación de estados hipotensivos, en la presente etapa se valida la predicción de AHEs mediante cadenas de Markov. La validación de la predicción de AHEs sigue el flujo descrito en la Figura 5.6. En este caso usando un código con una etiqueta *a priori* Y_n se compara contra la etiqueta que se obtiene al predecir el mismo código que posteriormente es mapeado H_n a estados binarios. Este proceso compara el valor real contra el valor predicho evaluando los estados (AHE y NoAHE) de ambos valores en Y_n y H_n.

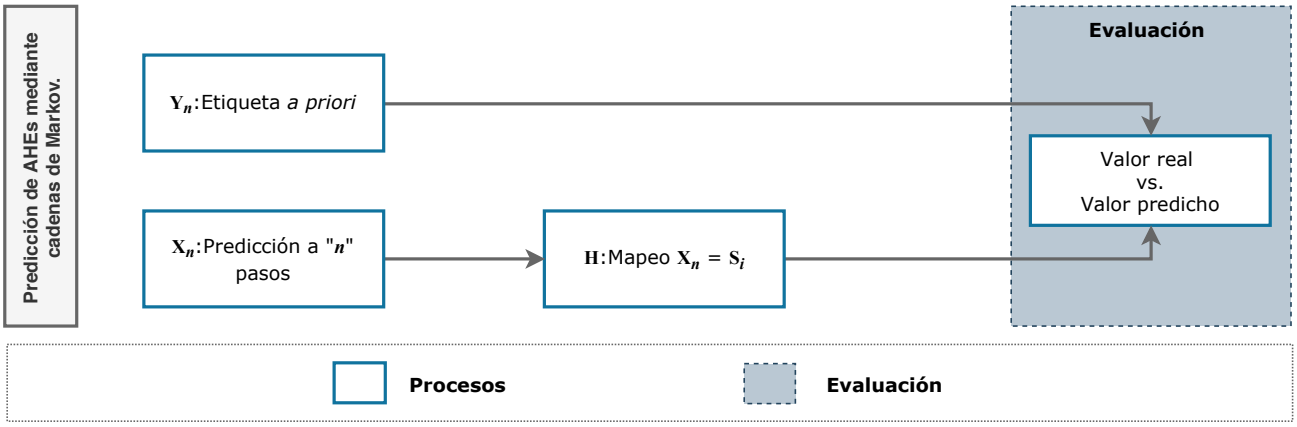

Figura 5.6: Pasos de la evaluación de la predicción de AHEs mediante cadenas de Markov.

Los resultados de predicción en esta sección se comparan evaluando los índices de tasa de error (ERRORRATE), sensibilidad (SEN), exactitud (ACC) y la especificidad (SPE), tal como se definieron en el Capítulo 2.

En esta parte de la experimentación se seleccionaron dos series de tiempo de forma aleatoria del conjunto de datos y se evaluó la predicción en el tiempo. El objetivo de este experimento consiste en validar el método propuesto usando la configuración obtenida en la etapa anterior para tener un primer acercamiento a los resultados que se pueden obtener. Las series de tiempo seleccionadas son: $id = \{306, 12\}$, en los dos casos se usó el total de muestras de las series de tiempo y siguiendo el método propuesto en el Capítulo 4 se realizó la predicción y se comparó contra la etiqueta *a priori*.

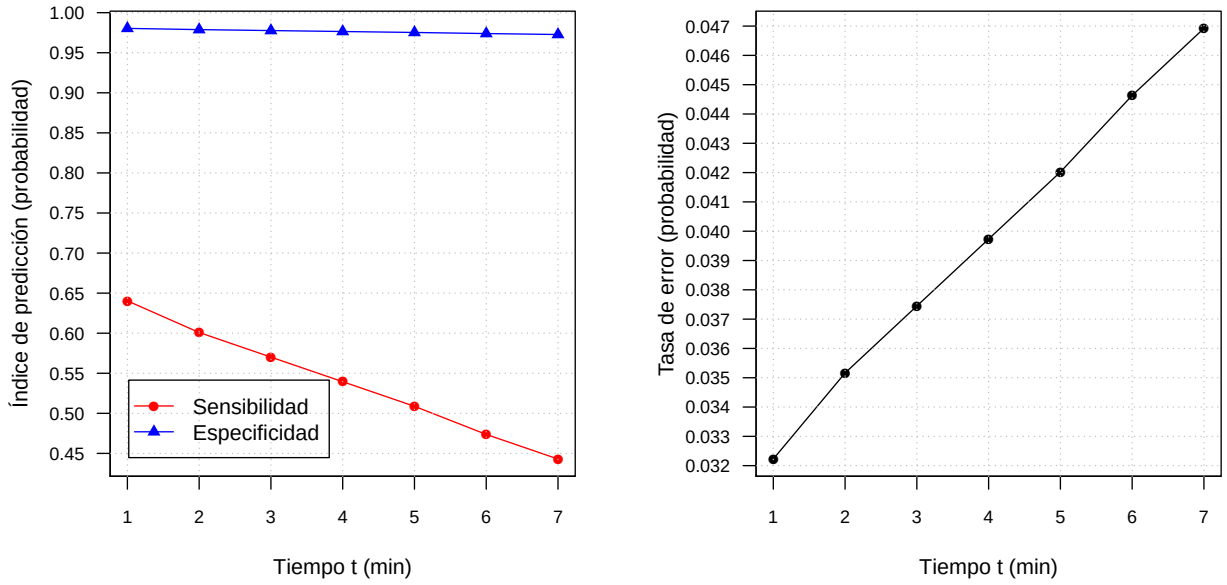

Figura 5.7: Índices de predicción para la serie de tiempo número 306.

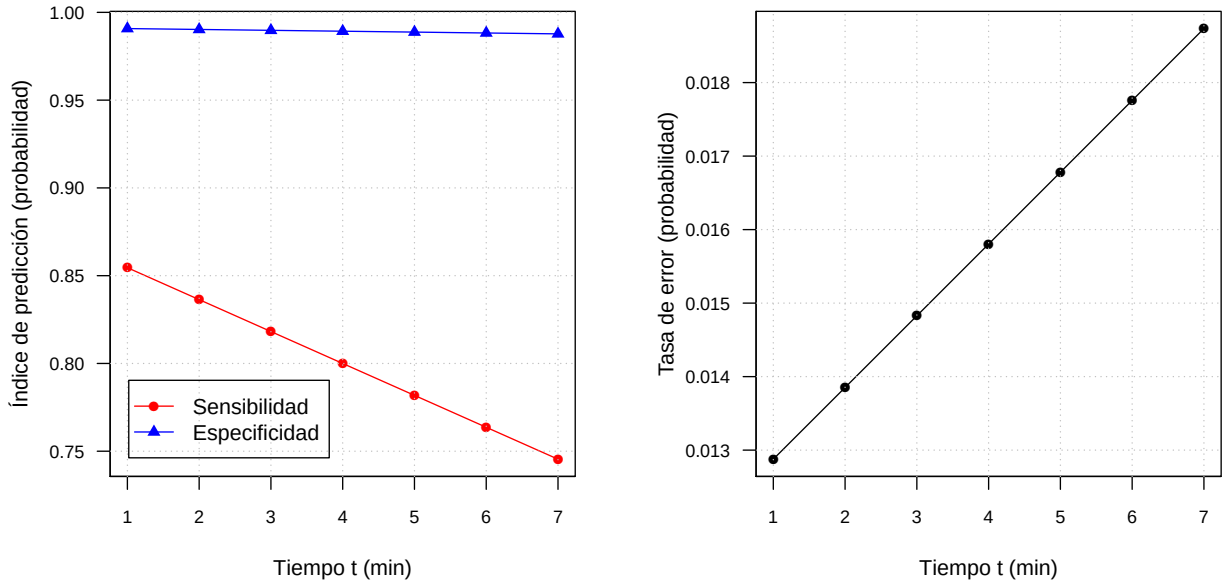

Figura 5.8: Índices de predicción para la serie de tiempo número 12.

Los resultados de la comparación se observan en la Figura 5.7 para la serie de tiempo con $id = 306$ y en la Figura 5.8 para la serie de tiempo con $id = 12$. En ambos casos se evaluó

usando las métricas de la SEN, la SPE y la ERRORRATE. También en ambos casos se presenta un comportamiento esperado en términos de la disminución de los índices de SEN y de ESP dado que a mayor tiempo de predicción aumenta la dificultad de predecir lo cual se ve reflejado en el aumento de la ERRORRATE. Por otra parte, algo importante a señalar es que el desbalanceo de las clases produce que la ESP disminuya en un rango de $0{,}03$ de probabilidad de predicción para las series de tiempo con $id = \{306, 12\}$ mientras que la SEN decae con mayor rapidez en rangos de $0{,}10$ y $0{,}20$ de probabilidad para las series de tiempo con $id = \{30, 612\}$, respectivamente.

En esta parte de la experimentación se seleccionaron un total de 173 series de tiempo y usando los modelos de la etapa de codificación e identificación de estados hipotensivos que se detallan en el Anexo A se realizó la predicción a n pasos. El número de códigos total generados para las 173 series de tiempo corresponde a 767,234. Estas series de tiempo tienen como particularidad que registran en sus etiquetas a *priori* al menos diez AHE. Esta restricción es debido a que las series de tiempo con un bajo número de etiquetas *a priori* presentan en algunos casos ESP de uno lo cual dificulta evaluar de forma correcta la predicción.

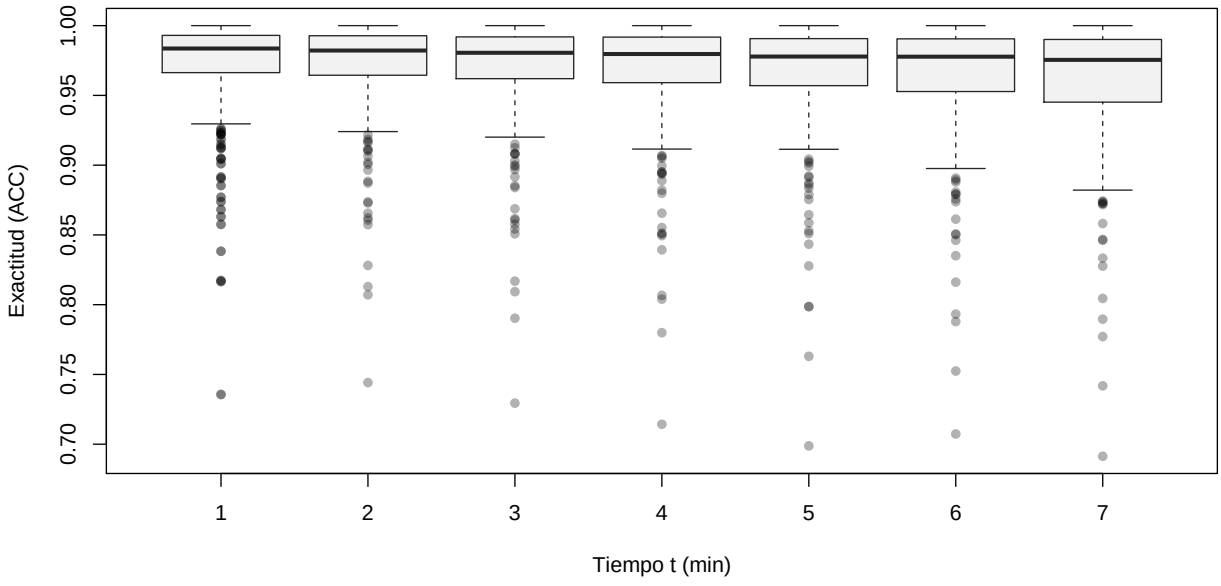

Figura 5.9: Exactitud de predicción para $k = 4$ grupos.

En la Figura 5.9 se ilustra en una gráfica de cajas la exactitud de la predicción en función del tiempo t. En esta experimentación se observó que la media obtenida en la predicción disminuye conforme se considera una ventana de tiempo mayor, lo cual es un comportamiento esperado debido a la dificultad de predecir a futuro. Para la obtención de dicho valor medio se consideraron los resultados de predicción para un total de 173 series de tiempo de MAP.

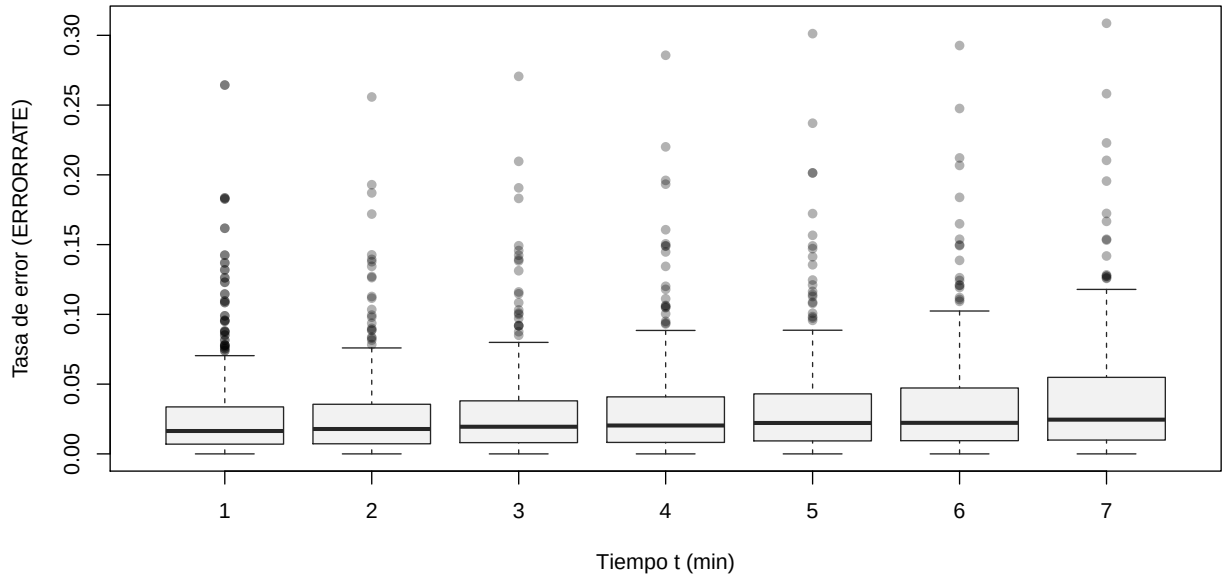

Figura 5.10: Error de predicción para $k = 4$ grupos.

En la Figura 5.10 se ilustra el error dado el mismo conjunto de 173 series de tiempo. Esta gráfica muestra la medida de tasa de error el cual es calculado para cada una de las series de tiempo.

En las dos últimas Figura 5.9 y 5.10 se muestra la distribución de los índices de exactitud y tasa de error obtenido al predecir diferentes series de tiempo. Si bien esta información es útil no permite observar la relación del número de eventos etiquetados *a priori* y el valor obtenido usando las métricas de evaluación. Por tal motivo se presenta a continuación cuatro Figuras 5.11, 5.12, 5.13 y 5.14, que ilustran la relación existente. En estas cuatro figura se observan siete gráficas donde cada una correspondiente a los siete momentos en los cuales se realizó la predicción $t = \{1, 2, 3, \ldots, 7\}$. Cada una de estas gráficas ilustra en el eje de abscisas el valor obtenido al evaluar una serie de tiempo

usando las métrica de clasificación y predicción correspondiente; y en el eje de ordenadas una escala del total de etiquetas *a priori* identificadas como AHE. Finalmente, estos siete conjunto de gráficas para las cuatro figuras disponibles denotan la media del eje de las abscisas con una línea.

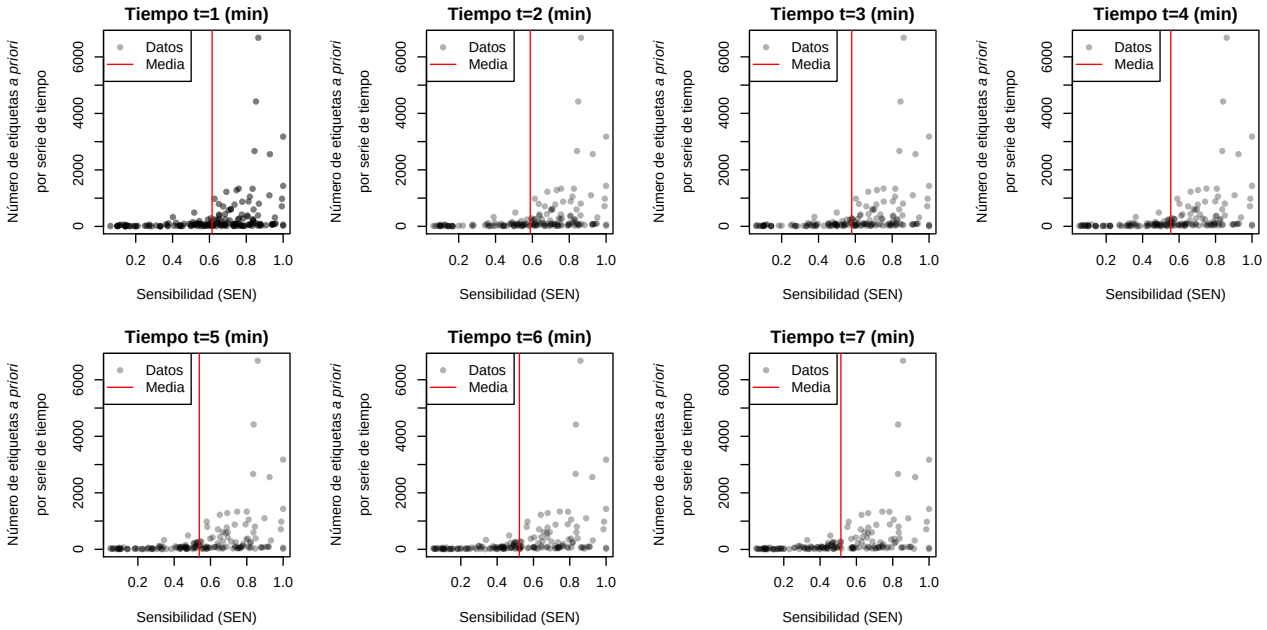

Figura 5.11: Sensibilidad vs. número de etiquetas por serie de tiempo para $k = 4$ grupos y tiempo $1 \leq t \leq 7$.

En la Figura 5.11, se ilustra la distribución de la SEN y se observa un comportamiento donde un bajo número de etiquetas *a priori* se predicen de forma irregular aunque conforme aumenta el número de etiquetas AHE disponible en la serie de tiempo aumenta el índice de sensibilidad ubicándose la media de los datos en 0.62 hasta 0.53 en los tiempo $t = 1$ y $t = 7$, respectivamente. Esta distribución de la SEN alcanzada es importante dado el desbalanceo de las etiquetas *a priori* en donde predomina la etiqueta NoAHE con un 0.97.

En la Figura 5.12 se ilustra la ESP y se observa el nivel de predicción alcanzado de las etiquetas NoAHE identificadas *a priori*. En este caso la media de la ESP se ubica en rangos superiores a 0.98 logrando predecir correctamente las etiquetas NoAHE. También se aprecia que contrario a la SEN la nube de puntos o intersecciones de los datos de la ESP se encuentra menos dispersa y se concentra

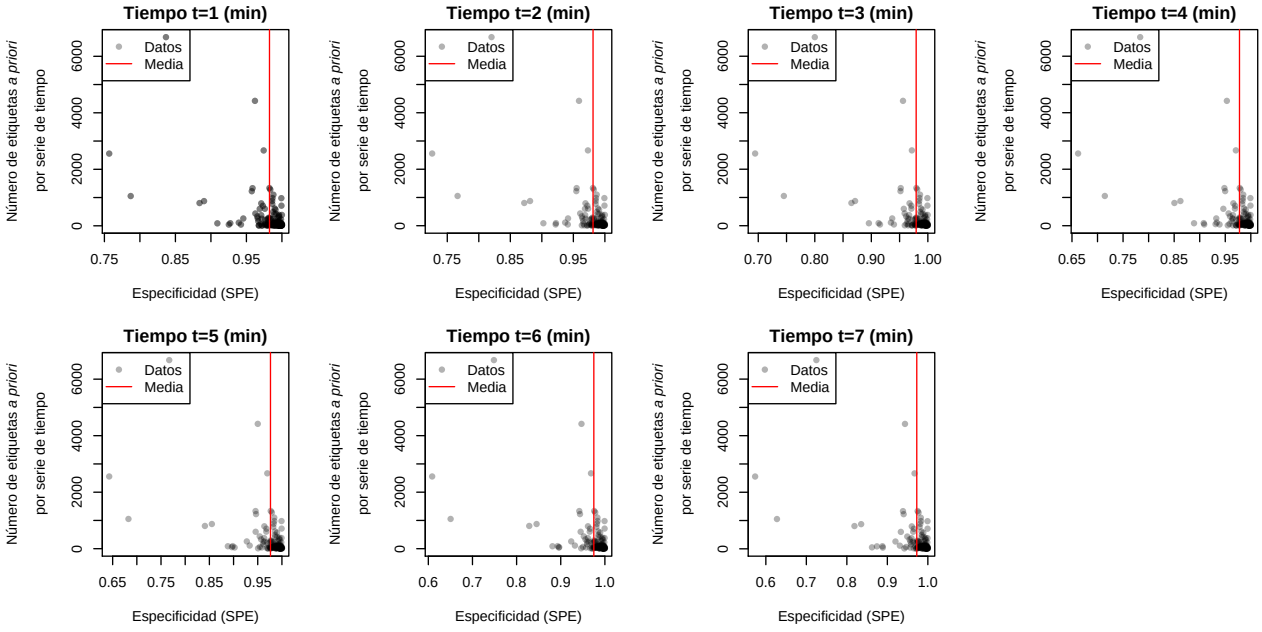

Figura 5.12: Especificidad vs. número de etiquetas por serie de tiempo para $k = 4$ grupos y tiempo $1 \leq t \leq 7$.

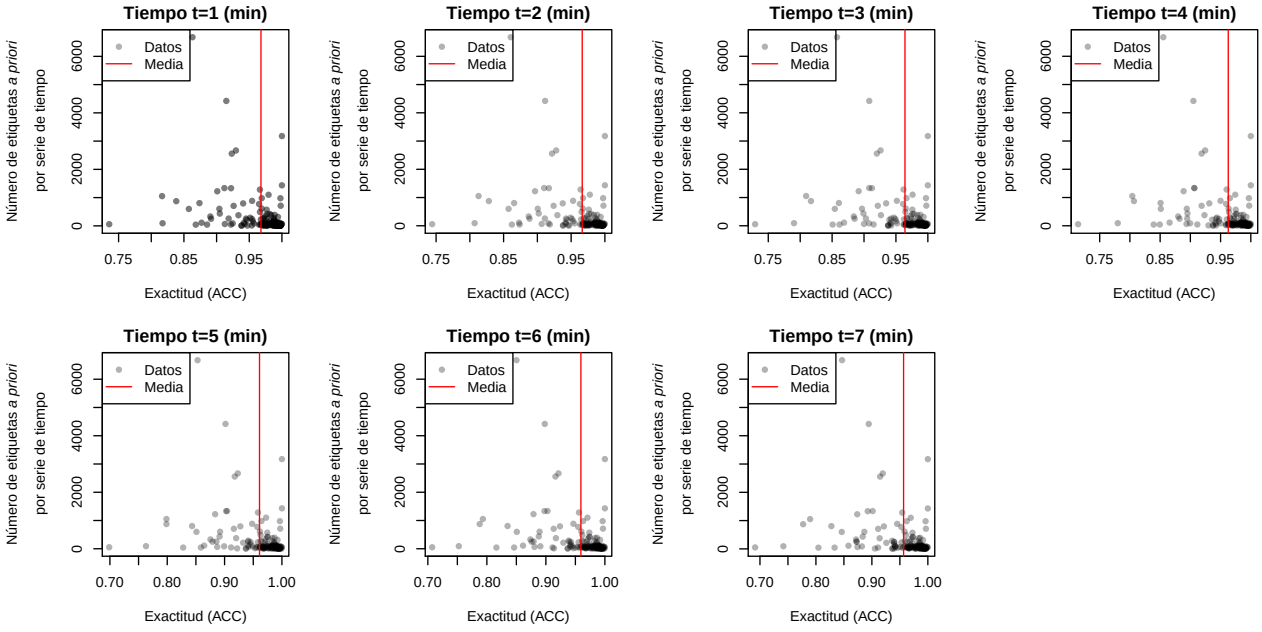

Figura 5.13: Exactitud vs. número de etiquetas por serie de tiempo para $k = 4$ grupos y tiempo $1 \leq t \leq 7$.

en la parte inferior izquierda lo que sugiere un comportamiento menos variante a un aumento o disminución del número de etiquetas presentes durante la evaluación de diferentes series de tiempo.

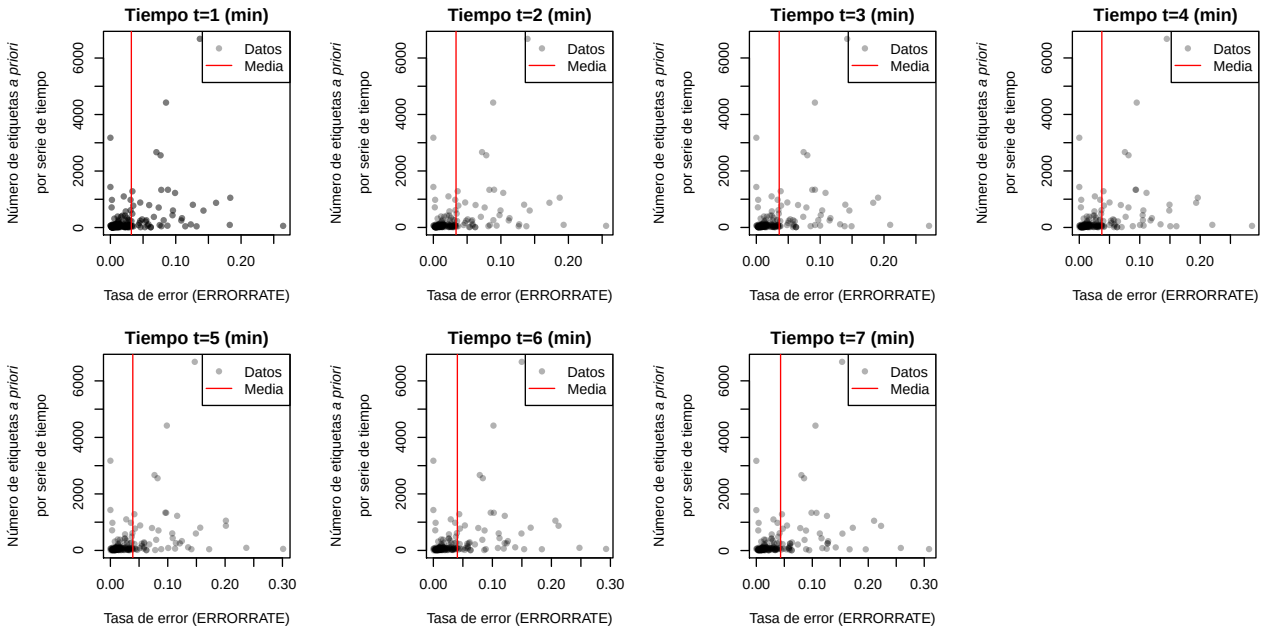

Figura 5.14: Tasa de error vs. número de etiquetas por serie de tiempo para $k = 4$ grupos y tiempo $1 \leq t \leq 7$.

Por último, en las Figuras 5.13 y 5.14, se observa en el eje de las abscisas el comportamiento que se muestra y se describe anteriormente usando las Figuras 5.9 y 5.10, respectivamente. Este comportamiento de las Figuras 5.13 y 5.14 se muestran como un espejo a su vez que la media de las métricas usadas (ACC y ERRORRATE) se acercan a una probabilidad de uno en el caso de ACC y a cero en el caso de la ERRORRATE.

5.5 Comparativa con enfoques de predicción de AHEs

Después de presentar los resultados obtenidos en la experimentación, en esta sección se discuten los resultados de predicción usando el enfoque propuesto contra los resultados alcanzados en la literatura. En este sentido, uno de los principales retos consiste en definir un marco de referencia

que permita una comparativa justa, si bien en la práctica esto es difícil de lograr debido a escenarios heterogéneos, las cifras presentadas en esta sección son una aproximación que nos permite intuir el comportamiento de los resultados.

Tabla 5.2: Comparativa con trabajos relacionados.

Autor	O_w (min)	G_n (min)	ACC	SEN	SPE
Propuesta	30	1	**0.97**	0.61	**0.98**
	30	3	**0.97**	0.58	**0.98**
	30	5	**0.96**	0.53	**0.98**
Feras Hatib [31]	30	5	NA	0.92	0.92
	30	10	NA	0.89	0.90
	30	15	NA	0.88	0.87
Dazhi Jiang [35]	60	0	0.89	0.92	0.88
	60	0	0.81	0.78	0.82
Sakyajit Bhatt [7]	45	120	0.94	0.94	0.95
	50	5	0.92	0.97	0.89
	20	115	0.92	0.84	0.96
Teresa Rocha [62]	140	0	NA	0.82	0.78
Joon Lee [45]	30	60	0.86	0.81	0.86
	60	60	0.86	0.83	0.86
	30	120	0.83	0.79	0.83
	60	120	0.83	0.79	0.83

Para esta comparativa se seleccionó del estado del arte consultado en esta tesis los resultados más representativos y se contrastaron algunos resultados de las métricas de evaluación publicados por los autores. En la Tabla 5.2 se observan las columnas de: autor, tamaño de la ventana de observación (O_w) en minutos, tamaño del intervalo entre las ventanas (G_n) en minutos, la exactitud (ACC), sensibilidad (SEN) y especificidad (SPE). En estos destacamos la ACC y la SPE obtenida por nuestro método, con un tamaño de intervalo entre ventanas de cinco minutos nuestro modelo mostró un rendimiento de 96 % de ACC y 98 % de SPE. Y con un intervalo entre ventana de tres minutos mostró un rendimiento de 97 % de ACC y 98 % de SPE.

5.6 Resumen

En este capítulo se ha presentado y descrito la validación y evaluación de los resultados del método propuesto en el Capítulo 4. Este capítulo está organizado siguiendo la metodología planteada en las Figuras 4.1 y 4.1, donde cada una de las dos etapas son validadas. Finalmente, se presenta un conjunto de experimentación para un total de 173 series de tiempo con un total de 767234 códigos donde el 97 % corresponde a las etiquetas con estado NoAHE y el 3 % corresponde a las etiquetas AHE logrando índices de exactitud y especificidad de (0.97, 0.98), tres (0.97, 0.98) y cinco minutos (0.96, 0.98).

6

Conclusiones y trabajo a futuro

En este capítulo se presentan las conclusiones de la tesis y se detallan las principales contribuciones alcanzadas en la misma. Asimismo, se identifican posibles limitaciones del método propuesto así como posibles líneas de trabajo a futuro de la investigación realizada.

6.1 Conclusiones

En esta tesis se ha abordado el problema de representación de series de tiempo de MAP para la predicción de AHEs. La propuesta se ha centrado en obtener una representación que contenga información en el dominio de la frecuencia y tiempo. Y también en la selección de una técnica que modele procesos estocásticos a partir de las propiedades aleatorias de las series de tiempo de la MAP.

- El método propuesto cumple con el objetivo de obtener una representación que mediante una nueva codificación de los AHEs y el agrupamiento de los datos categóricos resultantes se mejore el rendimiento de predicción de AHEs usando un enfoque basado en cadenas de Markov.

- La representación usada contiene información en el dominio del tiempo y la frecuencia. La información contenida en el dominio del tiempo se logra al obtener un código de 30 dígitos que contiene valores de códigos anteriores ordenados. Estos valores corresponden a cero o uno resultado de la transformación binaria que contiene a su vez información en el dominio de la frecuencia.

- El método implementado es flexible y permite ajustar la representación de estados en diferentes valores k.

- La representación de códigos usada da indicios de la posible existencia de estados diferentes al binario (AHE/NoAHE), en particular se resalta los resultados obtenidos con $k = 4$.

6.2 Principales contribuciones

Como resultado del trabajo de investigación presentado se obtuvieron las siguientes contribuciones:

- Un método que permite codificar los episodios agudos hipotensivos (AHE, por sus siglas en inglés) de forma tal que los códigos generados preservan las propiedades de los componentes de frecuencia y tiempo de una serie de presión arterial media (MAP, por sus siglas en inglés).

- Validación de un algoritmo de agrupación de datos categóricos para agrupar los códigos resultantes e identificar los k grupos de interés para su uso en tareas de predicción.

- La integración del método de codificación en un esquema basado en cadenas de Markov permite predecir con precisión la ocurrencia de un AHE en una ventana de hasta 7 minutos de antelación.

6.3 Limitaciones del método propuesto

Este trabajo de tesis se utilizó un conjunto de datos de MAP de la MIMIC-II, al cual se le aplicaron diversas tareas de pre-procesamiento. Debido a que la información en esa base de datos de signos vitales no se encuentra etiquetada, es necesario asignar las marcas de clase en función del tipo de episodio adverso de interés. En este contexto, la aplicabilidad del método propuesto está limitado a la definición de AHE considerada, como se describe en el Capítulo 1, misma que fue usada para definir las etiquetas de clase en las series de tiempo de MAP consideradas en la presente tesis. En cualquier caso, el uso de otra definición de AHE puede ser empleado para generar los códigos correspondientes con la codificación propuesta.

6.4 Trabajo a futuro

En trabajos futuros se propone validar las posibles ventajas o desventajas asociadas al aumento del número de grupos para $k > 10$. De forma empírica se observó que cuando se agrupan los códigos en un mayor número de grupos, es posible predecir un AHE con mayor antelación en términos de una ventana de predicción mayor. Sin embargo, el aumento en el valor de la ventana de predicción ocasiona una disminución en el rendimiento de la clasificación y precisión de la predicción. Por tanto, el compromiso que debe existir entre un rendimiento aceptable y una ventana de predicción de cierta longitud se establece como parte del trabajo a futuro de la presente tesis.

A

Modelado de predicción usado durante la experimentación

A continuación se presentan los valores obtenidos durante la agrupación de códigos y el mapeo a estado binario. Esta configuración se encontró para un valor de $k = 4$ grupos. En la Tabla A.1, se muestran los valores de los centroides C_K obtenidos para los cuatro grupos, el número de columnas o dimensiones D se fija de forma automática y corresponde al número de elementos o caracteres del código en representación decimal. Los valores de los códigos en decimal oscilan en un rango de 0 a $2^{30} = 1,073,741,824$.

Tabla A.1: Valores de los centroides asociados a los grupos C_K.

C_K	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
1	1	0	7	3	7	4	1	7	9	1
2	1	0	7	3	7	4	1	8	2	3
3	0	0	0	0	0	0	0	0	0	0
4	1	0	7	3	7	7	9	0	7	5

En la tabla A.2, se presenta los valores de la matriz de estados S que contiene los vectores C_K para los cuatros grupos y el vector M_K que representa el mapeo a estado binario de cada grupo. En este caso el estado S_3 que hace referencia al grupo tres que contiene los estados etiquetados como AHE.

Tabla A.2: Matriz de estados S.

	S_1	S_2	S_3	S_4
C_K	1	2	3	4
M_K	0	0	1	0

Por último, en la Tabla A.3 se muestra los valores de la matriz de probabilidades de transición entre estados (S) que fue calculado a partir de los códigos previamente suministrados que se extraen de un conjunto de 173 series de tiempo que son usados durante toda la experimentación y resultados.

Tabla A.3: Matriz de probabilidades de transición entre estados A.

	S_1	S_2	S_3	S_4
S_1	0.644578967792006	0.233597206053551	0.029553744664338	0.092270081490105
S_2	0.018405832009586	0.974906626258068	0.001199452410643	0.005488089321703
S_3	0.099449571659148	0.020485866821173	0.875367297107147	0.004697264412532
S_4	0.380943956977073	0.357415793942825	0.009163600339655	0.252476648740447

B

Algoritmo y funciones de agrupación de códigos

En este anexo se muestran tres algoritmos usados para la agrupación de los códigos. El Algoritmo 1 y Algoritmo 2 corresponde a las funciones de cálculo de la distancia de Chebyshev y la inicialización de los centroides, estos algoritmos reciben parámetros desde la función $categoricalKMeans()$ en el Algoritmo 3.

El Algoritmo 3 es el algoritmo que contiene la función principal $categoricalKMeans(k, X)$, el parámetro k corresponde al número de grupos y el parámetro X contiene una matriz de tamaño n x 10, donde n corresponde al número de códigos y 10 al número de caracteres o dígitos en los cuales se puede dividir un código en notación decimal.

Algorithm 1 Cálculo de la distancia de Chebyshev.

```
function chebyshevDistance(x, y)
    d = 0
    Inicializar diff[|x|]
    for i = 1 to |x| do
        diff[i] = abs(x[i] − y[i])
    end for
    d = argmax(diff)
    return d
end function
```

Algorithm 2 Inicialización de los centroides.

```
function categoricalCentersInitialization(k, d)
    Inicializar c_[|k|,|d|]
    Inicializar diff[|x|]
    max = d − 1
    min = 0
    for i = 1 to |k| do
        for j = 1 to |d| do
            Se genera un valor aleatorio R^ℕ|R ∈ {min, ..., max}
            c[i, j] = R
        end for
    end for
    return c
end function
```

Algorithm 3 Algoritmo para agrupación categórica usando k-means.

function CATEGORICALKMEANS(k, X)
    Inicializar $ndim = |X[1, :]|$
    Inicializar $ncol = |X[:, 1]|$
    Inicializar $l[ncol]$
    $c = categoricalCentersInitialization(k, ndim)$
    $t = 5$
    while $n <= t$ **do**
        for $j = 1$ to $ncol$ **do**
            $d[k]$
            for $i = 1$ to k **do**
                $d[i] = chebyshevDistance(X[j, :], c[i, :])$
            end for
            Índice p del argumento mínimo $argmin(d)$
            $l[j] = p$
        end for
        for $i = 1$ to k **do**
            Índice a donde $i == l$
            $dc = X[a, :]$
            $c[i, :] = dc$
        end for
    end while
end function
$categoricalKMeans(k, X)$

Bibliografía

[1] Abbasinia, M. et al. (2016). Predicting acute hypotensive episode by using hybrid features and a neuro-fuzzy network. *Turkish journal of electrical engineering & computer sciences*, 24:3335–3344.

[2] Adibuzzaman, M. et al. (2014). The mixing rate of the arterial blood pressure waveform markov chain is correlated with shock index during hemorrhage in anesthetized swine. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3268–3271.

[3] AM, M. et al. (1950). The normal blood pressure range and its clinical implications. *Journal of the American Medical Association*, 143(17):1464–1470.

[4] Ascha, M. et al. (2018). Pulse oximetry and arterial oxygen saturation during cardiopulmonary exercise testing. *Medicine and science in sports and exercise*, 50(10):19921997.

[5] Barnett, V. (1994). *Outliers in Statistical Data*. Wiley.

[6] Begum, S. et al. (2011). Case-based reasoning systems in the health sciences: A survey of recent trends and developments. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41(4):421–434.

[7] Bhattacharya, S. et al. (2014). A novel classification method for predicting acute hypotensive episodes in critical care. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '14, pages 43–52, New York, NY, USA. ACM.

[8] Bhattacharya, S., Huddar, V., Rajan, V., and Reddy, C. K. (2018). A dual boundary classifier for predicting acute hypotensive episodes in critical care. *PLOS ONE*, 13(2):1–17.

[9] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

[10] Boughorbel, S., Jarray, F., and El-Anbari, M. (2017). Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PLOS ONE*, 12(6):1–17.

[11] Bright, T. J. et al. (2012). Effect of Clinical Decision-Support Systems: A Systematic Review. *Annals of Internal Medicine*, 157(1):29–43.

[12] Broaddus, V. (2016). *Murray 'i&' Nadel's Textbook of Respiratory Medicine*. Elsevier/Saunders, Philadelphia, PA.

[13] Brockwell, P. J. (2016). *Introduction to Time Series and Forecasting (springer Texts in Statistics)*. Springer.

[14] Chen, X. et al. (2009). Forecasting acute hypotensive episodes in intensive care patients based on a peripheral arterial blood pressure waveform. *Computers in Cardiology (CinC)*, 36:545–548.

[15] Clifford, G. D. et al. (2010). User guide and documentation for the mimic ii database (version 2, release 1). Technical report, Department of Engineering Science and Kellogg College, University of Oxford. N.B. Dr Clifford is now based at the Department of Engineering Science and Kellogg College, University of Oxford.

[16] Cohen, A. (1998). Hidden markov models in biomedical signal processing. In *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Vol.20 Biomedical Engineering Towards the Year 2000 and Beyond (Cat. No.98CH36286)*, volume 3, pages 1145–1150 vol.3.

[17] da S. Luz, E. J. et al. (2016). Ecg-based heartbeat classification for arrhythmia detection: A survey. *Computer Methods and Programs in Biomedicine*, 127:144–164.

[18] Daskalaki, S. et al. (2006). Evaluation of classifiers for an uneven class distribution problem. *Applied Artificial Intelligence*, 20(5):381–417.

[19] Dernoncourt, F. et al. (2015). Gaussian process-based feature selection for wavelet parameters: Predicting acute hypotensive episodes from physiological signals. In *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*, pages 145–150.

[20] Donald, R. et al. (2019). Forewarning of hypotensive events using a bayesian artificial neural network in neurocritical care. *Journal of Clinical Monitoring and Computing*, 33(1):39–51.

[21] Fan, Z. et al. (2015). Prediction of acute hypotensive episodes using random forest based on genetic programming. In *2015 IEEE Congress on Evolutionary Computation (CEC)*, pages 688–694.

[22] Forkan, A. R. M. et al. (2017). Visibid: A learning model for early discovery and real-time prediction of severe clinical events using vital signs as big data. *Computer Networks*, 113:244–257.

[23] Ghaffari, A. et al. (2010a). A methodology for prediction of acute hypotensive episodes in icu via a risk scoring model including analysis of st-segment variations. *Cardiovascular Engineering*, 10(1):12–29.

[24] Ghaffari, A. et al. (2010b). Parallel processing of ecg and blood pressure waveforms for detection of acute hypotensive episodes: A simulation study using a risk scoring model. *Computer Methods in Biomechanics and Biomedical Engineering*, 13(2):197–213. PMID: 19697181.

[25] Ghassemi, M. (2011). Methods and models for acute hypotensive episode prediction. Master's thesis, Oxford University, UK.

[26] Ghosh, S. et al. (2014). Risk prediction for acute hypotensive patients by using gap constrained

sequential contrast patterns. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2014:1748–1757. 25954447[pmid] PMC4419954[pmcid].

[27] Ghosh, S. et al. (2016). Hypotension risk prediction via sequential contrast patterns of icu blood pressure. *IEEE Journal of Biomedical and Health Informatics*, 20(5):1416–1426.

[28] Gómez, J. Z. V. (2017). Detección Temprana de Episodios Hipotensivos Agudos Basada en un Modelo de Aprendizaje Automático. Master's thesis, Cinvestav unidad Tamaulipas.

[29] Group, N. P. (cited Oct 2019a). Kernel density estimation.

[30] Group, N. P. (cited Oct 2019b). Predictive medicine.

[31] Hatib, F. et al. (2018). Machine-learning algorithm to predict hypotension based on high-fidelity arterial pressure waveform analysis. *Anesthesiology*.

[32] Jiang, D. et al. (2014). Detection of acute hypotensive episodes via empirical mode decomposition and genetic programming. In *2014 International Conference on Identification, Information and Knowledge in the Internet of Things*, pages 225–228.

[33] Jiang, D. et al. (2015). An approach for prediction of acute hypotensive episodes via the hilbert-huang transform and multiple genetic programming gariclassifier. *International Journal of Distributed Sensor Networks*, 11(8):354807.

[34] Jiang, D. et al. (2017a). Prediction of acute hypotensive episodes using emd, statistical method and multi gp. *Soft Computing*, 21(17):5123–5132.

[35] Jiang, D. et al. (2017b). Probability distribution pattern analysis and its application in the acute hypotensive episodes prediction. *Measurement*, 104:180–191.

[36] Johnson, A. et al. (2016). MIMIC-III, a freely accessible critical care database. `http://www.nature.com/articles/sdata201635`. [Online; accessed 7-Jun-2018].

[37] Kang, H. et al. (2015). Indoor localization of dron using vision based sensor fusion. *2015 15th International Conference on Control, Automation and Systems ICCAS*, pages 932–934.

[38] Karlen, W. et al. (2010). Capnobase: Signal database and tools to collect, share and annotate respiratory signals. 2010 annual meeting of the Society for Technology in Anesthesia (STA); Conference Location: West Palm Beach, FL, USA; Conference Date: January 13-16, 2010; .

[39] Kawamoto, K. et al. (2005). Improving clinical practice using clinical decision support systems: A systematic review of trials to identify features critical to success. *BMJ*, 330(7494):765.

[40] Keogh, E. et al. (2001). An online algorithm for segmenting time series. In *Proceedings 2001 IEEE International Conference on Data Mining*, pages 289–296.

[41] Kim, S.-H. et al. (2016). HeartCast: Predicting acute hypotensive episodes in intensive care units. *Statistical Methodology*, 33:1–13.

[42] Kim, Y. B. et al. (2014). Large-scale methodological comparison of acute hypotensive episode forecasting using mimic2 physiological waveforms. In *2014 IEEE 27th International Symposium on Computer-Based Medical Systems (CBMS)*, volume 00, pages 319–324.

[43] Kong, G. et al. (2008). Clinical decision support systems: A review on knowledge representation and inference under uncertainties. *International Journal of Computational Intelligence Systems*, 1(2):159–167.

[44] Laboratory For Computational Physiology, M. (2011). The mimic ii clinical database.

[45] Lee, J. and Mark, R. G. (2010). An investigation of patterns in hemodynamic data indicative of impending hypotension in intensive care. *BioMedical Engineering OnLine*, 9(1):62.

[46] Lehman, L. H. et al. (2008). Similarity-based searching in multi-parameter time series databases. In *2008 Computers in Cardiology*, pages 653–656.

[47] Marín, J. (cited Nov 2019). Cadenas de markov.

[48] Matthews, B. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451.

[49] McGregor, J. C. et al. (2006). Impact of a Computerized Clinical Decision Support System on Reducing Inappropriate Antimicrobial Use: A Randomized Controlled Trial. *Journal of the American Medical Informatics Association*, 13(4):378–384.

[50] Merino, A. P. and C., M. A. R. D. (2006). *Analisis De Datos Con Spss 13 Base (spanish Edition)*. McGraw-Hill Interamericana.

[51] Moghadam, M. C. et al. (2019). A machine learning approach to predict hypotensive events in icu settings. *bioRxiv*.

[52] Moody, G. and Lehman, L. (2009). Predicting acute hypotensive episodes: The 10th annual physionet/computers in cardiology challenge. In *2009 36th Annual Computers in Cardiology Conference (CinC)*, pages 541–544.

[53] Murdoch, T. and Detsky, A. (2013). The inevitable application of big data to health care. *JAMA : the journal of the American Medical Association*, 309:1351–2.

[54] Palaniappan, S. and Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. In *2008 IEEE/ACS International Conference on Computer Systems and Applications*, pages 108–115.

[55] Pathinarupothi, R. K. and Rangan, E. S. (2017). Consensus motifs as adaptive and efficient predictors for acute hypotensive episodes. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1688–1691.

[56] Paul, A. K. et al. (2016). Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease. In *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, pages 145–150.

[57] Petkus, V. et al. (2016). Method for prediction of acute hypotensive episodes. *Elektronika ir Elektrotechnika*, 22(1).

[58] Pickering, T. G. et al. (2006). Ambulatory blood-pressure monitoring. *New England Journal of Medicine*, 354(22):2368–2374. PMID: 16738273.

[59] Piper, I. et al. (2010). The brain monitoring with information technology (brainit) collaborative network: Ec feasibility study results and future direction. *Acta Neurochirurgica*, 152(11):1859–1871.

[60] Reason, J. (1995). Understanding adverse events: Human factors. *Quality and Safety in Health Care*, 4(2):80–89.

[61] Rocha, T. et al. (2010). Wavelet based time series forecast with application to acute hypotensive episodes prediction. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pages 2403–2406.

[62] Rocha, T. et al. (2011). Prediction of acute hypotensive episodes by means of neural network multi-models. *Computers in Biology and Medicine*, 41(10):881–890.

[63] Romano, M. J. and Stafford, R. S. (2011). Electronic Health Records and Clinical Decision Support Systems: Impact on National Ambulatory Care QualityClinical Decision Support and Ambulatory Care. *JAMA Internal Medicine*, 171(10):897–903.

[64] Rosenthal, D. I. et al. (2006). Radiology order entry with decision support: Initial clinical experience. *Journal of the American College of Radiology*, 3(10):799–806.

[65] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.

[66] Shaffer, F. and Ginsberg, J. P. (2017). An overview of heart rate variability metrics and norms. *Frontiers in Public Health*, 5:258.

[67] Singh, A. et al. (2010). Hidden markov models for modeling blood pressure data to predict acute hypotension. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 550–553.

[68] Soberanes RL, Salazar EDC, C. C. (2016). Morbimortalidad en 10 aï¿½os de atenciï¿½n en la unidad de cuidados intensivos del hospital general agustï¿½n oï¿½horan de mï¿½rida, yucatï¿½n. *Medicina Crï¿½tica*, 20:65–68.

[69] Song, L. et al. (2014). Ndna-prot: Identification of dna-binding proteins based on unbalanced classification. *BMC Bioinformatics*, 15(1):298.

[70] Sun, H. et al. (2013). A method for prediction of acute hypotensive episodes in icu via pso and k-means. In *2013 Sixth International Symposium on Computational Intelligence and Design*, volume 1, pages 99–102.

[71] Vincent, J.-L. et al. (2018). Mean arterial pressure and mortality in patients with distributive shock: A retrospective analysis of the mimic-iii database. *Annals of intensive care*, 8(1):107–107. 30411243[pmid], PMC6223403[pmcid].

[72] Wang, C. et al. (2015). Imdc: An ensemble learning method for imbalanced classification with mirna data. *Genetics and Molecular Research*, 14:123–133.

[73] Wasson, J. H. et al. (1985). Clinical prediction rules. *New England Journal of Medicine*, 313(13):793–799. PMID: 3897864.

[74] Yamamoto, A. et al. (2017). Usefulness of pulse oximeter that can measure spo(2) to one digit after decimal point. *Yonago acta medica*, 60(2):133–134. 28701897[pmid] PMC5502226[pmcid].

[75] York, S. N. (2008). *Time Series*, pages 536–539. Springer New York, New York, NY.

[76] Zenati, M. S. et al. (2002). A brief episode of hypotension increases mortality in critically ill trauma patients. *Journal of Trauma and Acute Care Surgery*, 53(2).

[77] Zhang, Y. et al. (2010). Outlier detection techniques for wireless sensor networks: A survey. *IEEE Communications Surveys Tutorials*, 12(2):159–170.

[78] Zhou, Y. et al. (2013). Prediction of acute hypotensive episode in icu using chebyshev neural network. *JSW*, 8:1923–1931.