ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERIA DE SISTEMAS
UNIDAD DE TITULACION

EVOLUCION DE UN METODO BASADO EN SIMBOLOS PARA CLASIFICAR SERIES TEMPORALES USANDO MINERIA DE DATOS

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL GRADO DE MAGISTER EN COMPUTACIÓN

SANDRA ELIZABETH GALARZA PARRA
sandra.galarza@epn.edu.ec

Director: MARCO MOLINA BUSTAMANTE, PHD
marco.molinab@epn.edu.ec

CO-DIRECTORA: MONSERRATE INTRIAGO PAZMIÑO, MSC
monserrate.intriago@epn.edu.ec

Julio, 2021

APROBACIÓN DEL DIRECTOR

Como director del trabajo de titulación “EVOLUCION UN METODO BASADO EN SIMBOLOS PARA LA CLASIFICACION DE SERIES TEMPORALES USANDO MINERIA DE DATOS” desarrollado por Sandra E. Galarza Parra, estudiante de la Maestría en Ciencias de la Computación, habiendo supervisado la realización de este trabajo y realizado las correcciones correspondientes, doy por aprobada la redacción final del documento escrito para que prosiga con los procedimientos correspondientes para apoyar la Defensa oral.

PhD Marco Molina
DIRECTOR

APROBACIÓN DE LA CO-DIRECTORA

Como codirector del trabajo de titulación “EVOLUCION UN METODO BASADO EN SIMBOLOS PARA LA CLASIFICACION DE SERIES TEMPORALES USANDO MINERIA DE DATOS” desarrollado por Sandra E. Galarza Parra, estudiante de la Maestría en Ciencias de la Computación, habiendo supervisado la realización de este trabajo y realizado las correcciones correspondientes, doy por aprobada la redacción final del documento escrito para que prosiga con los procedimientos correspondientes para apoyar la Defensa oral.

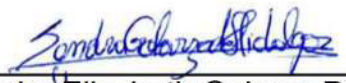
MARIA MONSERRATE INTRIAGO PAZMIÑO

Digitally signed by MARIA MONSERRATE INTRIAGO PAZMIÑO
DN: C=EC, OU=EPN, CN=MARIA MONSERRATE INTRIAGO PAZMIÑO, E=monserrate.intriago@epn.edu.ec
Date: 2021.07.29 18:01:34-05'00'
Foxit PDF Reader Version: 11.0.0

Msc. Monserrate Intriago
CO-DIRECTORA

DECLARACIÓN DE AUTORÍA

Yo, Sandra Elizabeth Galarza Parra, declaro bajo juramento que el trabajo aquí descrito es mi responsabilidad; que no ha sido previamente presentado para ningún título o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento. La Escuela Politécnica Nacional puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.



Sandra Elizabeth Galarza Parra

DEDICATORIA

Dedico este trabajo de titulación a mi familia que son el pilar que me sostiene y por los que lucho y sigo adelante, en especial a mi esposo quien me alentó a seguir la carrera.

AGRADECIMIENTOS

Agradezco a Dios por permitir que termine este proyecto de titulación, a mi familia por su apoyo incondicional en cada etapa de mi formación como magister, en especial a mi esposo Marco y mis hijos David, Sandra y Daniel que son el soporte y motor de mi vida impulsándome a seguir adelante, a mi madre María, mi padre Gerardo, mi hermana Gabriela por su apoyo en cada paso que he dado.

Agradezco al PhD Marco Molina director de la presente tesis, por todo su apoyo y guía en la realización de este proyecto de investigación, por compartir sus conocimientos y contribuir de esta manera para que sea una mejor profesional.

De igual manera agradezco a la Msc. Monserrate Intriago codirectora de este trabajo de investigación, por todo su apoyo para la realización de esta tesis.

Contenido

Índice de Tablas

Índice de Figuras

RESUMEN

El análisis del diagnóstico de patologías relacionados con el sistema auditivo evaluados con potenciales evocados auditivos de tronco cerebral (PEATC), se ha realizado por el especialista de forma manual. Para suplir la falta de procesos automáticos de evaluación de los PEATC, se propuso un método basado en símbolos, *Symbolic Pattern-based Classification* (SPC), para soportar el diagnóstico de las patologías mencionadas, que se basa en el análisis de los PEATC convertidos en series temporales simbólicas. En el presente trabajo se modificó el algoritmo para la transformación numérica a simbólica utilizando splines cúbicos para suavizar las series temporales eliminando el ruido residual en la toma de estas pruebas y el algoritmo de clasificación series temporales simbólicas. El método de clasificación proporcionó resultados muy alentadores: una exactitud de 99.4%, sensibilidad de 97.8% y especificidad del 100%, los cuales son muy similares a los obtenidos con el método original, sin embargo, el tiempo de ejecución de los algoritmos implementados supera con mucho al de los originales.

Palabras Clave: Series temporales simbólicas, machine learning, minería de datos, clasificación de series temporales simbólicas, Potenciales Evocados Auditivos de Tronco Cerebral (PEATC).

ABSTRACT

The diagnostic analysis of pathologies related to the auditory system evaluated with brainstem auditory evoked potentials (BAEPs) has been performed manually by the specialist. To make up for the lack of automatic processes for the evaluation of BAEPs, a symbol-based method, Symbolic Pattern-based Classification (SPC), was proposed to support the diagnosis of the mentioned pathologies, which is based on the analysis of BAEPs converted into symbolic time series. In the present research, the algorithm for the numerical to symbolic transformation was modified using cubic splines to smooth the time series by eliminating the residual noise in the taking of these tests and the symbolic time series classification algorithm. our classification method gave very promising results: 99.4% accuracy, 97.8% sensitivity and 100% specificity, which are very similar to the results obtained with the original method, however the execution time of the implemented algorithms are much better than that of the original ones.

Keywords: symbolic time series, machine learning, data mining, symbolic time series classification, Brainstem Auditory Evoked Auditory Potentials (BAEP).

CAPITULO 1

1 Introducción

Los PEATCs son pruebas que se realizan para el diagnóstico de patologías relacionadas con el sistema auditivo. Es uno de los procesos más utilizados dentro del ámbito de la audiología ya que no son invasivos y se los puede realizar aun con el paciente sedado, por lo que es ideal para realizarlos con bebes; también se los realiza a niños y adultos. La utilización de estas pruebas permite la detección temprana de trastornos auditivos como pérdida de la audición y tumores vasculares. Para la realización de las pruebas, se sigue un riguroso procedimiento utilizando equipos especializados. El análisis de los resultados de estos procedimientos es visual, lo cual significa que la experiencia del experto es el medio necesario en el proceso para determinar los resultados. Por lo tanto, resulta sumamente importante automatizar el proceso de diagnóstico, pues sería un gran aporte. Para ello se han propuesto varios tipos de análisis uno de ellos y el menos costoso computacionalmente puede ser transformar los PEATCs numéricos, que en informática son series temporales numéricas, a series temporales simbólicas.

El procedimiento consiste en colocar sensores fijos en tres puntos de la cabeza, luego se envían varios estímulos sonoros (clics) a través del conducto auditivo, lo que provoca la respuesta del tronco cerebral, mediante una onda eléctrica denominada Respuesta del Tronco Cerebral Auditivo (*Auditory Brainstem Response*, ABR), en forma de series temporales. Cada ABR tiene una duración de

0 a 15 ms, dependiendo de la calibración del aparato (Bukard, Don, & Eggermont, 2007) (Molina, Perez, & Valente, 2016).

Una serie temporal es una secuencia T = (t1, t2, ..., t*n*) que es un conjunto ordenado de *n* números reales. Típicamente, el orden es en el tiempo; sin embargo, otros tipos de datos, como distribuciones de color (Martínez, Chuvieco, Aguado, & Salas, 2017), formas (Molina et al., 2016) y espectrógrafos, también tienen un orden bien definido, y pueden considerarse como "series temporales", al aplicar algoritmos de aprendizaje (Eamonn Keogh, 2017) (Ivanovic & Kurbalija, 2016).

En la minería de datos, el manejo de series temporales constituye un desafío importante debido al hecho de que tienen, en general, un orden cronológico, que debe tenerse en cuenta al analizarlas; de igual manera se debe considerar la presencia de formas que no contribuyen a la comprensión del contenido de la serie y, por lo tanto, deben ser reconocidas y eliminadas.

Uno de los problemas más conocidos en el análisis de series temporales es la gran dimensionalidad, para lo cual se han propuesto muchas soluciones que convergen en la segmentación de las series.

En la presente tesis se utiliza el método de simbolización temporal, que al tiempo que disminuye la dimensionalidad, contribuye a la comprensión del significado de la serie temporal (Deng, Wang, & Xu, 2016) (Molina et al., 2016).

La simbolización temporal es el proceso de transformación de la serie temporal numérica en una secuencia temporal asociada a un alfabeto finito, que se conoce también como serie temporal simbólica, a la que se puede aplicar técnicas

especializadas de minería de datos como: indexación, clasificación, agrupamiento, reglas de asociación y detección de anomalías.

En esta investigación se aplicarán métodos de minería de datos sobre series temporales tales como la transformación de las series temporales a secuencias temporales simbólicas, existen varias formas de simbolizar las series como segmentarlas, por motivos, su morfología, entre otras; para la presente tesis se toma en cuenta la forma, la duración, la amplitud, entre otros parámetros, así como también la similitud entre ellas con el cálculo de la distancia. Como dominio para este estudio se escogió los potenciales evocados auditivos del tronco cerebral (PEATC).

Existe un método basado en símbolos para automatizar el diagnóstico de las patologías relacionadas con el sistema auditivo según (Molina et al., 2016). Este método se basa en el análisis de respuestas auditivas de tronco cerebral (ABR), recogidas como series temporales numéricas y convertidas en secuencias temporales simbólicas que incluyen el dominio mediante las cuales se realiza el proceso de clasificación. En el presente trabajo se presenta modificaciones que se realizaron al método original, que buscan mejorar la eficacia de la simbolización y la precisión de la clasificación.

1.1 Formulación del problema

Existe un método en el que se simboliza las series temporales numéricas y se clasifica estas series temporales simbolizadas que representan potenciales evocados auditivos de tronco cerebral y se observó que se puede modificar los

algoritmos para la transformación de series numéricas a simbólicas y el de clasificación de las mismas con el fin de mejorarlo y de ser posible obtener mejores resultados que los obtenidos en el método original.

1.2 Objetivos de la Investigación

1.2.1 Objetivo General

Revisar y modificar el algoritmo de transformación de una serie temporal numérica a una secuencia simbólica; y el algoritmo usado para clasificación del método *Symbolic-based Patern Classification*, SPC.

1.2.2 Objetivo Especifico

- Realizar una revisión de literatura para determinar el estado del arte relativo a la clasificación de PEATC mediante series temporales simbólicas.
- Proponer nuevos algoritmos de transformación simbólica y clasificación de las series temporales de PEATC basados en la concepción central de los existentes en el método SPC.
- Comparar el rendimiento de los algoritmos propuestos, con respecto a los algoritmos originales.

1.3 Hipótesis

El uso de nuevos avances teóricos en *machine learning* y *data mining*; el aprovechamiento de técnicas algorítmicas más modernas; y, el uso de lenguajes de

programación especializados, aportarán al incremento de la eficiencia y precisión del método basado en símbolos para la clasificación de series temporales simbólicas.

1.4 Esquema de la tesis

El primer capítulo describe una idea global de la presente investigación; en el capítulo 2, se detallan los pasos para la revisión de literatura sobre series temporales simbólicas y sus métodos de clasificación, usando minería de datos para el dominio de los PEATCs. En el capítulo 3, se detalla el método utilizado en la presente propuesta. El capítulo 4, presenta la experimentación realizada para el mejoramiento del método SPC y los resultados obtenidos. El capítulo 5, despliegan las conclusiones del trabajo realizado y, finalmente, el capítulo 6 contiene las referencias bibliográficas.

CAPITULO 2

2 Revisión de literatura y Marco teórico

Este capítulo aborda una revisión de trabajos relacionados, según la metodología de Bárbara Kitchenham. La revisión de la literatura permitió conocer que no hay muchos trabajos relacionados con la simbolización considerando la forma de las series temporales y en cuanto a la clasificación uno de los métodos que expone mayor interpretabilidad es la utilización de los árboles de decisión, sin embargo, cuando se incorpora el dominio los autores prefieren incorporar sus propios clasificadores.

Adicionalmente, se presentan los conceptos fundamentales para el presente estudio, por ejemplo, qué es una serie temporal simbólica, se detalla algunas técnicas de simbolización, algunos métodos de clasificación supervisada y las métricas de clasificación más usadas, entre otros temas.

2.1 Revisión de literatura

2.1.1 Introducción

Las series temporales son grandes conjuntos de datos ordenados cronológicamente de una variable observable en diferentes momentos (Mauricio, 2009), éstas se presentan en muchos dominios tales como la medicina, agricultura, la industria entre otros. Por su alta dimensionalidad son particularmente estudiadas y al tratarlos con técnicas de minería de datos se ha logrado resultados muy alentadores.

Una de estas técnicas es la transformación de series temporales numéricas a secuencias temporales simbólicas, se refiere al proceso mediante el cual se realizan agregaciones temporales a las series y estas se asocian a un alfabeto finito, este procedimiento permite que haya una reducción de dimensionalidad y, en consecuencia, también de costo computacional.

Hay técnicas que realizan la segmentación de las series como PAA (*Piecewise Aggregate Approximation)* (Eamonn Keogh, Chakrabarti, Pazzani, & Mehrotra, 2001) con lo que se reduce la serie significativamente con una perdida razonable de información lo cual no afecta en algunos dominios. Otras discretizan la serie con lo que facilita el manejo de las mismas para encontrar información interesante.

Dentro de la revisión de literatura realizada se encontró que hay transformaciones simbólicas que incorporan el dominio, lo cual hace que la clasificación sea interpretable y ayuda, de mejor manera, a la toma de decisiones. Con la utilización de esta técnica se ha aprovechado más eficientemente la información contenida en las series temporales.

Para la clasificación de las series temporales simbólicas hay una gran variedad de métodos utilizados: unos permiten más interpretabilidad de los datos que otros; algunos estudios solo se han centrado en mejorar la precisión de la clasificación, mientras que otros como (Zalewski, Silva, Maletzke, & Ferrero, 2016), (Molina et al., 2016), (Sharabiani, Sharabiani, & Darabi, 2016) a más de la precisión revisan la interpretabilidad de los datos.

2.1.2 Metodología

Para la presente revisión de literatura referente a la transformación de series temporales numéricas a secuencias temporales simbólicas para diagnóstico y su clasificación, se utilizó la metodología de (Kitchenham, 2004). La selección y extracción de información se detalla a continuación:

- Preguntas de investigación
- Método de revisión
- Estudios incluidos y excluidos

2.1.2.1 Preguntas de Investigación

Las preguntas de investigación son:

RQ1: ¿Qué algoritmos de transformación simbólica de series o secuencias temporales existen para diagnóstico, que usen métodos de Data Mining?

RQ2: ¿Qué algoritmos de clasificación para series temporales o secuencias simbólicas existen para diagnóstico, que usen métodos de Data Mining?

RQ3: ¿Cuáles son las ventajas y desventajas de los algoritmos encontrados?

2.1.2.2 Método de revisión

2.1.2.2.1 Fuentes de investigación

Las siguientes bases de datos fueron utilizadas para extraer la información:

- Springer,
- Science direct,

- IEEE Xplorer,
- ACM,
- ArXiv,
- SCOPUS,
- PubMed

2.1.2.2.2 Cadenas de búsqueda

La cadena de búsqueda utilizada fue:

(Temporal abstraction OR Symbolic transformation) AND (time-series OR time-sequences) AND (diagnost*) AND (Data mining OR medical intelligen* OR Classificat*) AND (Method* OR Algorithm* OR technique*)

2.1.2.2.3 Criterios de selección de estudios primarios

Criterios de inclusión:

- Que se proponga algoritmos o métodos de conversión de series temporales numéricas a simbólicas.
- Que este publicado en una revista o congreso científico.
- Que trate sobre los procesos involucrados en el diagnóstico.
- Que trate sobre series o secuencias temporales simbólicas y diagnóstico.
- Que trate sobre clasificación de series o secuencias temporales simbólicas.
- Solo artículos científicos que estén en idioma inglés.
- Que trate de análisis de series temporales con métodos de data mining.

Criterios de exclusión:

- Que sean series temporales multivariante.
- Si el trabajo fue publicado antes del año 2000.
- Cuando el trabajo no sea accesible, duplicado o incompleto.
- Que traten de transformaciones de domino(tiempo-frecuencia) (descomposición de señales).
- Que trate de clasificadores estadísticos de series temporales.
- Revisar el título y el resumen para descartar documentos que no respondan a las preguntas de investigación.
- Revisar información relevante en la introducción o conclusión.

2.1.2.2.4 Extracción de información

Para cada artículo seleccionado se considera al menos uno de los siguientes componentes:

- Métodos o técnicas de transformación simbólica en dominios referentes a diagnóstico.
- Métodos eficientes para clasificación de series temporales simbólicas.
- Resultados.
- Conclusiones relevantes.

2.1.2.3 Estudios incluidos y excluidos

Al colocar la cadena de búsqueda en las diferentes bases de datos dio como resultado 12852 artículos relacionados con el tema de investigación, posteriormente

se incluyó los criterios de inclusión y exclusión antes mencionados y se redujo significativamente a 165 artículos a los cuales se realizó otro proceso de selección dando como resultado los datos que se muestran en la tabla 2.1 que se muestra a continuación:

Tabla 2-1: Cuadro de artículos incluidos y excluidos

	Springer	Science direct	IEEE Xplorer	ACM	ArXiv	Scopus	PubMed	TOTALES
ARTICULOS PARA REVISAR	54	62	17	5	4	11	12	165
REPETIDOS Y NO RELEVANTES	32	46	12	2	2	9	8	111
INCOMPLETOS	20	8	1	1	1	0	3	34
RELEVANTES	2	8	4	2	1	2	1	20

2.1.3 Resultados

De los estudios seleccionados se obtuvo información relevante para contestar las preguntas de investigación, las cuales se detallan a continuación:

RQ1: ¿Qué algoritmos de transformación simbólica de series o secuencias temporales existen para diagnóstico, que usen métodos de Data Mining?

Los resultados obtenidos en la revisión de literatura sobre los algoritmos para transformación de series o secuencias temporales simbólicas para series temporales univariantes; se detallan a continuación: en la mayoría de los artículos seleccionados se realiza la simbolización mediante la técnica SAX (*Symbolic Aggregate approXimation*) (Lin, Keogh, Lonardi, & Chiu, 2003) en la que la serie temporal se la divide en segmentos equidistantes utilizando PAA (*Piecewise*

Aggregate Approximation) y se saca los valores medios de cada segmento, utiliza la distribución gaussiana para dividir la serie en regiones estas pueden ser: 3, 5, 7, etc., regiones cada región corresponde a un símbolo la unión de estos forman una sola palabra (Lin et al., 2003); los artículos que presentan la utilización de esta técnica son: (Wang et al., 2016), (D. Li, Li, Bissyandé, Klein, & Traon, 2016), (Duan et al., 2016), (Sharabiani et al., 2016), (J. Yin, Xu, & Zheng, 2019), (Georgoulas, Karvelis, Loutas, & Stylios, 2015a), (Ordoñez, Schwarz, Figueroa-Jiménez, Garcia-Lebron, & Roche-Lima, 2016), también se ha tratado de mejorar esta técnica incorporando otros parámetros, por ejemplo, en (Taktak, Triki, & Kamoun, 2017) se propone un alfabeto ampliado, con la finalidad de incorporar la tendencia de la subsecuencia común más larga de tal manera que las letras minúsculas representan la caracterización media y las mayúsculas la tendencia de la subsecuencia común más larga de la siguiente manera (aA, bB, cC,…); en (Tamura & Ichimura, 2017) se propone una representación simbólica hibrida denominada MHSAX, es una combinación de SAX y el histograma de convergencia media móvil (MACD) con lo que se logra capturar la variación local y global de las series temporales. En (Yu, Zhu, Wan, Liu, & Zhao, 2019) se presenta la aproximación agregada simbólica de la característica de tendencia (TFSAX). Una vez segmentada la serie utilizando la técnica PAA, se extrae la característica de tendencia en cada segmento posteriormente se construye reglas de mapeo simbólico para discretizar la tendencia en símbolos. En (Yahyaoui & Al-Daihani, 2019) se propone una forma de representación simbólica que captura las tendencias de una serie temporal basándose en los puntos de cambio brusco y en la variación de los datos, la cual

los autores denominaron SAX_CP, en esta representación el tamaño de los segmentos varía de acuerdo con los puntos de cambio de la serie temporal.

Otra de las técnicas utilizadas son los shapelets son pequeños segmentos de formas de las series temporales, más representativos los cuales mediante una ventana deslizante recorren las series temporales viendo la coincidencia de la forma del shapelet con la de las series temporales cuando coinciden se mide la distancia entre el shapelet y la serie con lo que se puede clasificarlas (Ye & Keogh, 2009). En (Ahmadi, Aminshahidy, & Shahrabi, 2017) utiliza *fast shapelet* que es una representación simbólica hibrida ya que utiliza SAX en una primera etapa y posteriormente escoge los shapelets candidatos midiendo la calidad de los mismos utilizando el historial de colisiones. En (Zalewski et al., 2016) presenta tres enfoques exhaustivo, relajado y reducido de estos tres se toma el último para hacer la experimentación, se combina reducido shapelet (ST+) con filtro basado en consistencia(CS), selección de características basado en correlación (CFS), filtro basado en correlación rápida (FCBF). En (Guiling Li, Yan, & Wu, 2019) propone la utilización de shapelets de poda, con puntos clave, a la que se la denomina como (PSKP) por sus siglas en inglés, primero encuentra los puntos clave en las series temporales de acuerdo con la desviación estándar de cada tic de tiempo de las series temporales y luego extrae los candidatos a shapelet, con estos puntos clave; de esta manera se eliminan los shapelets redundantes.

Otra forma de realizar la simbolización de las series temporales que se propone en la literatura científica es tomando en cuenta la forma de la serie temporal, así se tiene que (Sevcech & Bielikova, 2015) propone una representación simbólica denominada ISC (*Incremental Subsequences Clustering*) que utiliza grupos de

subsecuencias similares como símbolos, además, utiliza un algoritmo incremental y codicioso para formar los grupos y esto lo hace aplicable en el procesamiento de datos de flujo. Esta agrupación de subsecuencias disminuye la distancia media mínima del centro del grupo, pero es causada por la reducción del espacio y no por la aleatoriedad de las secuencias formadas, ya que se forman a partir de las formas básicas de la serie temporal original. En (Alonso, Martínez, Pérez, Santamaría, & Valente, 2006) se propone simbolizar las series temporales conservando la morfología de las mismas para lo cual se define los siguientes símbolos: ascendente, descendente, pico, hundimiento, curvatura, transición e incorpora el dominio en los símbolos. En (Molina et al., 2016) se propone la simbolización de las series temporales en dos fases, la primera fase es independiente del dominio y utiliza los símbolos: pico, valle, subida, bajada, constante; en la segunda fase incorpora el dominio, en el cual cada símbolo corresponde a un carácter y sólo se mantienen los símbolos, de la fase anterior, que sean relevantes para el dominio, es decir, todos los símbolos irrelevante son eliminados, con lo cual se reduce significativamente la secuencia a ser analizada.

RQ2: ¿Qué algoritmos de clasificación para series temporales o secuencias simbólicas existen para diagnóstico, que usen métodos de Data Mining?

En la revisión de literatura realizada se encontró los métodos para clasificación de series temporales detallados en la tabla 2.2

Tabla 2-2: Métodos de clasificación

Artículo	Referencia	Método
A1	(Molina et al., 2016)	SPC (*Symbolic Pattern Clasification*)
A2	(Wang et al., 2016)	SVM (*Support Vector Machine*)
A3	(D. Li et al., 2016)	DSCo (*Domain Series Corpus)*
A4	(Yu et al., 2019)	kNN *(Nearest Neighbor*)
A5	(Sevcech & Bielikova, 2015)	---
A6	(Duan et al., 2016)	SVM (*Support Vector Machine*)
A7	(Sharabiani et al., 2016)	BR (*Bayesian Rules) and Chain Rule*
A8	(Taktak et al., 2017)	SVM (*Support Vector Machine)*
A9	(Tamura & Ichimura, 2017)	kNN (*Nearest Neighbor*)
A10	(Shknevsky, Shahar, & Moskovitch, 2017)	---
A11	(Yahyaoui & Al-Daihani, 2019)	kNN (*Nearest Neighbor*)
A12	(Guiling Li et al., 2019)	DT (*Decision Trees*)
A13	(Zalewski et al., 2016)	DT (*Decision Trees*)
A14	(J. Yin et al., 2019)	---
A15	(Georgoulas et al., 2015a)	kNN (*Nearest Neighbor*)
A16	(Lin et al., 2003)	kNN (*Nearest Neighbor*) y DT (*Decision Trees*)
A17	(Ye & Keogh, 2009)	DT (*Decision Trees*)
A18	(Ordoñez et al., 2016)	kNN (*Nearest Neighbor*)
A19	(Ahmadi et al., 2017)	RF (*random forest*), LR *(logistic regression*), SVM

Artículo	Referencia	Método
		(*support vector machine*) y PNN (*probabilistic neural network*)
A20	(Alonso et al., 2006)	GP (*Genetic Programming*)

De los artículos seleccionados el método de clasificación más utilizado es el de los vecinos más cercanos kNN, por sus siglas en inglés, según lo reportado en la literatura este método es el que mejor precisión se ha obtenido tomando en cuenta solo los datos, pero es poco interpretable, por lo que no es recomendable para dominios como el diagnóstico médico entre otros.

Los árboles de decisión o métodos de clasificación que incorporen reglas de decisión de algún tipo, son buenos candidatos para métodos de clasificación que incorporan el dominio.

RQ3: ¿Cuáles son las ventajas y desventajas de los algoritmos encontrados?

Los resultados obtenidos sobre las ventajas y desventajas de los algoritmos de transformación simbólica se detallan en la tabla 2.3 y de los algoritmos de clasificación en la tabla 2.4 respectivamente.

Tabla 2-3: Ventajas y desventajas de los algoritmos de transformación simbólica

Transformación Simbólica	Ventajas	Desventajas
SAX	Permite una considerable reducción de la dimensionalidad	No incorpora el dominio.

	manteniendo la calidad de los resultados. Permite que la extracción de los datos sea rápida.	No considera la tendencia o dirección de la serie.
FTSAX, SAX_CP, SAX AVANZADO, MHSAX	Tienen las ventajas del SAX original y además incorporan la tendencia de la serie temporal.	No incorporan el dominio
Shapelets	Permiten que la clasificación sea interpretable, con mayor precisión y rapidez.	Si no hay una adecuada definición de los parámetros se podría omitir shapelets representativos Requiere mucho tiempo de ejecución
Fast Shapelet, shapelets + puntos clave	Mejor precisión. Menoran el tiempo de ejecución.	No son interpretables
Transformaciones que toman en cuenta la forma de las series temporales	Son interpretables. Incorporan el dominio	Requieren de mayor estudio del dominio.

Tabla 2-4: Ventajas y Desventajas de los clasificadores

CLASIFICADOR	VENTAJAS	DESVENTAJAS
k-NN	• Es simple de implementar. • No tiene supuestos paramétricos. • Funcionan muy bien cuando el conjunto de entrenamiento contiene múltiples combinaciones de los predictores	• El proceso de encontrar los vecinos más cercanos en un conjunto de entrenamiento grande puede tener un alto costo computacional en cuanto al tiempo de procesamiento. • Es un aprendiz perezoso. • Es un algoritmo • Pierde precisión cuando la cantidad de datos es grande. • Es poco interpretable

CLASIFICADOR	VENTAJAS	DESVENTAJAS
Arboles de decisión	• Una vez construidos son computacionalmente de bajo costo incluso con grandes cantidades de datos. • Son interpretables. • Resistentes a valores atípicos y capaces de manejar valores perdidos. • El pre procesamiento de los datos es más fácil porque no se tiene que escalarlos. • No es necesario transformar los datos • No son lineales ni paramétricos.	• Es propenso al sobreajuste. • Si el conjunto de datos tiene incoherencias el modelo se puede acoplar a ellas dando como resultado clasificaciones deficientes. • Son sensibles a los cambios en los datos. • Puede perder relación entre predictores.
Support Vector Machine (SVM)	• Tiene varios kernels en las que se divide el hiperplano para separar las clases clasificadas. • Es eficiente con datos de alta dimensión. • Precisión media a alta	• Requiere definir varios parámetros para que la precisión sea óptima. • Baja capacidad de interpretación.
Reglas bayesianas	• Cada regla puede representar una parte del conocimiento del dominio	• Se basa en una aproximación
PNN	• Alta precisión. • Aprovecha la clasificación optima de Bayes.	• Requiere más espacio en memoria para almacenar el modelo.
SPC	Es interpretable	Sirve para pequeños conjuntos de datos

2.2 Marco teórico

2.2.1 Series temporales

Según (Mauricio, 2009) define las series temporales como un conjuntos de datos numéricos ordenados y equidistantes cronológicamente sobre una característica (serie univariante o escalar) o sobre varias características (serie multivariante o vectorial) de una unidad observable en diferentes momentos.

A lo largo del tiempo las series temporales evolucionan y muchos investigadores están interesados en conocer su evolución a través del tiempo. Para conseguir este fin se puede utilizar la función de auto correlación muestral y su inferencia se la conoce como dominio del tiempo y otros casos a través de la frecuencia mediante la utilización de la densidad espectral y su inferencia es conocida como dominio de la frecuencia (Gras, 2001). Para la presente tesis se toma en cuenta el tiempo y la forma de las series temporales.

La Figura 2.1 muestra un ejemplo de serie temporal en donde muestra la variación del precio de un activo financiero.

Figura 2-1: Ejemplo de Serie temporal que representa el precio de un activo financiero ("Tipos de series temporales - Qué es, definición y concepto | 2021 | Economipedia," n.d.).

Uno de los objetivos del análisis de las series temporales es predecir o pronosticar el comportamiento de una variable a través del tiempo en base al conocimiento que se posea de esta de las observaciones que se hayan realizado; suponiendo que no varía estructuralmente, mediante el análisis de modelos realizados previamente. Un segundo objetivo es la descripción, para lo cual se debe graficar la serie para visualizar las descripciones básicas de la misma considerando la tendencia, si los datos presentan forma creciente; la estacionalidad cuando existe influencia de ciertos periodos de cualquier unidad de tiempo; si aparecen observaciones extrañas o discordantes valores atípicos. Tercer objetivo es la simulación se utiliza cuando el proceso es muy complicado y resulta complejo su estudio de forma analítica. Otro objetivo es el control de procesos: en este se trata de seguir la evolución de una

variable definida para lograr regular su resultado, es utilizado en medicina en centros de control de enfermedades (Ecured contributors, 2019).

Una serie temporal puede ser continua cuando los valores de la variable que se está considerando para el análisis fluye de forma permanente en el tiempo, o puede ser discreta cuando los valores de las observaciones se presentan por intervalos de tiempo generalmente homogéneos, en resumen, depende de cómo sean las observaciones(Mauricio, 2009).

Puede ser determinista, si los resultados se los predice de forma exacta, ya que pueden seguir un determinado patrón o esquema; y si no siguen ningún patrón especifico pueden ser aleatorias o también llamadas estocásticas. Con este último tipo de series los resultados se los puede predecir parcialmente, puesto que dependen de las observaciones pasadas, es decir, tienen un componente de probabilidad (García, 2016) (Mauricio, 2009).

Hay cuatro aspectos o componentes que deben ser tomados en cuenta en toda serie temporal, debido a que se considera que su comportamiento es producto de ellos (Mauricio, 2009). A continuación, se hace una breve descripción de cada uno:

- *Tendencia*: representa la evolución de la serie en periodos prolongados de tiempo, esta se identifica con el movimiento o dirección de la serie a largo plazo.
- *Estacionalidad*: son variaciones periódicas en donde se colectan comportamientos de la serie de forma regular y repetitiva, las mismas que pueden ser producidas por varias causas como factores climáticos; estas

pueden ser medidas y, de ser el caso que no sean necesarias, también podrían ser eliminadas.

- *Variación* cíclica: son variaciones periódicas en lapsos de tiempo grandes no necesariamente regulares, para poder encontrarlas se requiere que sean series temporales largas.
- *Aleatoriedad*: se refiere a movimientos irregulares de las series, ocurridos al azar, pero recuperables posteriormente. Cuando se considera solo el componente aleatorio se puede calcular el error de estimación.

2.2.2 Minería de datos en series temporales

La minería de datos en series temporales es la aplicación de la minería de datos a un conjunto de datos extenso ordenado cronológicamente generalmente equidistante que ocurre en un intervalo de tiempo.

Las técnicas de minería de datos en series temporales son diversas entre las cuales se tiene: el agrupamiento o *clustering*, descubrimiento de reglas de asociación, descubrimiento de patrones, agregaciones simbólicas, entre otras.

El análisis de las series temporales es complejo ya que son conjuntos de datos con dimensionalidad alta y de formas variadas. Para tratar la gran dimensionalidad de las series se han realizado procesos de segmentación, así como también de simbolización con buenos resultados. La similitud entre las series es otro aspecto importante que se debe considerar en su análisis se lo realiza mediante el cálculo de la distancia entre ellas.

Los campos de aplicación de series temporales son muy variados como en la industria, agricultura, medicina, finanzas, donde es posible que se utilicen técnicas estadísticas o de minería de datos, para encontrar valor agregado y con ello utilizar de la mejor manera el conocimiento encontrado para dar las soluciones más adecuadas a los problemas que se puedan presentar en los diferentes dominios. Con el análisis de las series temporales utilizando técnicas de minería de datos se puede explicar, pronosticar o diagnosticar varios sucesos o acontecimientos del diario vivir como, por ejemplo: el pronóstico de cosechas, el diagnostico en el ámbito de la medicina y de la industria.

2.2.3 Series temporales simbólicas

También llamadas secuencias temporales simbólicas o abstracciones temporales son series temporales numéricas transformadas a secuencias de caracteres asociadas a un alfabeto finito; este tipo de series temporales representan un aporte sustancial en la reducción de la dimensionalidad que es uno de los desafíos a los que se enfrentan las personas que trabajan con series temporales puesto que, en muchos dominios, debido a la gran cantidad de datos, se dificulta su análisis. Al aplicar data Mining, mediante un proceso de transformación de las series temporales, se ataca el problema de la dimensionalidad. En la transformación simbólica se debe considerar que no se pierdan las características esenciales, ni la calidad de la información de la serie temporal original.

Para el proceso de transformación simbólica se debe tomar en cuenta tres aspectos (Boucheham, 2012):

1. El tamaño de la serie aproximada debe ser menor que el de la serie original
2. Los extremos de las series coinciden.
3. La distancia entre la serie original y la abstracción es menor que un umbral.

La representación simbólica de series temporales ha demostrado ser útil para facilitar la manipulación de datos discretos en muchas áreas, por ejemplo, clasificación de imágenes médicas de rayos X (Rajaei, Dallalzadeh, & Rangarajan, 2015), diagnóstico de fallas de maquinaria (Duan et al., 2016), verificar la consistencia en los datos médicos de ECG (D. Li et al., 2016) y, detección movimiento con censores en el sector ganadero (H. Yin, Yang, Zhu, Ma, & Zhang, 2015).

Existen varias técnicas de transformaciones simbólicas que han sido estudiadas y propuestas en la literatura, a continuación, se menciona algunas de ellas:

- ***PAA (Piecewise Aggregate Approximation)*** esta técnica fue propuesta casi simultáneamente por dos equipos de investigación de forma independiente , una fue propuesta por (Eamonn Keogh et al., 2001) la denomino *Piecewise Aggregate Approximation* PAA y la otra fue propuesta por (Yi & Faloutsost, 2000) la denomino *Segmented-means*, en la literatura científica es más referida como PAA.

  En esta técnica se divide la serie en un conjunto finito de segmentos de igual tamaño y almacena las medias de los valores correspondientes a los puntos que corresponden a cada segmento (ver figura 2.2). A pesar de su sencillez, logra muy buenos resultados en términos de exactitud y de eficiencia.

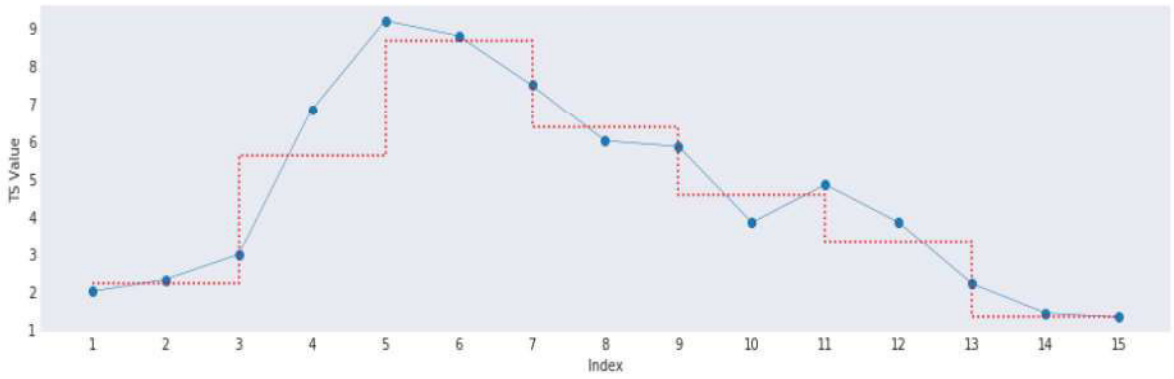


Figura 2-2: Representación PAA de una serie temporal (Krish, 2018)

- ***APCA (Adaptive Piecewise Constant Approximation)*** Esta técnica fue presentada en (E. Keogh, Chakrabarti, Mehrotra, & Pazzani, 2001), a diferencia de la técnica anterior ésta permite segmentos de diferente tamaño, para esto se toma en cuenta donde hay más uniformidad en la serie se toma como un segmento, es decir se toma todos los puntos en donde hay menor actividad como un segmento, así como también todos los puntos donde hay mayor actividad como otro segmento (ver figura 2.3).

  Según el estudio realizado por (E. Keogh et al., 2001) la reducción de dimensionalidad es la mitad de la obtenida con PAA.

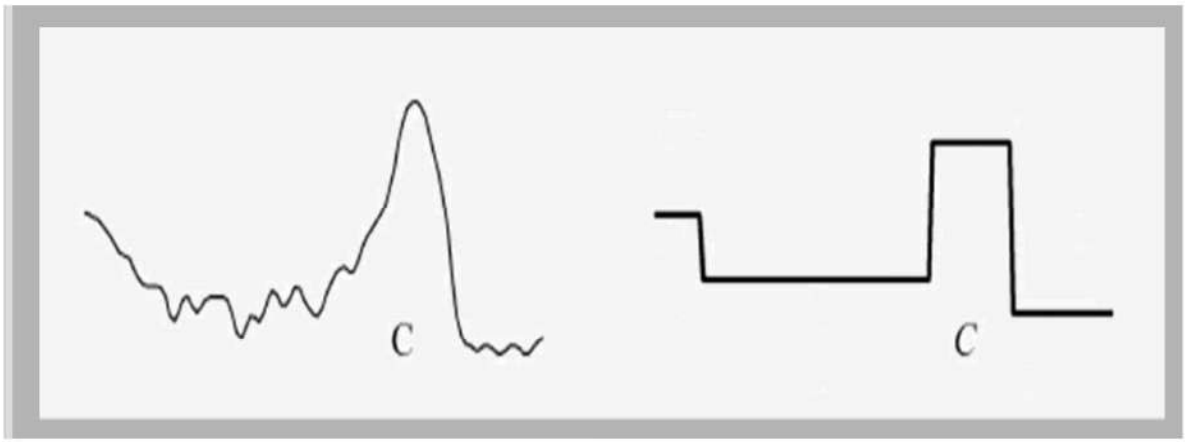


Figura 2-3: Representación de una serie temporal con APCA

- ***SHAPELETS*** esta técnica fue propuesta por (Ye & Keogh, 2009) para series temporales es una sub secuencia de una serie de tiempo que se alcanza a través de una búsqueda exhaustiva de cada sub secuencia posible entre valores predefinidos para longitudes mínima y máxima (ver figura 2.4). Para la selección del shapelet se genera candidatos, se realiza el cálculo de la distancia entre shapelets y posteriormente se hace la evaluación de los mismos quedando los más representativos. Los shapelets son apropiados para series temporales de gran tamaño. Se puede ver como una forma de algoritmo supervisado de descubrimiento de motivos, utiliza una ventana deslizante para encontrar los shapelets.

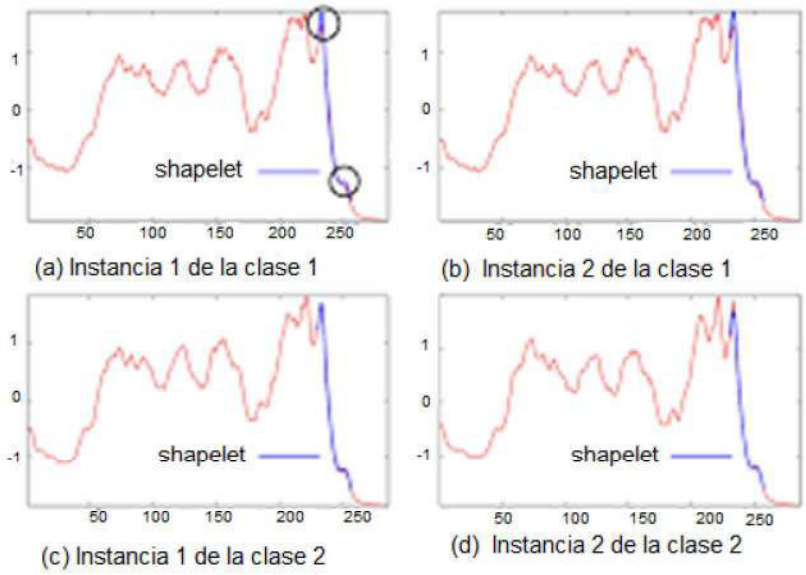


Figura 2-4: Representación por Shapelet (Guozhong Li et al., 2020).

- ***SAX (Symbolic Aggregate approXimation)*** Esta técnica fue publicada por (Lin et al., 2003) desde su publicación ha sido ampliamente utilizada, así como también se han propuesto algunas variantes de la misma, en el contexto de la literatura científica. Esta técnica transforma una serie temporal en una secuencia de símbolos que forman una sola palabra con una longitud definida. Para lograr este fin se aplica PAA a la serie temporal, posteriormente transformarla a símbolos discretos con puntos de ruptura determinados y una cantidad finita de símbolos (ver figura 2.5).

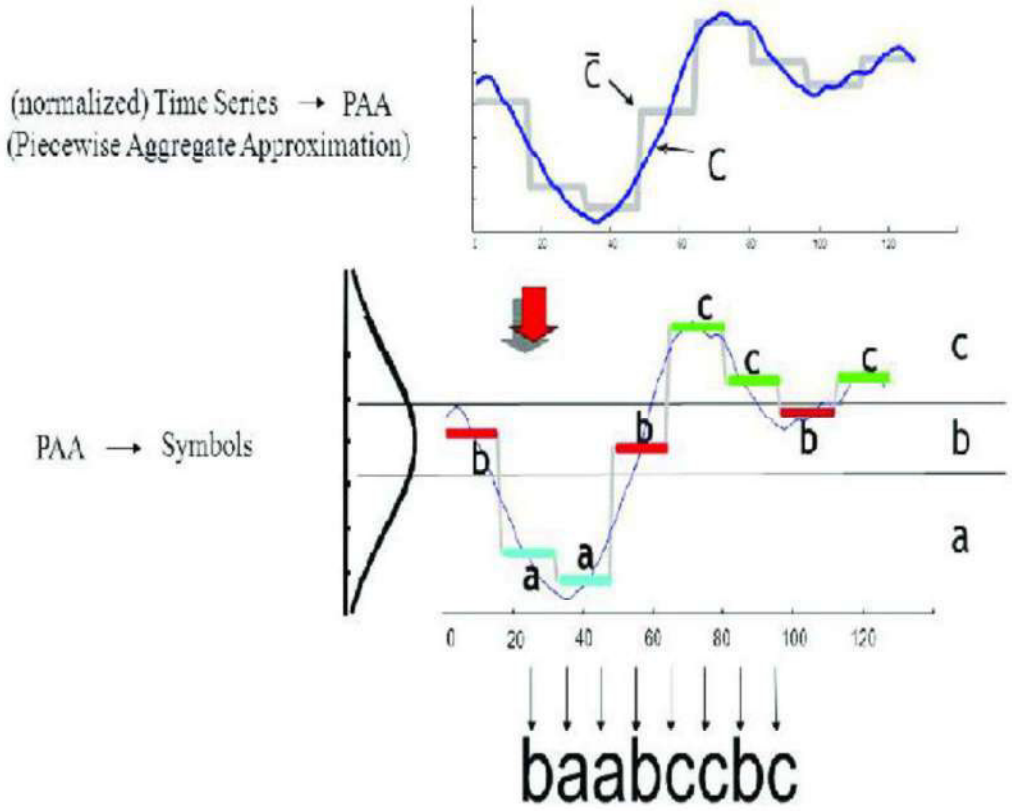


Figura 2-5: Proceso de transformación simbólica de una serie temporal usando SAX (Caviedes, Li, & Jammula, 2020).

SAX ha sido utilizado ampliamente en la literatura en varios trabajos de investigación, desde su aparición en el 2003, tales como (N. Kumar et al., 2005), (Lin & Li, 2009), (Ordóñez, DesJardins, Feltes, Lehmann, & Fackler, 2008), (Pham, Le, & Dang, 2010), (Senin & Malinchik, 2013), (Georgoulas, Karvelis, Loutas, & Stylios, 2015b), (D. Li et al., 2016), (J. Yin et al., 2019) entre otros, así como también se han creado variaciones del mismo como en (Lkhagva, Suzuki, & Kawagoe, 2006) SAX extendido, (Pham et al., 2010) SAX adaptativo indexable, (Tamura & Ichimura, 2017) MHSAX y (Yu et al., 2019).

- ***SDL (Shape Definition Language)*** Esta técnica fue publicada por (Agrawal, Psaila, Wimmers, & Zaït, 1995) y está ampliamente descrita en (Santamaría Falcón, 2011). Esta técnica convierte elementos numéricos en sus símbolos correspondientes y compara las secuencias de símbolos sin tener en cuenta los valores numéricos de los elementos originales; en consecuencia, dos valores muy próximos se podrían considerar como símbolos diferentes (Kim, Yoon, Park, & Won, 2006).

- ***STF-Mine*** Esta técnica fue propuesta por (Batal, Sacchi, Bellazzi, & Hauskrecht, 2009). En este algoritmo se propone que la transformación de los símbolos se realice en dos partes: abstracciones por valor (Alto, Normal, Bajo) y por tendencia (Creciente, decreciente, estable) y el alfabeto de símbolos se obtiene de la combinación de las abstracciones obtenidas de acuerdo a la forma de la serie (ver figura 2.6).

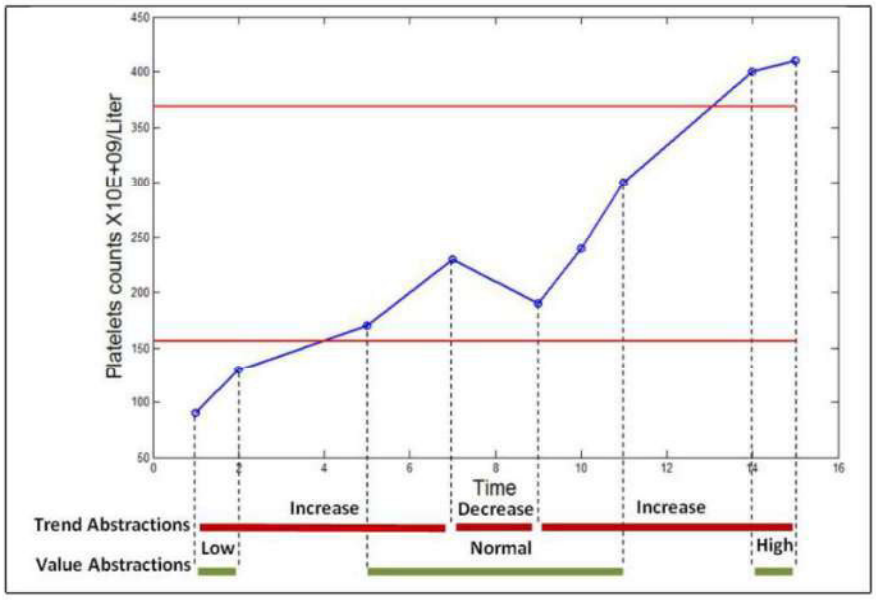


Figura 2-6 : Representación de STF-Mine (Batal et al., 2009).

- En (Santamaría Falcón, 2011) se propone transformar la serie temporal por una secuencia de caracteres en donde cada carácter corresponde a un símbolo que incorpora una parte de la semántica de la serie que tiene significado para el experto del dominio. Tiene como alfabeto: subida, bajada, hundimiento, pico, transición y curvatura. También presenta una propuesta para establecer una distancia simbólica entre las series transformadas que se deriva de la distancia de edición, la cual se detalla más adelante.

- En (M. Kumar & Kalia, 2012) se propone una abstracción simbólica fácil de interpretar y ayuda en la búsqueda de un patrón de conjunto. Utiliza tres símbolos: Up, Down y Neutral.

- En (Shabtai, 2016) se propone una abstracción simbólica considerando la duración por intervalos para la detección de anomalías. Utiliza como símbolos: muy corto, corto, medio, largo y muy largo.

- En (Ma, Yan, Li, & Nord, 2018) se propone una abstracción simbólica basada en formas y la agrupación jerárquica. Se utilizaron otras técnicas como dendrograma, mapa de calor y vista de calendario para ayudar a comprender los comportamientos de uso de energía del edificio. Se utiliza los símbolos: *stable, jump, up, down, plunge*.

2.2.4 Interpolación y suavizado de series temporales

2.2.4.1 Interpolación de Newton

Para N centros x_0 ,…, x_N, los polinomios de Newton se pueden construir con una secuencia recursiva como se indica a continuación:

$P_1(x) = a_0 + a_1 (x - x_0)$,

$P_2(x) = a_0 + a_1 (x - x_0) + a_2 (x - x_0)(x - x_1)$,

.
.
.

Por lo tanto,

$$P_N(x) = a_0 + a_1 (x - x_0) + a_2 (x - x_0)(x - x_1) + a_3 (x - x_0)(x - x_1)(x - x_2) + \text{---} +$$

$$a_N (x - x_0)(x - x_1)(x - x_2) \text{---} (x - x_{N-1})$$

donde el polinomio $P_N(x)$ se obtiene de $P_{N-1}(x)$ usando la relación recursiva

$$P_N(x) = P_{N-1}(x) + \text{---} + a_N (x - x_0)(x - x_1)(x - x_2) \text{---} (x - x_{N-1})$$

Contenido tomado de (Mathews , John H. and Fink, 2000)

2.2.4.2 Interpolación o suavizado de splines cúbicos

Los splines cúbicos constan de varios segmentos en donde cada uno de ellos son polinomios cúbicos unidos por puntos especiales llamados nudos. Para que la unión de estos segmentos sea una curva suave los polinomios y su primera y segunda derivada deben ser continuos en los nudos y tener un gradiente uniforme; lo que implica que la continuidad de la primera derivada de la curva no contiene esquinas

y para la continuidad de la segunda derivada el radio de la curvatura está dado por cada punto o nudo (Kimball, 1976)(Mathews , John H. and Fink, 2000).

(Mathews , John H. and Fink, 2000) define la interpolación de splines cúbicos como: Suponiendo que se tiene N + 1 puntos, como $\{x_h, y_h\}$ donde h = 0,1,2,…,N, cuyas abscisas están ordenadas ascendentemente. Se dice que una función S(x) es una interpolante cúbico segmentario para dichos datos si existen N polinomios cúbicos $S_h(x)$, definidos como:

$S(x) = S_k(x) = S_{h,0} + S_{h,1}(x - x_h) + S_{h,2}(x - x_h)^2 + S_{h,3}(x - x_h)^3$, para $x \in [x_h, x_{h+1}]$ y h = 0,1, 2, … , N-1 que verifican las siguientes propiedades:

$S(x_h) = y_h$, para h = 0, 1, 2, …, N, (interpola datos)

$S_h(x_{h+1}) = S_{h+1}(x_{h+1})$ para h = 0, 1, 2, …, N -2,

$S'_h(x_{h+1}) = S'_{h+1}(x_{h+1})$ para h = 0, 1, 2, …, N -2,

$S''_h(x_{h+1}) = S''_{h+1}(x_{h+1})$ para h = 0,1,2, …, N -2, (cumplen que la primera y segunda derivadas existen y son continuas)

2.2.5 Distancias entre series temporales para medir su similitud/disimilitud

Las series temporales son un tipo de dato complejo y como tal se la debe considerar como un todo ya que sus características van cambiando con el tiempo, una de las consideraciones importantes que se hace con series temporales es que la similitud se considera como una aproximación ya que es prácticamente imposible que exista

dos series temporales iguales. Hay distancias numéricas entre las más utilizadas están la distancia euclidiana y la distorsión dinámica del tiempo (*Dinamic time Warping* - DTW), así como también distancias entre caracteres como las distancias de edición se detallan brevemente a continuación.

2.2.5.1 Distancias numéricas

Distancia euclidiana entre las distancias aritméticas más utilizadas esta la distancia euclidiana, esta ha sido ampliamente utilizada en varios campos como en (Lee, Lim, Kim, Yang, & Lee, 2014), (San Segundo, Tsanas, & Gómez-Vilda, 2017), (Patel & Upadhyay, 2020); es la distancia en línea recta entre dos puntos en el espacio euclidiano. Se aplica generalmente en series temporales transformadas, cuando hay diferencias entre las series como por ejemplo que no tienen la misma longitud puede haber diferencias en la distancia, este inconveniente se supera normalizando los datos, se calcula la distancia euclidiana con la siguiente ecuación:

$$E = \sqrt{\sum_{p=1}^{n} (x_p - y_p)^2}$$

Ecuación 1

La distorsión dinámica del tiempo (*Dinamic time Warping* - DTW) es muy utilizada ya que a diferencia de la distancia euclidiana este si considera la distorsión del tiempo, es decir, se puede alinear datos desfasados en el tiempo. En 1994 D. Berndt, J. Clifford propusieron el siguiente algoritmo para DTW, el mismo que se describe en (Ge & Chen, 2020):

Dadas dos series temporales X =(x1, x2,x3,…,xn) y =(y1,y2,y3,….,ym). La distancia DTW (X,Y) se calcula de la siguiente manera:

$$DTW(X,Y) = \sqrt{D(n,m)}$$

Donde n y m son las longitudes de las series y D(n,m) se calcula mediante

$$\mathrm{D(i,j)} = (\mathrm{xi} - \mathrm{yj})2 + \min[\mathrm{D(i-1,j)}, \mathrm{D(i,j-1)}, \mathrm{D(i-1,j-1)}]$$

Con

$$\mathrm{D(0,0)} = 0, \mathrm{D(i,0)} = \mathrm{D(0,j)} = \infty, \qquad \mathrm{i} = 1,2,\ldots,\mathrm{n}; \quad \mathrm{j} = 1,2,\ldots,\mathrm{m}$$

DTW tiene un alto costo computacional especialmente en series temporales largas (Ge & Chen, 2020) (Corres, Esteban, García, & Zárate, 2009).

2.2.5.2 Distancia de edición – distancia simbólica

Para el cálculo de la distancia de edición, entre dos secuencias temporales, es necesario que se realice un conjunto de operaciones para convertir la primera secuencia en la segunda. Las operaciones permitidas son inserción, borrado y substitución de un carácter; a cada una de esas operaciones se le asigna un costo y, finalmente, los costos se suman para obtener la distancia de edición total (Amón & Jiménez, 2010; Cohen, Ravikumar, & Fienberg, 2003).

La **distancia de edición de Levenstein** (Levenshtein, 1966) es la más simple ya que asigna como costo de cada operación una unidad. La suma de las operaciones mínimas necesarias para transformar una secuencia en otra es el valor de la distancia.

La **distancia Damerau-Levenshtein** incorpora la trasposición de caracteres en las secuencias dentro de las operaciones de distancia (Amón & Jiménez, 2010).

La **distancia Needleman - Wunsch** (Needleman & Wunsch, 1970) introdujeron la variabilidad de los costos de las operaciones, es así que para inserción y borrado tiene un costo fijo y para sustitución y copia tiene otro costo que depende del dominio que se esté trabajando (Santamaría Falcón, 2011).

La **distancia de brecha afín** tiene como objetivo dar solución al fallo de la identificación de secuencias equivalentes, mediante la penalización de la inserción/borrado de k caracteres consecutivos con bajo costo, mediante una función afín p(k) = g+h(k-1), donde *g* es el costo de iniciar una brecha, *h* el costo de extenderla un carácter, y *h*<<*g* (Amón & Jiménez, 2010).

La **distancia Smith- Waterman** (Smith & Waterman, 1981) calcula la distancia tomando en cuenta los posibles alineamientos locales que puedan existir entre dos secuencias, en esta se define las misma operaciones de distancia de edición y además permite eliminar caracteres al principio o final de las secuencias, obteniendo secuencias equivalentes (Amón & Jiménez, 2010).

La **distancia de Hamming** solo toma en cuenta la coincidencia de caracteres en la misma posición, se utiliza con secuencias de misma longitud y tiene como única operación la sustitución con costo 1 (Alberca & Pensamient Matemátic, 2018) (Santamaría Falcón, 2011).

La **subcadena común más larga** ha sido analizada por (Apostolico & Guerra, 1987) así como también por (Friedman & Sideli, 1992) esta técnica se basa en un proceso

recursivo mediante el cual se encuentra y se elimina la subcadena común más larga entre dos cadenas comparadas, esta técnica es eficaz para manejar errores tipográficos menores, pequeñas variaciones y permutaciones de nombres.

La **distancia de bolsa** (Bartolini, Ciaccia, & Patella, 2002) es una aproximación barata a la distancia de edición que actúa como filtro para descartar cadenas no relevantes esta distancia es menor o igual a la distancia de edición.

La **distancia de edición con penalización real** propuesta (ERP) en (Chen & Ng, 2004) puede soportar el cambio de tiempo local, y es una métrica, se basa en la distancia de edición simple y *dynamic time wraping* (DTW), pone una penalización real cuando no hay gap en las dos secuencias, entendido como gap un símbolo añadido a la secuencia, y si algún carácter de las dos secuencias es gap, la penalización es una constante. Tiene como propiedad el poder indexarse a un árbol B+ estándar.

La **distancia de edición en secuencia real** (EDR) propuesta por (Chen, Özsu, & Oria, 2005), tiene las mismas propiedades que ERP pero esta además soporta ruido en los datos ya que tiene un umbral de coincidencia lo que permite la reducción de los efectos del ruido.

La **Distancia simbólica** propuesta por (Santamaría Falcón, 2011) define a la distancia como una variante de la distancia Needleman- Wunsch puesto que permite costos variables para las operaciones de distancia, y se le puede incorporar conocimiento del dominio ya que para las operaciones de sustitución y copiado el costo depende del dominio de aplicación; para definir los costos de las operaciones

se utiliza la estructura de grafo ya que los símbolos pueden ser compuestos (comportamiento + tipo); el costo principal de sustitución de dos símbolos está definido por el comportamiento y el tipo de ambos símbolos matiza dicho costo. El costo de inserción y borrado se unifican para asegurar que la comparación de las dos series es simétrica. Este proceso se realiza para cada subsecuencia, posteriormente se realiza el proceso de normalización que se basa en dividir la distancia obtenida de cada subsecuencia para la mayor distancia posible entre las subsecuencias; finalmente se realiza la media aritmética de las n distancias normalizadas.

2.2.6 Clasificación

La clasificación constituye una de las tareas básicas de análisis de datos y más utilizada en sistemas inteligentes, esta es supervisada ya que se tiene conocimiento previo de la clase o clases con sus respectivas etiquetas. Para la realización de la clasificación el conjunto de datos se divide en dos subconjuntos: el primero para el entrenamiento el cual debe tener un tamaño lo suficientemente grande, para que el resultado del entrenamiento del modelo de clasificación tenga una alta confiabilidad, y el segundo para prueba, el mismo que tiene como objetivo verificar el nivel de confiabilidad del modelo. Hay varios modelos de clasificación, algunos de ellos se describen brevemente a continuación.

2.2.6.1 Modelos de clasificación

k-NN (k –vecinos más cercanos)

Es un método supervisado de clasificación, no paramétrico, en el que se pretende encontrar los k registros similares, dentro del conjunto de entrenamiento, con el nuevo registro que se desea clasificar; luego se asigna ese registro a la clase a la que pertenecen los registros predominantes del conjunto de k registros más cercanos y que se denomina conjunto de predictores.

Este método no hace suposiciones de la relación de pertenencia entre las clases y los predictores, sino que se extrae las similitudes de los predictores en el conjunto de datos en base a lo cual estima el valor de probabilidad de que un registro x pertenezca a una clase C tomando en cuenta la información del conjunto de predictores. Un registro x es asignado a una clase C si esta clase es más predominante entre los k vecinos más cercanos del entrenamiento. Por ejemplo, si se tiene dos clases y se asigna k = 3 y dentro de los tres predictores están dos de la clase 1 y uno de la clase 2, el nuevo registro se le asigna la clase 1 como se muestra en la figura 2.7.

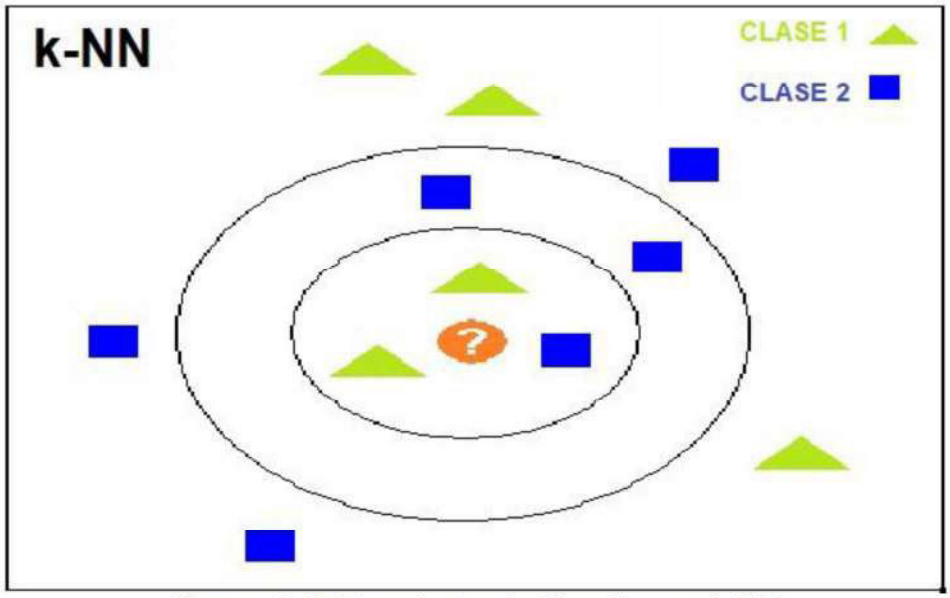


Figura 2-7: Ejemplo de clasificación con k-NN

El algoritmo de k-NN se basa en cálculos de distancia, la más utilizada es la distancia euclidiana, ya que es la de más bajo costo computacional, después de calcular las distancias entre el registro a clasificar y los registros existentes, se deben tener en cuenta los k vecinos que van a ser seleccionados, ya que cuando k tiene un valor grande reduce el efecto de ruido en la clasificación y por ende, el sobreajuste por ruido, si es muy bajo se puede adaptar al ruido pero si es demasiado alto se pierde la capacidad de capturar la estructura local de los datos, es decir, se crean límites entre clases parecidas. Por ende, la elección de un buen k depende fundamentalmente de la naturaleza de los datos, este puede ser seleccionado con procedimientos de optimización (Shmueli, Bruce, Gedeck, & Patel, 2020)(Parra Rodríguez, 2017).

Arboles de decisión

Este método de clasificación tiene su origen en el aprendizaje automático (*Machine Learning*) y la inteligencia artificial; crea diagramas de construcciones lógicas con las que se resuelven los problemas, también se le conoce como segmentación jerárquica. Este modelo es apropiado para conjuntos de datos grandes, es adaptable tanto para problemas de clasificación como de regresión, permite entender e interpretar fácilmente las decisiones tomadas por el modelo, ya que se basa en la separación de registros en subgrupos homogéneos mediante divisiones recursivas, el árbol de decisión se va armando con reglas lógicas, formando una estructura compuesta por nodos, ramas y hojas. La construcción del árbol se inicia por la creación del nodo raíz o inicial, que representa la variable

de mayor relevancia, se divide este nodo en dos o más subconjuntos homogéneos y mediante las ramas se conectan los nodos internos; este proceso se realiza recursivamente hasta que, finalmente, se llega a las hojas que representan los posibles resultados. Hay dos tipos de nodos: los nodos de decisión, son aquellos que se dividen y los nodos terminales u hojas (Parra Rodríguez, 2017).

Un árbol ajustado desembocará indudablemente en sobreajuste, lo cual puede ser evitado con procedimientos de poda en el árbol, que consiste en la eliminación de ramas que no aporten a la clasificación (Shmueli et al., 2020).

En la literatura, existen varios algoritmos para la construcción de árboles de decisión. A continuación, se hace una descripción de los detalles más relevantes de algunos de ellos:

CHAID (CHi-square Automatic Interaction Detection) Este algoritmo se basa en AID (*Automatic Interaction Detection)* fue denominado así porque no se concentra en la clasificación, sino más bien, en las interacciones entre las variables. El algoritmo AID realiza el análisis de la varianza entre las categorías de la variable independiente y las divisiones binarias de la variable dependiente; utiliza la prueba F para discernir las mayores diferencias posibles. Una de las limitaciones importantes de este algoritmo consiste en que la partición resultante depende de la primera variable que se escoja, pues esto condiciona al resto de particiones; otra limitación radica en que las particiones son exclusivamente binarias y, finalmente; el algoritmo tiende a seleccionar como variables

significativas a las que tienen mayor número de categorías sin importar si éstas son o no realmente relevantes para la clasificación.

CHAID surge como una mejora del método AID, ya que corrige muchas de estas limitaciones entre las limitaciones ya mencionadas. CHAID admite particiones o divisiones de más de dos nodos, se generan, tanto arboles de clasificación, como de regresión y utiliza un contraste estadístico para determinar la jerarquía del árbol conservando los valores distintos de las variables; si las variables son continuas utiliza la prueba F y si son categóricas, Chi-cuadrado; si el resultado es estadísticamente representativo, se realiza la división del nodo de lo contrario no se la realiza, con lo cual se disminuye el costo computacional (Shmueli et al., 2020) (Parra Rodríguez, 2017).

CART (Classification and Regression Trees) Este algoritmo fue propuesto en 1984 por Leo Breiman, Jerome Friedman, Richard Olshen y Charles Stone; con este algoritmo se construyen arboles de decisión binarios, es decir, con dos ramas; utiliza GINI índex para calcular la medida de impureza; cuando esta es cero, significa que el grupo es homogéneo (Parra Rodríguez, 2017) ("CART: Classification and Regression Trees," 2020).

QUEST (Quick, Unbiased, Efficient Statistical Tree) Este método fue propuesto en 1997 por Wei-Yin Loh y Yu-Shan Shih; consiste en escoger la variable más representativa para segmentar los datos y posteriormente realizar las divisiones para la construcción del árbol. Este método solo se lo puede utilizar si la variable

de salida es categórica nominal y permite particiones con más de dos nodos; se diferencia de los anteriores, en la manera con la cual se particionan los nodos.

Una variante de este método propuesto por los mismos autores es FACT, en la que se utiliza análisis discriminante para la construcción del árbol de decisión (Parra Rodríguez, 2017)("Métodos de selección dividida para árboles de clasificación en JSTOR," n.d.).

ID3 propuesto por (Quinlan, 1983) sirve para descubrir reglas de clasificación relevantes de una colección de objetos (datos), pertenecientes a dos clases, es decir, sirve para clasificaciones binarias, y construir un árbol de decisión conciso y correcto para la cual la selección de atributos útiles es imprescindible.

C5.0 y su versión no comercial C4.5 son los algoritmos de árboles de decisión para clasificación más utilizados, estos algoritmos resultan de la mejora del algoritmo ID3 perteneciente al mismo autor. Las variables de entrada para estos algoritmos pueden ser categóricas o continuas y su salida es solo categórica; la construcción del árbol de decisión cada atributo es evaluado con una prueba estadística que determina los elementos de entrenamiento, de estos selecciona el mejor atributo y se lo coloca como nodo raíz, los ejemplos de entrenamiento son colocados en los nodos descendientes siguientes de acuerdo al valor que tengan para el atributo que está siendo evaluado; el primer caso es con el nodo raíz. Para hacer este proceso se utiliza la ganancia de información para la evaluación del nodo más relevante a ser evaluado (Parra Rodríguez, 2017).

SVM (*Support Vector Machine*)

Es un método de aprendizaje supervisado que se utiliza tanto en clasificación como regresión. Es un método robusto que equilibra los datos de entrada en una dimensión más alta, que consta de un hiperplano, cuya función consiste en separar los datos, conocidos como vectores de soporte con un margen de distancia optimo entre las clases el cual se denomina margen funcional; las funciones o kernel utilizadas son: la función lineal, sigmoide, RGB función de base radial gaussiana, polinomio homogéneo y no homogéneo. Inicialmente solo se utilizaba la función lineal para separar las dos clases a las cuales se les asigna los valores de 1 y -1. En razón de que la naturaleza de los datos, estos no siempre son linealmente separables, fueron incorporadas las otras funciones anteriormente indicadas. La figura 2.8 muestra la clasificación usando SVM (Parra Rodríguez, 2017)(Cortes, 1995).

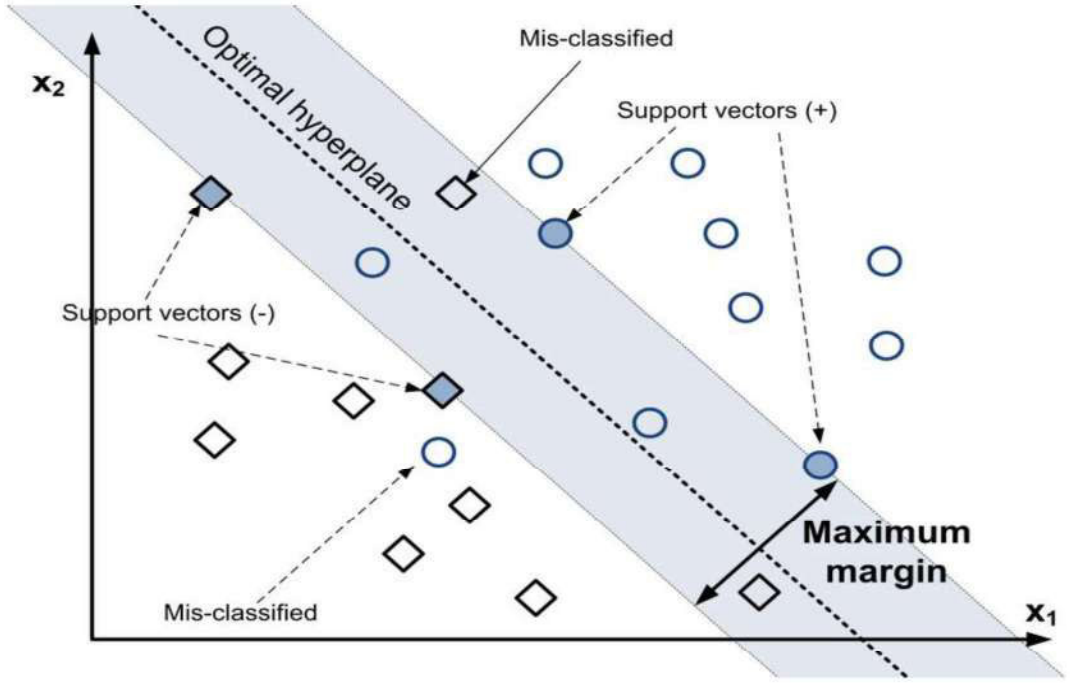


Figura 2-8: Esquema de clasificación usando SVM (Nguyen Duc, Kamwa, Dessaint, & Cao-Duc, 2017)

La clasificación multiclase se la realiza con varias clasificaciones binarias, se las puede hacer de dos maneras comparando una con todas las demás clases por ejemplo si se tiene tres clases se realiza la comparación de la primera con la segunda, la primera con la tercera; y comparando una a una, es decir la primera con la segunda, la segunda con la tercera y la primera con la tercera (Van Den Burg & Groenen, 2016).

2.2.7 Partición de los datos

Para crear un modelo de clasificación, usando una de varias técnicas supervisadas, es necesario realizar dos pasos:

1. Aprendizaje del modelo.

2. Probar la exactitud del modelo.

Para cada uno de los pasos se utiliza un subconjunto del conjunto de datos elegido, lo que quiere decir que la primera decisión que hay que tomar luego de haber elegido el método es la de dividir los datos en los dos subconjuntos mencionados: el primero para entrenamiento y el segundo para prueba.

2.2.7.1 División en porcentaje

El conjunto de datos se divide en dos subconjuntos, cada uno con un porcentaje del número total de filas: se puede mencionar que típicamente se dividen en porcentaje de 50-50, 60-40 o 70-30. Es bastante determinar, de inicio, cuál es la mejor partición, pero en medio de la experimentación se pueden probar y verificar los resultados, para decidir cuál sería la mejor forma de dividir los datos.

2.2.7.2 Validación Cruzada (*Cross Validation*)

Esta forma de dividir los datos para entrenamiento y prueba es muy utilizada en investigaciones en el ámbito de *data mining*, con cantidades limitadas de datos; consiste en dividir el conjunto de datos disponible en k pliegues o segmentos, y sobre esa base se utiliza k-1 pliegues para entrenamiento y el restante para prueba (ver figura 2.9). De hecho, se realizan k iteraciones con todo el conjunto de datos, para finalmente obtener el error medio de cada iteración y luego sumarizarlas para encontrar la exactitud global del modelo entrenado. Los datos de entrenamiento y prueba deben ser independientes, pero esta independencia depende del caso de uso de los datos, así se tiene que para diagnostico los datos

de entrenamiento y prueba deben tener registros de varios sujetos mientras que para pronostico se debe tener varios registros de un sujeto en varios intervalos de tiempo o épocas, por lo tanto depende de la aplicación (Parra Rodríguez, 2017) (Saeb, Lonini, Jayaraman, Mohr, & Kording, 2017).

Figura 2-9: Esquema de Validación Cruzada ("Aprendizaje automático (12) ejemplos de validación cruzada - programador clic," n.d.)

Validación cruzada dejar uno fuera*, Leave One Out Cross-Validation* **(LOOCV)** esta técnica de validación cruzada es una variante de *k-Cross Validation*, en donde k = n-1, siendo n el número de muestras, es decir, se entrena con las muestras disponibles del conjunto de datos excepto uno (ver figura ; se repite n veces, calculando el error en cada muestra; el error estimado utilizando esta técnica es el promedio de todos los errores calculados; es así que al final del

proceso se utilizan todos los datos para entrenamiento y prueba. Se recomienda para conjuntos de datos pequeños, por su alto costo computacional ya que se requiere que el modelo se reajuste y valide n veces. Puede ser usado por cualquier modelo predictivo, como por ejemplo regresión logístico o análisis discriminante lineal (Amat Rodrigo, 2020) (James, Witten, Hastie, & Tibshirani, 2017).

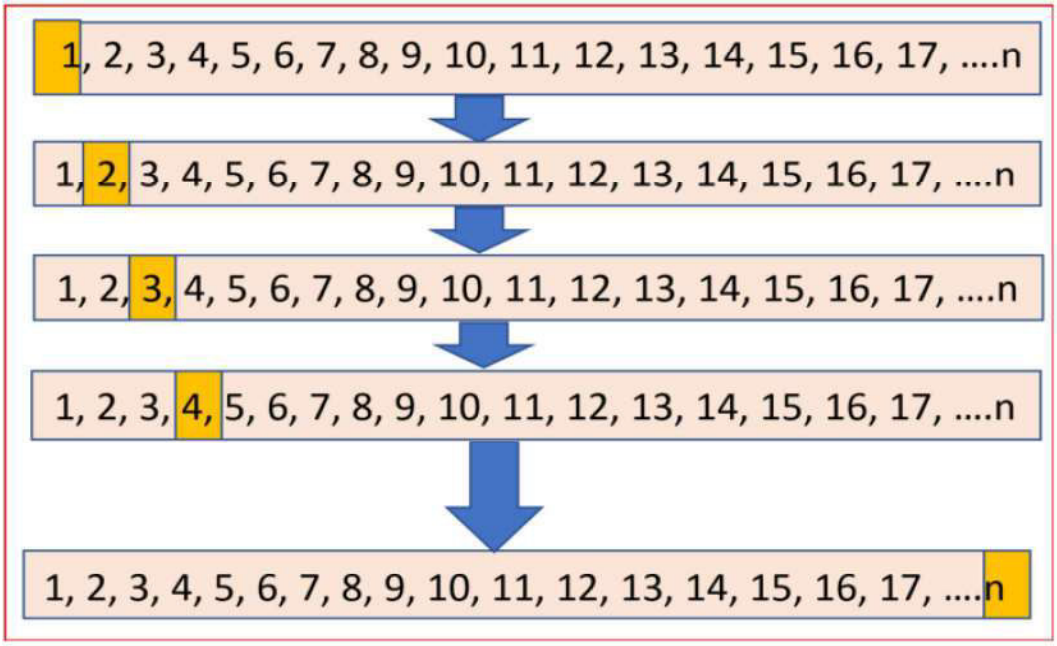


Figura 2-10: Esquema de validación cruzada dejar uno fuera ("K fold y otras técnicas de validación cruzada," 2020)

2.2.8 Métricas de evaluación de la clasificación

Para medir el grado de exactitud de un modelo de clasificación, existen varias métricas reportadas en la literatura científica. A continuación, se detallan algunas de ellas.

Según (Chawla, 2009) y (Provost & Kohavi, 1998) citado en (Ahmadi et al., 2017), un clasificador binario se evalúa normalmente con una matriz de confusión como la que se muestra en la figura 2.11. En esta matriz, las filas denotan los resultados reales dados por las etiquetas del conjunto de datos, mientras que en las columnas están los resultados predichos por el modelo aprendido y probado en el proceso. Cuando el valor positivo de una fila corresponde al positivo de la columna, estamos en el caso de los verdaderos positivos (VP), mientras que, si el valor negativo de una fila corresponde al positivo de la columna, estamos ante el caso de Falsos Positivos (FP). De forma similar, si tuviéramos positivo en la fila y negativo en la columna, estamos ante el caso de Falsos negativos (FN) y si tenemos negativos en la fila y negativos en la columna, estamos en el caso de Verdaderos Negativos (VN).

		Predicción	
		positivos	negativos
realidad	positivos (**P**)	verdaderos positivos (VP)	falsos negativos (FN)
	negativos (N)	falsos positivos (FP)	verdaderos negativos (VN)

Figura 2-11: Matriz de confusión

VP: Es el número de datos positivos clasificados como correctamente positivos.

FP: Es el número de datos negativos clasificados como incorrectamente positivos.

FN: Es el número de datos positivos clasificados como incorrectamente negativos.

VN: Es el número de datos negativos clasificados como correctamente negativos.

Estos valores, sirven de ingrediente para calcular ciertas métricas que son usadas por expertos en diferentes dominios, para evaluar los modelos de clasificación aprendidos y probados. Algunas de estas métricas se detallan a continuación, de manera breve:

Exactitud (*accuracy*) es el número de datos correctamente clasificados con respecto al total de datos de la muestra, esta métrica puede resultar engañosa cuando los datos no están equilibrados con respecto a sus clases (Žižka, Dařena, & Svoboda, 2019)(Shmueli et al., 2020).

$$accuracy = \frac{VP + VN}{\text{TP+TN+FP+FN}} \qquad \text{Ecuación 2}$$

Precisión (*precision*) es la división entre los datos correctamente clasificados para el numero de datos clasificados positivos tanto correctamente como incorrectamente que es en realidad el número de datos relevantes en un conjunto de datos asignados a una clase por el clasificador (Žižka et al., 2019).

$$precicion = \frac{VP}{VP + FP} \qquad \text{Ecuación 3}$$

Exhaustividad (*recall*) también llamado sensibilidad es la proporción de positivos reales que se han identificado correctamente (Žižka et al., 2019).

$$recall = \frac{VP}{VP + FN} \qquad \text{Ecuación 4}$$

Especificidad (*specificity*) es la capacidad del clasificador para descartar correctamente los datos negativos reales.

$$specificity = \frac{VN}{FP + VN} \qquad \text{Ecuación 5}$$

Error, es el número de datos clasificados incorrectamente con respecto al total de datos de la muestra.

$$error = \frac{FP + FN}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \qquad \text{Ecuación 6}$$

F-*measure* es la media armónica ponderada de la precisión y la recuperación:

$$F = \frac{1}{\alpha\left(\frac{1}{precision}\right) + (1-\alpha)\left(\frac{1}{recall}\right)} = \frac{(\beta^2 + 1) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall} \qquad \text{Ecuación 7}$$

donde $\beta^2 = \frac{1-\alpha}{\alpha}$

Cuando β aumenta, la precisión tiene mayor importancia. Cuando *precision* y *recall* tienen la misma importancia, β = 1, y la medida se denomina F1 o F-score, que es la media armónica de la precisión y recall (Žižka et al., 2019), definida en la siguiente formula:

$$F1 = F - score = \frac{2 \cdot precision \cdot recall}{precision + recall} = \frac{2 \cdot TP}{2 \cdot TP + FN + FP} \qquad \text{Ecuación 8}$$

CAPITULO 3

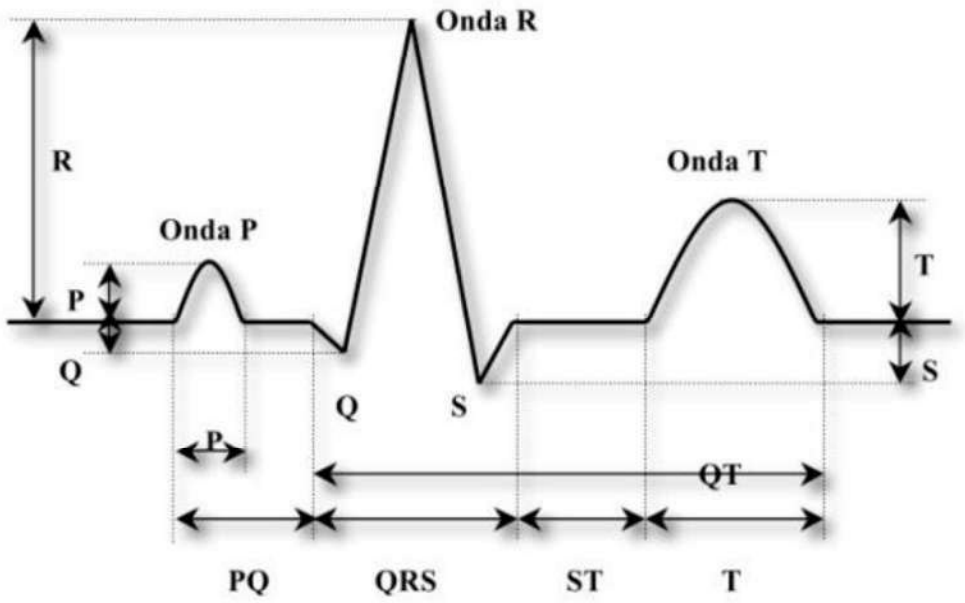
3 Método propuesto

En el presente capítulo se describe el método propuesto, que consiste se inicia con el suavizado de las series temporales, utilizando splines cúbicos para eliminar el ruido, a seguir, se realiza la transformación de las series temporales numéricas a secuencias simbólicas y finalmente se aplica el método supervisado de clasificación de las secuencias simbólicas obtenidas.

3.1 Transformación de series temporales numéricas a simbólicas

El primer paso del método propuesto consiste en transformar las series temporales numéricas en secuencias temporales simbólicas, que consiste en un proceso de abstracción de las series temporales, a las que se les asocian un conjunto de caracteres que, a su vez, forman parte de un alfabeto finito. Con este proceso, el análisis futuro de las series temporales, ganan significativamente en eficiencia, puesto que se consigue una alta reducción de dimensionalidad de las series temporales originales y, por ende, también se reduce el costo computacional.

Estas transformaciones permiten que las series temporales sean más interpretables, puesto que es verdad que el análisis de la forma de las series se ajusta mejor a la manera en que el experto las analiza en la realidad. Un ejemplo de este hecho tiene que ver con el análisis que realiza un cardiólogo a un electrocardiograma (ver figura 3.1): lo realiza observando las amplitudes y ubicación en el tiempo de las ondas generadas por la actividad eléctrica del corazón.



Amplitudes (mV)		Duraciones (msg)	
Onda P	0.25	Intervalo P-R	120 - 200
Onda R	1.6	Intervalo Q-T	350 - 440
Onda T	0.1 – 0.5	Segmento S-T	50 – 150
		Onda P	110
		Intervalo QRS	90

Figura 3-1: Ondas, segmentos, intervalos, amplitudes y duraciones de un ECG ("ECG. Amplitudes y duración de ondas, intervalos y segmentos.," 2018)

La transformación simbólica de la presente propuesta consta de dos pasos naturales:

1. La transformación independiente del dominio y,
2. La transformación dependiente del dominio.

A seguir, en las próximas subsecciones, se detallan estos pasos.

3.1.1 Transformación independiente del dominio

La transformación independiente del dominio, de una serie temporal, consiste en recorrer toda la serie en búsqueda de las formas relevantes que pudiera tener y cada forma encontrada es registrada como un símbolo independiente del dominio. En el presente estudio, las formas que se toman en cuenta son:

Pico, con los atributos: tiempo inicial (t_{ini}), valor inicial (v_{ini}), tiempo final (t_{fin}), valor final (v_{fin}), tiempo del pico (t_{max}) y valor del pico (v_{max}).

Valle, con los atributos: (t_{ini}), valor inicial (v_{ini}), tiempo final (t_{fin}), valor final (v_{fin}), tiempo del valle (t_{min}) y valor del valle (v_{min}).

Constante, con los atributos: tiempo inicial (t_{ini}), valor inicial (v_{ini}), tiempo final (t_{fin}), valor final (v_{fin}).

Subida, con los atributos: tiempo inicial (t_{ini}), valor inicial (v_{ini}), tiempo final (t_{fin}), valor final (v_{fin}).

Bajada, con los atributos: tiempo inicial (t_{ini}), valor inicial (v_{ini}), tiempo final (t_{fin}), valor final (v_{fin}).

A todas las formas descritas se las denomina símbolos, cuyos parámetros son almacenados para identificarlos y clasificarlos cuando fuere necesario.

A fin de identificar de la mejor manera los símbolos independientes del dominio, es preciso que se minimice el ruido que pudiera existir en las series temporales, mediante la utilización de splines cúbicos para suavizar la forma de la serie

temporal, lo cual permitiría eliminar las formas irrelevantes que resultaren sin significado válido entro del dominio.

La figura 3.2, mostrada a continuación, permite tener una idea de cuáles son las formas básicas que pueden ser parte de una serie temporal y que se denominan símbolos independientes del dominio. La tabla 3.1 resume los parámetros que se almacenan para cada símbolo.

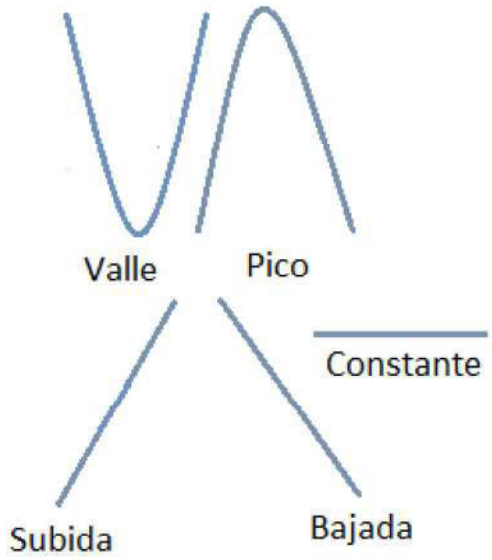


Figura 3-2: Símbolos utilizados en el método propuesto

Tabla 3-1: Parámetros de los símbolos propuestos

Símbolo	Valor 1	Tiempo 1	Valor 2	Tiempo 2	Valor Extr.	Tiempo Extr.
Pico	Vini	Tini	Vfin	Tfin	Vmax	Tmax
Valle	Vini	Tini	Vfin	Tfin	Vmin	Tmin

Símbolo	Valor 1	Tiempo 1	Valor 2	Tiempo 2	Valor Extr.	Tiempo Extr.
Subida	Vini	Tini	Vfin	Tfin		
Bajada	Vini	Tini	Vfin	Tfin		
constante	Vini	Tini	Vfin	Tfin		

Para la identificación de los símbolos se recorre la serie temporal mediante un puntero identificando si es pico o valle según la tendencia, este recorre la serie hasta que hay un cambio de tendencia cuando esto ocurre se guardan los datos de inicio y cambio, sigue recorriendo la serie hasta encontrar otro cambio de tendencia, se realizan los cálculos necesarios para caracterizar el símbolo y se los va almacena en una lista de símbolos como se muestra en un extracto del algoritmo de simbolización independiente del dominio.

Entrada: serie temporal (ts)
Salida: serie temporal simbolizada, simbolos

```
Import pandas
simbolos = []
While lim(a, ts):
    if pico:
        calcular (a, ts)
        simbolos.append(vini,tini,vfin,tfin) #subida
        simbolos.append(vini, tini,vmax,tmax,vini,tfin)#pico
    else:
        simbolos.append (vini, tini,vmax,tmax,vini,tfin)#pico
        simbolos.append (vini,tini,vfin,tfin)#bajada
    if valle:
        calcular(a, ts)
        simbolos.append (vini,tini,vmin,tmin,vfin,tfin) #valle
        simbolos.append (vini,tini,vfin,tfin) #subida
    else:
```

```
simbolos.append (vini,tini,vfin,tfin)#bajada
simbolos.append (vini,tini,vmin,tmin,vfin,tfin) #valle
```

Algoritmo 1: simbolización independiente del dominio

3.1.2 Transformación dependiente del dominio

Con la transformación independiente del dominio, se logra una significativa reducción de dimensionalidad, al tiempo que la serie temporal se convierte en una secuencia simbólica fácil de leer e interpretar. La única limitación de ese nivel de abstracción consiste en que las formas que contiene la secuencia no necesariamente son relevantes para el dominio; es por esta razón, que se requiere un nuevo paso para incorporar el criterio experto a fin de lograr una secuencia temporal, pero de ser posible, conformada solamente con símbolos relevantes para el dominio, según el experto.

Para lograr el objetivo delineado en el párrafo anterior, se propone la realización de un paso denominado Transformación dependiente del dominio, en el cual se identifica el o los símbolos independientes del dominio, que se convertirán en símbolos dependientes del dominio. En este paso, se toman en cuenta los valores representativos de cada símbolo como, por ejemplo: la amplitud, la duración, la posición relativa de cada símbolo, para poder caracterizar los símbolos que son relevantes para el dominio y, por lo tanto, para el análisis que se realiza.

Para la transformación dependiente del dominio se tiene como entrada los símbolos encontrados en el paso anterior y según la información que se tiene del dominio se codifica mediante heurísticas el conocimiento que se tiene del dominio, con lo que

se eliminan los símbolos irrelevantes para el dominio y aquellos que quedan se los denomina símbolos dependientes del dominio. Con esta selección de símbolos, se logra una nueva reducción de dimensionalidad de la serie y además una abstracción de muy alto nivel, donde el significado de la secuencia resultante es sumamente sencillo de interpretar por parte del experto. Por lo indicado, los símbolos encontrados en este paso son los que se utilizarán para la clasificación de las series temporales.

El algoritmo 2 aplica todos los conceptos explicados en los párrafos anteriores, para obtener la secuencia simbólica dependiente del dominio.

Entrada: símbolos Independientes del dominio
Salida: símbolos dependientes del dominio

```
Import pandas
picos = seleccionar_pico(símbolos)
for i in range(len(picos)):
    for j in range(len(picos[ i]
        tm = float(picos[ i] [ j]
        vm = float(picos[ i] [ j]
        vi = float(picos[ i] [ j]
        amplitud = vm - vi
        sd =aplicar_heuristica(tm, amplitud)
        SimbolosDominio.append(sd)

# Fragmento de heurística en el dominio PEATC
…
#sirve para verificar que tipo es la onda I
If i >=1.42 ms and i<=1.97 ms:
    sd.append('Onda I')
    sd.append('Normal')
elif i>= lim_inf and i<=1.41ms:
    sd.append('Onda I')
    sd.append('Adelantada')
elif i>=1.98 ms and i<=lim_sup:
    sd.append('Onda I')
    sd.append('Retardada')
….
```

Algoritmo 2: simbolización dependiente del dominio

3.2 Proceso de Clasificación

Posterior a la transformación de las series temporales numéricas en secuencias simbólicas dependientes del dominio, se procede a realizar el descubrimiento de patrones frecuentes en el conjunto de datos y estos son utilizados para la clasificación.

Los símbolos dependientes del dominio se los representa con caracteres provenientes de un alfabeto que se diseña específicamente para el caso de estudio que se esté tratando, el mismo que se utiliza para la búsqueda de patrones frecuentes.

Esta búsqueda se la realiza para todas y cada una de las clases identificadas en el conjunto de datos, con la finalidad de encontrar los patrones representativos y exclusivos de cada clase, para obtener estos, se elimina los patrones que se encuentren en más de una clase tomando en cuenta siempre la clase de control que en el dominio médico se considera los pacientes sanos.

Una vez obtenidos los patrones exclusivos de cada clase, se realiza un proceso de búsqueda de los conjuntos exclusivos de patrones de cada clase, dentro de la secuencia simbólica asociada al paciente.

Para este último paso del proceso se recibe dos parámetros para decidir la clase del nuevo individuo.

1. ***maxdist***: es la distancia máxima admitida para decidir que un individuo pertenece a una clase determinada, en caso de superar este parámetro, se

considera que el individuo no pertenece a esa clase y se procede a analizar su pertenencia a la clase siguiente.

Para el cálculo de la distancia se utilizó la distancia de edición que penaliza cada operación que se realiza en la secuencia analizada, con el objetivo de igualar a la secuencia objetivo; al final se suman las penalizaciones y el valor obtenido corresponde a la distancia total entre las secuencias, según este método.

2. *Porcentaje de patrones para clasificación* ***ppc***: este parámetro busca especificar el porcentaje en el cual deben coincidir las secuencias simbólicas representativas de una clase determinada y la secuencia analizada. Se trata de una restricción obligatoria de cumplimiento, pues una secuencia no será considerada de una clase, aunque esté dentro de los límites impuestos por el parámetro de la distancia mínima.

Después de realizar las fases anteriores, y de definir los valores para los parámetros expuestos, se utiliza la validación cruzada para la aplicación del método de clasificación expuesto.

CAPITULO 4

4 Experimentación

Para la experimentación se escogió el dominio de los potenciales evocados auditivos de tronco cerebral (PEATC), que se detalla en los siguientes párrafos. Se sigue los pasos del método propuesto y se detalla los resultados obtenidos.

4.1 Descripción del Dominio

Los Potenciales Evocados Auditivos de Tronco Cerebral (PEATC) son señales eléctricas que se generan en el tronco cerebral, como respuesta a un estímulo sonoro que ingresa por el sistema auditivo de una persona. Las señales eléctricas generadas contienen información sobre el estado del sistema auditivo, y a través de su forma es posible concluir si existe alguna patología de ese sistema o si se trata de un paciente saludable. Por lo anotado, el Examen de Potenciales Evocados Auditivos se ha convertido en una prueba necesaria para el diagnóstico de patologías relacionadas con el sistema auditivo.

La prueba se inicia colocando al paciente en posición horizontal, sedado o no – cuando se trata de un niño es necesario que esté sedado – se le conectan electrodos en el vertex y mastoides; la actividad generada por el sistema nervioso central en el tronco encefálico como respuesta neuroeléctricas a un estímulo acústico, produce una serie de ondas en los primeros 10 a 12 milisegundos después de generado el estímulo acústico, este estimulo mecánico se transforma en el Corti en un estímulo eléctrico que recorre toda la vía auditiva hasta llegar a la

corteza cerebral (Trinidad, Trinidad, & De La Cruz, 2008)(Manrique, Jaime, & Algarra, 2014).

Las señales generadas están constituidas por imágenes que representan, en dos dimensiones, la forma en la que varían los potenciales que son del orden de micro voltios.

Las imágenes obtenidas son utilizadas para evaluar el grado de pérdida auditiva; así como también para evaluar los niveles de audición en bebés y niños pequeños con sospecha de sordera y las latencias e interlatencias, que pueden ser obtenidas de esas imágenes, proporcionan información sobre la posible existencia de tumores (Babu, Shakeela, & Priyadarshini, 2017)(Bukard et al., 2007).

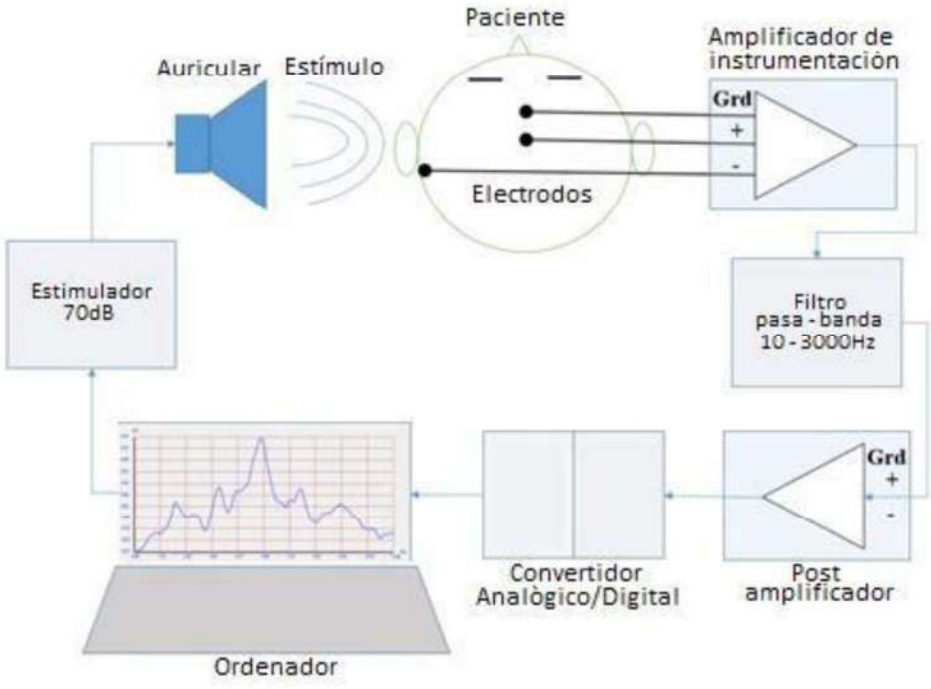


Figura 4-1: Esquema de montaje para la realización del examen de PEATC (Molina Bustamante, 2017)

El procedimiento que se muestra en la figura 4.2, consiste en definir previamente el número de clics, la duración, la intensidad en el convertidor analógico/digital que es el aparato que capta las señales eléctricas emitidas por el cerebro ver figura 4.1 en la que se muestra el montaje para la realización del examen de PEATCs; al paciente se le coloca sensores fijos en tres puntos de la cabeza, luego se envían varios estímulos sonoros (clics) a través del conducto auditivo por medio de auriculares que pueden ser audífonos o vibrador óseo, lo que provoca la respuesta del tronco cerebral, mediante una onda eléctrica denominada Respuesta Auditiva del Tronco Cerebral (*Auditory Brainstem Response*, ABR). Cada ABR está formado por una serie continua de puntos que pueden ser discretizados para obtener un conjunto de pares ordenados (tiempo, valor) y, en consecuencia, se pueden transformar en series temporales que son susceptibles de análisis. Cada ABR tiene una duración de 0 a 15 ms, dependiendo de la calibración del aparato (Trinidad et al., 2008)(Bukard et al., 2007).

Figura 4-2: Procedimiento- Potenciales Evocados Auditivos de Tronco Cerebral (Tejedor, 2016)

Cada ABR consiste en una señal compuesta por varios picos denominados ondas, en el lenguaje del dominio, picos estos que son numeradas con numeración romana; es así como existen las ondas I, II, III, IV, V, VI, etc., las ondas más representativas, para efectos del presente estudio son la onda I, onda III y onda V. Por otro lado, existen las distancias temporales entre las ondas, a las que se les denomina interlatencias, por lo tanto, se tienen la interlatencia I-III, la interlatencia III-V y la interlatencia I-V, como las más importantes para el diagnóstico de diferentes patologías y para el presente estudio. La figura 4.3 muestra un ABR con las características mencionadas tomadas con un estímulo acústico de 2000 clics y una intensidad de 70dB.

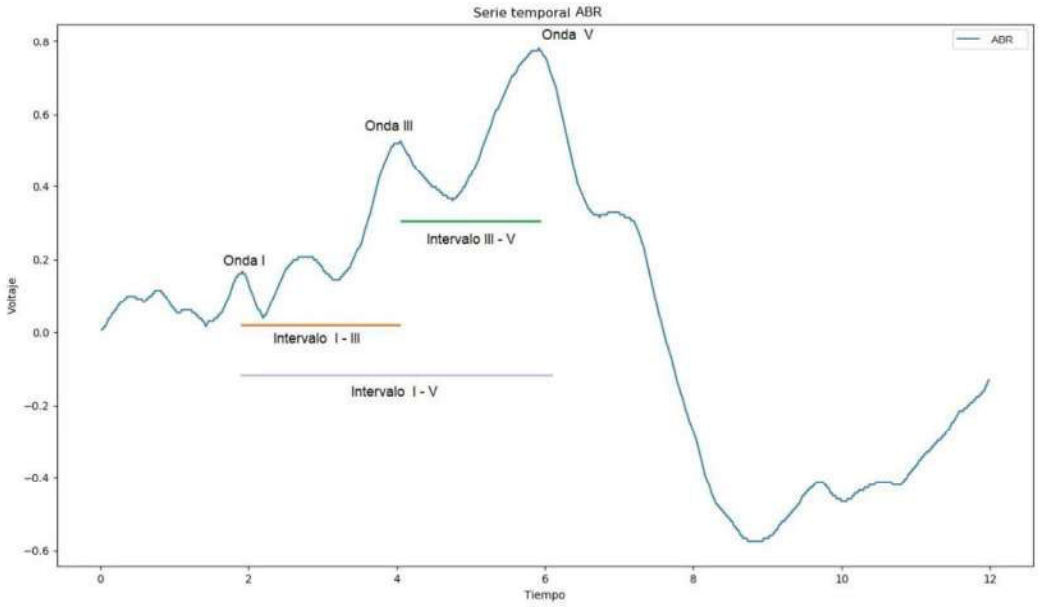


Figura 4-3: Parámetros relevantes de una serie temporal ABR a una intensidad de 70 dB y 2000 clics.

Es importante mencionar que un clic es suficiente para generar una señal que represente a los potenciales auditivos generados por el tronco cerebral, pero lo normal es que esa señal esté saturada de ruido, pues se trata de voltajes sumamente pequeños, y que no refleje lo que realmente los potenciales que están siendo emitidos por el tronco cerebral en ese momento. Para obtener una señal limpia (sin ruido), se requiere que se recojan varios ABRs y al final promediar los valores de los potenciales de cada punto en el tiempo, con lo cual se promedia también el ruido; al tratarse de un ruido aleatorio, éste a veces será positivo y otras será negativo, lo que provocará que su promedio tienda a minimizar el ruido y, por ende, a limpiar el ABR resultante.

Otro aspecto importante, a destacar, es la intensidad del clic. Consiste en el nivel de ruido de éste, el cual se mide en decibelios (dB). Mientras más decibelios tengan los clics, mayor será el nivel de ruido que se envía por el sistema auditivo del paciente. En los casos en que los pacientes tienen bajos niveles de audición, es necesario que se incremente la intensidad de los clics, siendo intensidades típicas: 30 dB, 5s dB, 70 dB, 90 dB y excepcionalmente la intensidad de 110 dB.

4.2 Descripción de los Datos para la experimentación

De la calibración que se realice al aparato dependerá las respuestas obtenidas, es así que uno de los parámetros requeridos es la definición del estímulo acústico que se lo realiza con clics y la intensidad de estos ya que cuanto más intenso sea el estímulo sonoro, mayor es la amplitud, las ondas se definen mejor, y la latencia es menor; de igual manera, a mayor número de clics/s, la latencia de las ondas se alarga y su amplitud es menor. (Trinidad et al., 2008).

Para la experimentación se utilizó los datos que se encuentran en: https://app.box.com/Auditory-Evoked-Potential-Data (Accessed: 28 March 2016). Los datos mencionados fueron obtenidos de exámenes realizados con una intensidad de 70 dB, y para el estímulo acústico se procedió a enviar 2000 clics, que son los parámetros más utilizados por especialistas, debido a que con ellos se obtienen ondas bien definidas.

4.3 Pre procesamiento de los datos

Cada clic produce un ABR, al ser 2000 clic que se utilizan para realizar las pruebas, se generan 2000 ABRs, que son promediadas obteniéndose un solo ABR con lo

cual se reduce el ruido; pero a pesar de eso y por otros factores puede haber ruido en las series, para minimizar el ruido que haya quedado se hizo un proceso de suavizado de las series utilizando splines cúbicos, con lo que se eliminó pequeñas irregularidades presentes en las series, la figura 4.4 muestra un PEATC al que se le suavizo utilizando splines cúbicos.

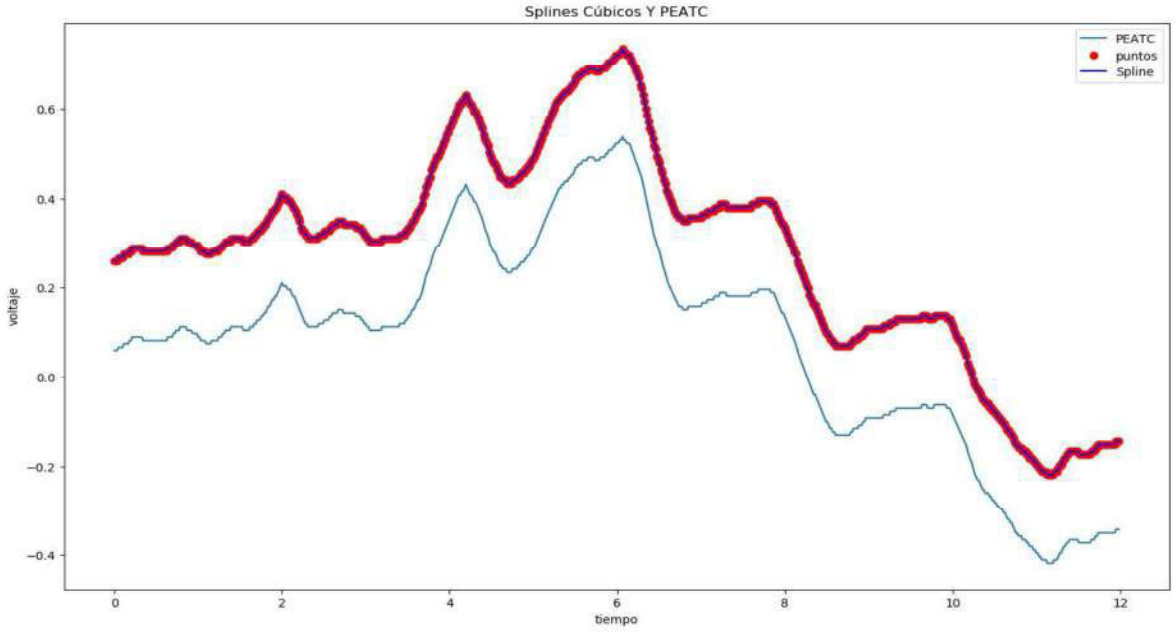


Figura 4-4: Aplicación de splines cúbicos en un PEATC

4.4 Aplicación de simbolización independiente del dominio

Una vez minimizado el ruido de las series se aplicó, a todas y cada una de ellas, el algoritmo para obtener los símbolos independientes del dominio como se muestra en la figura 4.5.

Como se esperaba, para cada serie simbolizada se obtuvo un conjunto de picos, subidas, bajadas y constantes, cuyos valores de atributos fueron almacenados en la base de datos creada para ello.

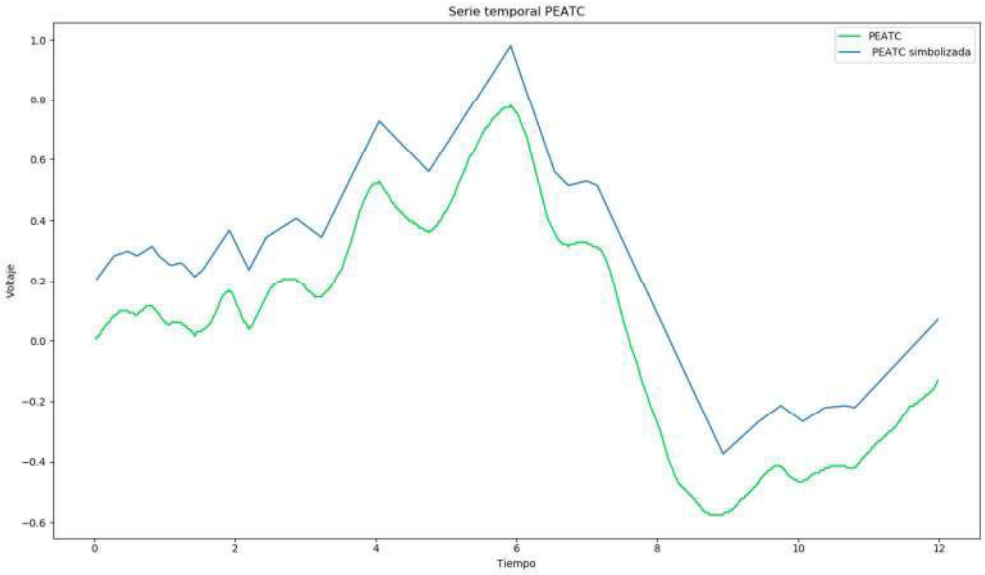


Figura 4-5: Serie temporal PEATC simbolizada independiente del dominio

4.5 Aplicación de simbolización dependiente del dominio

Las secuencias simbólicas independientes del dominio obtenidas en el paso anterior sirvieron de entrada para el procedimiento que incorporaría el conocimiento experto en ellas y, con ello, obtener nuevas secuencias simbólicas, pero ahora dependientes del dominio, es decir, que contienen conceptos del dominio que permiten incrementar el nivel de abstracción de los datos y, por ende, hacerlos más comprensibles y fáciles de analizar. La incorporación de los conceptos del dominio permitió encontrar las ondas I, III y V y los intervalos interlatencias según constan

en las tablas que utilizan los expertos en el dominio y aplicando otras consideraciones aplicadas por profesionales con vasta experiencia en esa área.

La tabla 4.1 muestra los símbolos dependientes del dominio con sus respectivos tipos.

Tabla 4-1: Cuadro con los símbolos dependientes del dominio

Símbolo	Tipo
Onda I	Adelantada Normal Retardada
Onda III	Adelantada Normal Retardada
Onda V	Adelantada Normal Retardada
Intervalo I-III	Corto Normal Largo
Intervalo III-V	Corto Normal Largo
Intervalo I-V	Corto Normal Largo

Tanto los nombres de los símbolos como de sus tipos se eligieron de tal manera que los usuarios del dominio PEATC se sientan familiarizados con ellos y no constituyan una sobrecarga para intervenir activamente en la construcción del modelo.

La figura 4.6 muestra el proceso que se siguió para conseguir las representaciones simbólicas de las series temporales originales:

La primera serie, de abajo hacia arriba, en color verde claro corresponde a la serie temporal original.

Siguiendo en el mismo sentido, la segunda señal es una secuencia simbólica cuyos símbolos son independientes del dominio.

Las dos gráficas siguientes, corresponden a las secuencias simbólicas con el conocimiento del dominio incorporado. En la cuarta gráfica las ondas I, III y V y los intervalos entre ellas pueden ser observados claramente y también se puede apreciar el alto nivel de abstracción conseguido, puesto que, prácticamente se materializan los elementos indispensables que forman parte del dominio y se ha eliminado todo aquello que pudiera considerarse distractor.

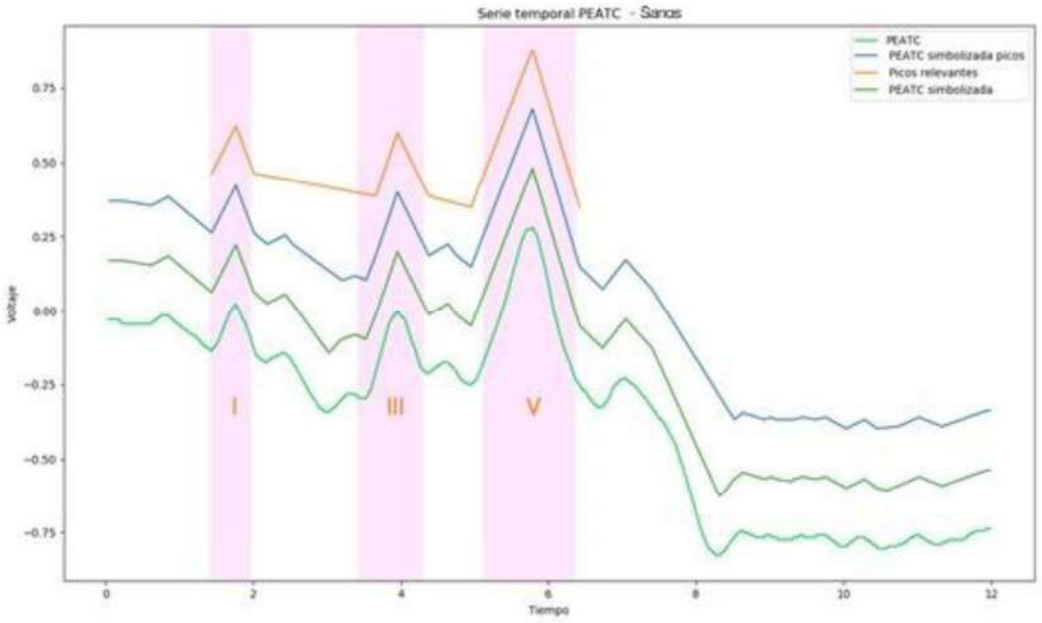


Figura 4-6: PEATC de un paciente sano simbolizado con incorporación del dominio

OBS: En las figuras se han colocado unas franjas de color rosa, a fin de marcar los intervalos de normalidad, para que sirvan de guía en la interpretación de las mismas a la hora de analizarlas.

De acuerdo con la investigación realizada sobre el dominio, cuando las ondas tienen latencias normales y los intervalos entre ellas son normales, el paciente está sano.

Del análisis de los datos y del conocimiento extraído del experto, así como también de la documentación publica consultada se concluye que cuando la onda I es normal, la onda III es normal o retardada, la onda V es retardada y los intervalos entre las ondas son largos se trata de un shwannoma vestibular con implicación en el tronco cerebral, este es un tumor benigno que crece lentamente.

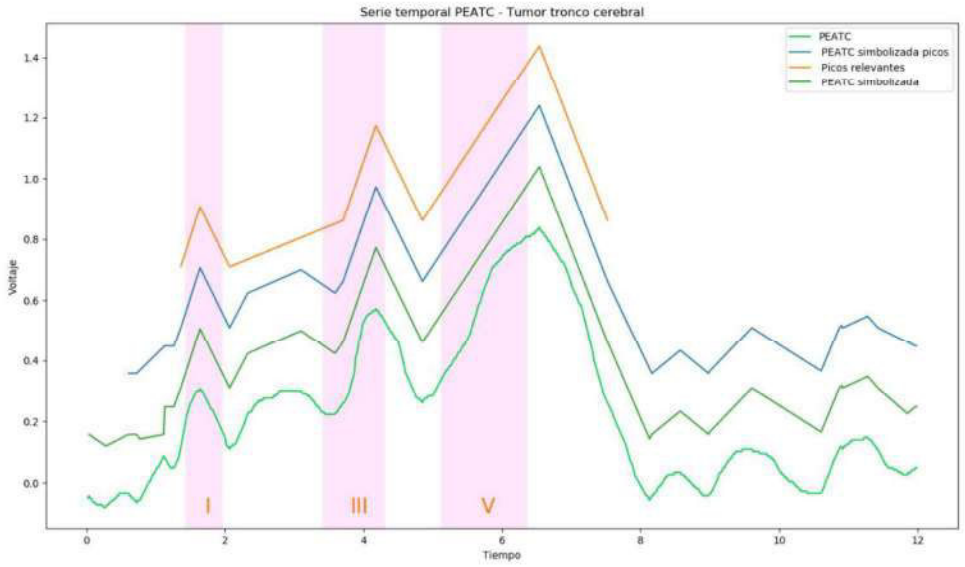


Figura 4-7: PEATC de un paciente con shwannoma vestibular con implicación en el tronco cerebral simbolizado con incorporación del dominio.

La figura 4.7 muestra las diferentes gráficas del ABR de un paciente con shwannoma vestibular con implicación en el tronco cerebral. Se puede observar claramente que la secuencia dependiente del dominio (en naranja) es fácilmente analizable, puesto que está conformada por aquellos símbolos relevantes.

De igual modo que los casos anteriores, después de hacer el análisis de los datos se tiene que cuando la onda I es normal, la onda III es retardada, la onda V es normal en algunos casos puede ser retardada; puede haber ausencia de ondas con más frecuencia la onda III que la I y V; cuando hay ausencia de la onda I o III la onda V es retardada, la ausencia de la onda I implica una probable pérdida auditiva y tumor presente; los intervalos I-III y I-V son largos y el intervalo III-V es normal o

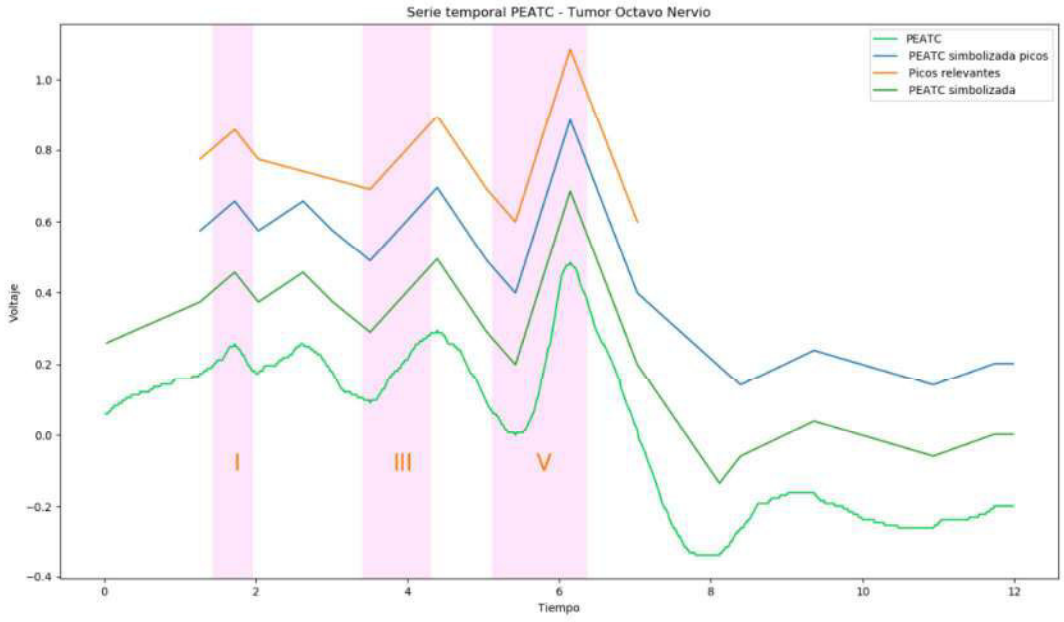


Figura 4-8: PEATC de un paciente con shwannoma vestibular con implicación en el octavo nervio simbolizado con incorporación del dominio.

corto se trata de un paciente con shwannoma vestibular con implicación en el octavo nervio.

En la figura 4.8 se puede observar que el ABR del paciente cumple con los criterios expresados en el párrafo anterior y por ello, es un paciente que padece de un shwannoma vestibular con implicación en el octavo nervio.

4.6 Aplicación del Proceso de Clasificación

Luego de transformar las series temporales de los individuos en secuencias simbólicas dependientes del dominio, se procedió a realizar el descubrimiento de patrones frecuentes en el conjunto de datos, en el que se cuenta con 91 secuencias pertenecientes a pacientes divididos en cuatro clases:

1. Pacientes sanos (48)
2. Pacientes con Pérdida Auditiva Conductiva (18)
3. Pacientes con Shwannoma Vestibular con implicación del octavo nervio (16)
4. Pacientes con Shwannoma Vestibular con implicación del tronco cerebral (9)

La experimentación, realizada hasta llegar a esta instancia, ha develado que cada una de las secuencias temporales dependientes del dominio es una abstracción de la secuencia temporal original en apenas un conjunto de seis símbolos, lo que significa que se ha llegado a una muy significativa disminución de la dimensionalidad del dataset original y, por lo tanto, es posible realizar los pasos que restan de la experimentación, con un nivel de complejidad muy inferior.

Los símbolos dependientes del dominio se los representa con caracteres provenientes de un alfabeto diseñado específicamente para el caso que estamos abordando en el presente estudio y que se muestra en la figura 4.9.

El alfabeto, que se ha definido, servirá para transformar la secuencia de símbolos en una simple secuencia de caracteres y como tal será tratada durante la búsqueda de patrones frecuentes.

CARACTERES DEL ALFABETO DE SIMBOLOS			
Símbolo	**Carácter**	**Símbolo**	**Carácter**
Onda I Adelantada	A	Intervalo I-III Corto	J
Onda I Normal	B	Intervalo I-III Normal	K
Onda I Retardada	C	Intervalo I-III Largo	L
Onda III Adelantada	D	Intervalo III-V Corto	M
Onda III Normal	E	Intervalo III-V Normal	N
Onda III Retardada	F	Intervalo III-V Largo	O
Onda V Adelantada	G	Intervalo I-V Corto	P
Onda V Normal	H	Intervalo I-V Normal	Q
Onda V Retardada	I	Intervalo I-V Largo	R

Figura 4-9: Alfabeto de símbolos dependientes del dominio.

Para el descubrimiento de patrones frecuentes, se usó el mismo algoritmo usado en el método Symbolic Patern-based Classification (SPC) y, en primer lugar, se procedió a encontrar los patrones representativos de los pacientes sanos, que conforman el grupo de control. Para este grupo se descubrieron varios patrones, los que se listan a seguir los patrones frecuentes encontrados para el grupo de pacientes sanos:

- 1patrones: B, E, H, K, N, Q.

- 2patrones: BE, BH, BK, BN, BQ, EH, EK, EN, EQ, HK, HN, HQ, KN, KQ, NQ.
- 3patrones: BEH, BEK, BEN, BEQ, BHK, BHN, BHQ, BKN, BKQ, BNQ, EHK, EHN, EHQ, EKN, EKQ, ENQ, HKN, HKQ, HNQ, KNQ.
- 4patrones: BEHK, BEHN, BEHQ, BEKN, BEKQ, BENQ, BHKN, BHKQ, BHNQ, BKNQ, EHKN, EHKQ, EHNQ, EKNQ, HKNQ.
- 5patrones: BEHKN, BEHKQ, BEHNQ, BEKNQ, BHKNQ, EHKNQ
- 6patrón: BEHKNQ

En total fueron encontrados 63 patrones, los mismos que servirán como representativos del grupo de pacientes sanos y que deberán ser extraídos de los patrones que se encuentren en las secuencias temporales de los pacientes con patologías auditivas, a fin de encontrar el grupo de patrones exclusivos de cada uno de los grupos. En la figura 4.10 se ilustra el proceso de clasificación de un nuevo paciente, cuando ya se tiene construido el modelo para la clase Sano y las clases con trastorno auditivo.

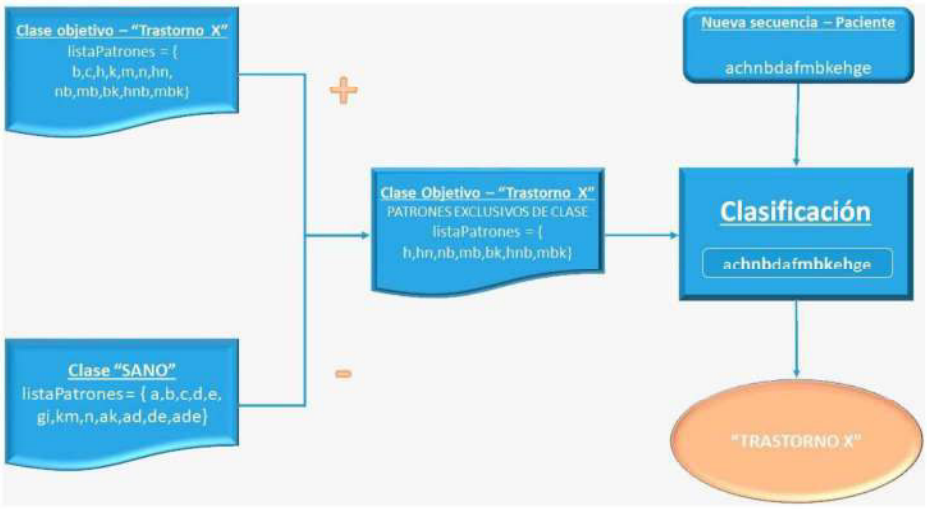


Figura 4-10: Proceso de clasificación a partir de patrones frecuentes.

Como siguiente paso en el proceso, se procedió a encontrar los patrones de cada una de las clases “Trastorno x”.

Para cada clase se obtuvo conjuntos de patrones, de forma similar a como lo fueron para los pacientes sanos y, de cada conjunto de patrones, representativos de cada clase, se eliminaron los patrones que también pertenecían a la clase “Sano” y, de esta manera, se obtuvieron conjuntos de patrones exclusivos de cada clase, con lo cual se asegura que entre clase y clase existan las distancias adecuadas para que sean discriminadas de forma efectiva.

A partir de este punto, la clasificación propiamente dicha es un proceso de búsqueda de los conjuntos exclusivos de patrones de cada clase, dentro de la secuencia simbólica asociada al paciente.

Este último paso del proceso recibe dos parámetros para decidir la clase del nuevo paciente a ser diagnosticado.

3. ***maxdist***: es la distancia máxima admitida para decidir que un paciente pertenece a la clase objetivo, en caso de superar este parámetro, se considera que el paciente no pertenece a esa clase y se procede a analizar su pertenencia a la clase siguiente.

   Para el cálculo de la distancia se utilizó la distancia de edición que penaliza cada operación que se realiza en la secuencia analizada, con el objetivo de igualar a la secuencia objetivo; al final se suman las penalizaciones y el valor

obtenido corresponde a la distancia total entre las secuencias, según este método.

4. *Porcentaje de patrones para clasificación* ***ppc***: este parámetro busca especificar el porcentaje en el cual deben coincidir las secuencias simbólicas representativas de la clase objetivo y la secuencia analizada. Se trata de una restricción obligatoria de cumplimiento, pues una secuencia no será considerada de una clase, aunque esté dentro de los límites impuestos por el parámetro de la distancia mínima.

Luego de cumplir estas fases, se realizó el proceso de clasificación con todos los individuos del dataset, usando una validación cruzada de cinco pliegues, con los siguientes valores para los parámetros del método:

- *ppc*: 0.95
- maxdist: 0.05.

Por tratarse de un dominio del área médica, es importante un alto grado de coincidencia para asegurar el correcto diagnóstico de los pacientes, pues los falsos negativos deben ser evitados de forma prioritaria.

Los resultados obtenidos de esta experimentación se detallan en la siguiente sección.

4.7 Resultados

De la ejecución del método de clasificación, se obtuvo una matriz de confusión que expresa de forma sucinta los resultados en lo referente a los cuatro tipos de resultados posibles a la hora de probar un clasificador, que son:

- Verdadero Positivo (*vp*): se refiere al elemento que fue clasificado como positivo y en realidad es un positivo.
- Verdadero Negativo (*vn*): se trata de aquel elemento que fue clasificado como negativo y en realidad es un negativo.
- Falso Positivo (*fp*): indica que el elemento fue clasificado como positivo, pero en la realidad es un negativo.
- Falso Negativo (*fn*): indica que el elemento fue clasificado como negativo, pero en la realidad es un positivo.

Tabla 4-2 : Matriz de Confusión de la Clasificación

Clasificado como / Real	Sano	PAC	ShV8C	ShVTC	Desc	Total
Sano	47	0	0	0	1	**48**
PAC	0	18	0	0	0	**18**
ShV8N	0	0	16	0	0	**16**
ShVTC	0	0	0	8	1	**9**
Total	**47**	**18**	**16**	**8**	**2**	**91**

La tabla 4.1. muestra la matriz de confusión obtenida luego de ejecutar la clasificación con cuatro clases, de las secuencias temporales simbólicas. En esta matriz se observan esas cuatro clases, representadas por siglas y, además, se puede ver que hay una adicional que corresponde a los elementos no clasificados:

- **Sano**: Corresponde a la clase sano.
- **PAC**: Corresponde a la clase Pérdida Auditiva Conductiva.
- **ShV8N**: Corresponde a la clase Shwannoma Vestibular con implicación del Octavo Nervio.
- **ShVTC**: Corresponde a la clase Shwannoma Vestibular con implicación del Tallo Cerebral.
- **Desc**: Corresponde a los individuos cuya clase no fue reconocida por el clasificador.

Hay que mencionar, como casos especiales a dos individuos que no fueron reconocidos por el método de clasificación y más bien terminaron comportándose como *outliers*, en el sentido de que no obedecieron a ninguno de los conjuntos de patrones que habían sido reconocidos durante el proceso de descubrimiento de tales patrones.

Para medir el rendimiento del método utilizado para la clasificación de las secuencias temporales simbólicas, se tomaron las medidas de rendimiento para clasificación con múltiples clases definidas en el trabajo de (Sokolova & Lapalme, 2009) y que son las siguientes:

1. Exactitud promedio (AA)

   Permite obtener la exactitud promedio por cada una de las clases contenidas en el dataset.

$$\frac{\sum_{i=1}^{n} \frac{tp_i + tn_i}{tp_i + fn_i + fp_i + tn_i}}{n} \qquad \text{Ecuación 9}$$

2. Sensitividad$_M$ (SE_M)

   Se usa para el cálculo de la sensitividad promedio por cada clase. Asocia igual peso a cada clase individual.

$$\frac{\sum_{i=1}^{n} \frac{tp_i}{tp_i + fn_i}}{n} \qquad \text{Ecuación 10}$$

3. Sensitividad$_\mu$ (SE_μ)

   Esta medida calcula la sensitividad con un peso promedio calculado a partir de los pesos individuales, tomando en consideración el tamaño de las clases y asociando igual peso a cada clasificación individual.

$$\frac{\sum_{i=1}^{n} tp_i}{\sum_{i=1}^{n} tp_i + fn_i} \qquad \text{Ecuación 11}$$

4. Especificidad$_M$ (SP_M)

   Se usa para el cálculo de la especificidad promedio por cada clase, para lo cual asocia igual peso a cada clase individual.

$$\frac{\sum_{i=1}^{n} \frac{tn_i}{tn_i + fp_i}}{n} \qquad \text{Ecuación 12}$$

5. Especificidad$_\mu$ (SP_μ)

   Permite calcular la especificidad con el peso promedio que es obtenido con relación al tamaño de cada una de las clases.

$$\frac{\sum_{i=1}^{n} tn_i}{\sum_{i=1}^{n} tn_i + fp_i} \qquad \text{Ecuación 13}$$

A continuación, se muestran los rendimientos calculados con las fórmulas correspondientes a las medidas descritas arriba.

Tabla 4-3: Resultados de la clasificación

Clase	tp	fp	fn	tn	AA	SE_M	SEμ	SP_M	SPμ
Sano	47	0	1	43	0.994	0.967	0.978	1.000	1.000
PAC	18	0	0	73					
ShV8N	16	0	0	75					
ShVTC	8	0	1	82					

CAPITULO 5

5 Conclusiones y trabajos futuros

5.1 Conclusiones

El trabajo realizado permite extraer algunas conclusiones que se pueden transformar en lecciones aprendidas para futuras investigaciones en el área de la minería de datos en series temporales. En el presente capítulo se presentan esas conclusiones a las que se ha llegado a partir de la investigación y las experimentaciones realizadas en este este trabajo.

1. De la revisión sistemática de la literatura que se realizó, se desprende que muy poco trabajo existe sobre el tema central tratado en esta tesis y es el de la simbolización, basada en la forma, de las series temporales. Cuando se intentó encontrar literatura científica referida a la simbolización de series, los resultados obtenidos fueron, casi en su totalidad, trabajos que abordaban el tema de la simbolización sin entrar en el aspecto de la forma de la serie, sino más bien como un artificio, bien elaborado, para la disminución de la dimensionalidad de aquella, siendo el trabajo más relevante el referido a SAX.
2. Al realizar el estudio del estado del arte, nos encontramos con diferentes maneras de tratar la clasificación de las series temporales, sin que casi ninguna de ellas incluyera el conocimiento del experto del dominio, lo cual resultó ser de suma importancia, en este trabajo, a la hora de decidir sobre cuáles serían los símbolos relevantes para representar a las series en el

proceso de clasificación. Hace falta mucho trabajo aún para sentar, de forma sólida, este concepto en la literatura científica.

3. La transformación de las series temporales numéricas en secuencias simbólicas resulta de suma importancia, puesto que, se obtienen múltiples beneficios a la hora de analizar el comportamiento de las series o de clasificarlas como ha sido el caso que nos atañe. En primer lugar, se logra una disminución muy significativa de la dimensionalidad de las series; segundo, se entiende mejor el significado de los datos, al tratarse con formas; tercero, se incorpora el conocimiento del dominio y, con ello, la terminología familiar para los usuarios y para el experto; cuarto se facilita el proceso de clasificación, debido a que se termina clasificando cortas secuencias de caracteres, en lugar de grandes series de números.
4. Los resultados obtenidos, fueron medidos sobre la base de las medidas de rendimiento para clasificadores y que son usadas con mucha frecuencia, especialmente en el área médica. Los valores obtenidos, para tales medidas, son muy parecidos a los del trabajo original y son muy buenos valores, si los comparamos con valores típicos obtenidos de procesos de clasificación similares.
5. De acuerdo con los objetivos, el presente trabajo pretendía mejorar los resultados obtenidos en el trabajo original sobre este tema y, de cierto modo se logró, pues, aunque es cierto que las medidas de rendimiento no fueron superadas, a pesar de que son muy cercanas, también es verdad, que el tiempo de ejecución de los algoritmos implementados son mucho mejores, puesto que, en el trabajo original fueron implementadas usando el lenguaje

C# y para la presente tesis se utilizó el lenguaje Python con las librerías Numpy, Pandas y Scipy.

5.2 Trabajos futuros

1. Dada la falta de suficientes aportes en los aspectos centrados en la simbolización de series temporales, que al tiempo que toman en cuenta la forma, introduzcan también el criterio del experto en el dominio, para lograr no solamente una reducción de dimensionalidad, sino también una mejor comprensión de los símbolos obtenidos; es necesario que se encuentren dominios donde este tipo de procedimientos permitan la obtención de buenos resultados.
2. Usar los dominios encontrados para experimentar con distintos métodos de clasificación, con el afán de incrementar la eficiencia, evitando siempre el sobreajuste.
3. Crear métodos genéricos que funcionen con diversos dominios, y que permitan la construcción de un framework para tratar los diferentes problemas a través de la introducción de parámetros.

6 REFERENCIAS

Agrawal, R., Psaila, G., Wimmers, E. L., & Zaït, M. (1995). Querying Shapes of Histories. *Proceeding of the International Conference on Very Large Data Bases*, 502–514.

Ahmadi, R., Aminshahidy, B., & Shahrabi, J. (2017). Well-testing model identification using time-series shapelets. *Journal of Petroleum Science and Engineering*, *149*, 292–305. https://doi.org/10.1016/J.PETROL.2016.09.044

Alberca, A. S., & Pensamient Matemátic, G. I. E. (2018). Investigación Una nueva taxonomía de colecciones y de funciones de similitud para su comparación A new taxonomy of collections and similarity functions for comparing them. In *Pensamiento Matemático, ISSN-e 2174-0410, Vol. 8, Nº. 2, 2018* (Vol. 8). Retrieved from Grupo de Innovación Educativa website: https://dialnet.unirioja.es/servlet/articulo?codigo=6636692&info=resumen&idioma=ENG

Alonso, F., Martínez, L., Pérez, A., Santamaría, A., & Valente, J. P. (2006). Symbol extraction method and symbolic distance for analysing medical time series. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4345 LNBI*, 311–322. https://doi.org/10.1007/11946465_28

Amat Rodrigo, J. (2020). Validación de modelos predictivos (machine learning): Cross-validation, OneLeaveOut, Bootstraping. Retrieved June 23, 2021, from Cienciadedatos website: https://www.cienciadedatos.net/documentos/30_cross-validation_oneleaveout_bootstrap

Amón, I., & Jiménez, C. (2010). Funciones de Similitud sobre Cadenas de Texto: Una Comparación Basada en la Naturaleza de los Datos. *International Conference on Information Resources Management*, 13. Retrieved from http://aisel.aisnet.org/confirm2010

Apostolico, A., & Guerra, C. (1987). The longest common subsequence problem revisited. *Algorithmica*, *2*(1–4), 315–336. https://doi.org/10.1007/BF01840365

Aprendizaje automático (12) ejemplos de validación cruzada - programador clic. (n.d.). Retrieved May 2, 2021, from https://programmerclick.com/article/8295250019/

Babu, N., Shakeela, B., & Priyadarshini, D. (2017). *Evaluation of Brainstem Auditory Evoked Potentials in Chronic Kidney Disease, Hemodialysis and Renal Transplantation Patients*. Kilpauk Medical College, Chennai, India.

Bartolini, I., Ciaccia, P., & Patella, M. (2002). String matching with metric trees using an approximate distance. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *2476*, 271–283. https://doi.org/10.1007/3-540-45735-6_24

Batal, L., Sacchi, L., Bellazzi, R., & Hauskrecht, M. (2009). Multivariate time series classification with temporal abstractions. *Proceedings of the 22nd International Florida Artificial Intelligence Research Society Conference, FLAIRS-22*, 344–349. Retrieved from www.aaai.org

Boucheham, B. (2012). PLA data reduction for speeding up time series comparison. *International Arab Journal of Information Technology*, *9*(5), 459–464.

Bukard, R. F. ., Don, M., & Eggermont, J. J. (2007). *Auditory Evoked Potential: Basic Principles and Clinical Application* (Frst editi; L. W. & Wilkins, Ed.).

CART: Classification and Regression Trees. (2020). In *The Top Ten Algorithms in Data Mining* (pp. 193–216). https://doi.org/10.1201/9781420089653-17

Caviedes, J. E., Li, B., & Jammula, V. C. (2020). Wearable Sensor Array Design for Spine Posture Monitoring during Exercise Incorporating Biofeedback. *IEEE Transactions on Biomedical Engineering*, *67*(10), 2828–2838. https://doi.org/10.1109/TBME.2020.2971907

Chawla, N. V. (2009). Data Mining for Imbalanced Datasets: An Overview. In *Data Mining and Knowledge Discovery Handbook* (pp. 875–886). https://doi.org/10.1007/978-0-387-09823-4_45

Chen, L., & Ng, R. (2004). *On The Marriage of Lp-norms and Edit Distance*.

Chen, L., Özsu, M. T., & Oria, V. (2005). Robust and fast similarity search for moving object trajectories. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 491–502. https://doi.org/10.1145/1066157.1066213

Cohen, W. W., Ravikumar, P., & Fienberg, S. E. (2003). A comparison of string distance metrics for name-matching tasks. *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web (IIWeb-03)*. Retrieved from www.aaai.org

Corres, G., Esteban, A., García, J., & Zárate, C. (2009). Análisis de series temporales. *Revista Ingeniería Industrial*, *8*(1), 21–33. Retrieved from https://bookdown.org/franciscoparrod/analisis_series/Analisis_Series.html

Cortes, C. (1995). *Support-Vector Networks* (Vol. 20).

Deng, W., Wang, G., & Xu, J. (2016). Piecewise two-dimensional normal cloud representation for time-series data mining. *Information Sciences*, *374*, 32–50. https://doi.org/10.1016/j.ins.2016.09.027

Duan, L., Zhang, Y., Zhao, J., Wang, J., Wang, X., & Zhao, F. (2016). A hybrid approach of SAX and bitmap for machinery fault diagnosis. *2016 International Symposium on Flexible Automation (ISFA)*, 390–396. https://doi.org/10.1109/ISFA.2016.7790195

ECG. Amplitudes y duración de ondas, intervalos y segmentos. (2018). Retrieved June 23, 2021, from https://lamochiladelresi.wordpress.com/2018/06/27/ecg-amplitudes-y-duracion-de-ondas-intervalos-y-segmentos/

Ecured contributors. (2019). Series Temporales -EcuRed. Retrieved April 14, 2021, from 03-09-2019 website: https://www.ecured.cu/Series_Temporales

Friedman, C., & Sideli, R. (1992). Tolerating spelling errors during patient validation. *Computers and Biomedical Research*, *25*(5), 486–509. https://doi.org/10.1016/0010-4809(92)90005-U

García, D. J. C. (2016). Predicción en el dominio del tiempo análisis de series temporales para ingenieros. In *Journal of Chemical Information and Modeling* (Vol. 53).

Ge, L., & Chen, S. (2020). Exact Dynamic Time Warping calculation for weak sparse time series. *Applied Soft Computing Journal*, *96*, 106631. https://doi.org/10.1016/j.asoc.2020.106631

Georgoulas, G., Karvelis, P., Loutas, T., & Stylios, C. D. (2015a). Rolling element bearings diagnostics using the Symbolic Aggregate approXimation. *Mechanical Systems and Signal Processing*, *60–61*, 229–242. https://doi.org/10.1016/J.YMSSP.2015.01.033

Georgoulas, G., Karvelis, P., Loutas, T., & Stylios, C. D. (2015b). Rolling element bearings diagnostics using the Symbolic Aggregate approXimation. *Mechanical Systems and Signal Processing*, *60*, 229–242. https://doi.org/10.1016/j.ymssp.2015.01.033

Gras, J. A. (2001). *Diseños de Series Temporales: Técnicas de Análisis* (J. A. Gras, Ed.). Retrieved from https://books.google.es/books?hl=es&lr=&id=lGptN_0cXMwC&oi=fnd&pg=PA21&dq=Gras,+2001&ots=Sc1g_QcEGi&sig=gLLYwgrw6U2168mdfyYAn3Zyv5c#v=onepage&q=Gras%2C2001&f=false

Ivanovic, M., & Kurbalija, V. (2016). Time series analysis and possible applications. *2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2016 - Proceedings*, 473–479. https://doi.org/10.1109/MIPRO.2016.7522190

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). An introduction to Statistical Learning whit applications in R. In S. N. Y. H. D. London (Ed.), *Current medicinal chemistry* (Volumen 10, Vol. 103). https://doi.org/10.1007/978-1-4614-7138-7

K fold y otras técnicas de validación cruzada. (2020). Retrieved June 23, 2021, from ichi.pro website: https://ichi.pro/es/k-fold-y-otras-tecnicas-de-validacion-cruzada-118762864262942

Keogh, E., Chakrabarti, K., Mehrotra, S., & Pazzani, M. (2001). Locally adaptive dimensionality reduction for indexing large time series databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 151–162. https://doi.org/10.1145/375663.375680

Keogh, Eamonn. (2017). Time Series. In *Encyclopedia of Machine Learning and Data Mining* (pp. 1274–1275). https://doi.org/10.1007/978-1-4899-7687-1_972

Keogh, Eamonn, Chakrabarti, K., Pazzani, M., & Mehrotra, S. (2001). Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. *Knowledge and Information Systems*, *3*(3), 263–286. https://doi.org/10.1007/pl00011669

Kim, S. W., Yoon, J., Park, S., & Won, J. I. (2006). Shape-based retrieval in time-series databases. *Journal of Systems and Software*, *79*(2), 191–203. https://doi.org/10.1016/j.jss.2005.05.004

Kimball, B. A. (1976). Smoothing Data with Cubic Splines [1]. *Agronomy Journal*, *68*(1), 126–129. https://doi.org/10.2134/agronj1976.00021962006800010033x

Kitchenham, B. (2004). *Procedures for Performing Systematic Reviews*.

Krish, V. (2018). Aproximación agregada por partes. Retrieved April 9, 2021, from https://vigne.sh/posts/piecewise-aggregate-approx/

Kumar, M., & Kalia, A. (2012). Preprocessing and symbolic representation of stock data. *Proceedings - 2012 2nd International Conference on Advanced Computing and Communication Technologies, ACCT 2012*, 83–88. https://doi.org/10.1109/ACCT.2012.89

Kumar, N., Lolla, V. N., Keogh, E., Lonardi, S., Ratanamahatana, C. A., & Wei, L. (2005). Time-series bitmaps: A practical visualization tool for working with large time series databases. *Proceedings of the 2005 SIAM International Conference on Data Mining, SDM 2005*, 531–535. https://doi.org/10.1137/1.9781611972757.55

Lee, S. H., Lim, J. S., Kim, J. K., Yang, J., & Lee, Y. (2014). Classification of normal and epileptic seizure

EEG signals using wavelet transform, phase-space reconstruction, and Euclidean distance. *Computer Methods and Programs in Biomedicine*, *116*(1), 10–25. https://doi.org/10.1016/j.cmpb.2014.04.012

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*(8), 707–710.

Li, D., Li, L., Bissyandé, T. F., Klein, J., & Traon, Y. Le. (2016). DSCo: A language modeling approach for time series classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9729*, 294–310. https://doi.org/10.1007/978-3-319-41920-6_22

Li, Guiling, Yan, W., & Wu, Z. (2019). Discovering shapelets with key points in time series classification. *Expert Systems with Applications*, *132*, 76–86. https://doi.org/10.1016/J.ESWA.2019.04.062

Li, Guozhong, Choi, B., Bhowmick, S. S., Wong, G. L. H., Chun, K. P., & Li, S. (2020). Visualet: Visualizing Shapelets for Time Series Classification. *International Conference on Information and Knowledge Management, Proceedings*, 3429–3432. https://doi.org/10.1145/3340531.3417414

Lin, J., Keogh, E., Lonardi, S., & Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms. *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, DMKD '03*, 2–11. https://doi.org/10.1145/882082.882086

Lin, J., & Li, Y. (2009). Finding structural similarity in time series data using bag-of-patterns representation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5566 LNCS*, 461–477. https://doi.org/10.1007/978-3-642-02279-1_33

Lkhagva, B., Suzuki, Y., & Kawagoe, K. (2006). Extended sax: Extension of symbolic aggregate approximation for financial time series data representation. *DEWS2006 4A-I8*. Retrieved from http://www.ieice.or.jp/iss/de/DEWS/DEWS2006/doc/4A-i8.pdf

Ma, Z., Yan, R., Li, K., & Nord, N. (2018). Building energy performance assessment using volatility change based symbolic transformation and hierarchical clustering. *Energy and Buildings*, *166*, 284–295. https://doi.org/10.1016/j.enbuild.2018.02.015

Manrique, M., Jaime, R., & Algarra, M. (2014). *Audiología Audiología ▲entinema* (Primera ed; S. A. CYAN, Proyectos Editoriales, Ed.). España: Ponencia Oficial de la Sociedad Española de Otorrinolaringología y Patología Cérvico-Facial.

Martínez, S., Chuvieco, E., Aguado, I., & Salas, J. (2017). *Número especial*. *49*, 17–32. https://doi.org/10.4995/raet.2017.7182

Mathews , John H. and Fink, K. D. (2000). *Metodos Numericos Con Matlab* (tercera ed; P. HALL, Ed.). Retrieved from http://hpclab.ucentral.edu.co/assets/mns.pdf

Mauricio, J. Al. (Universidad C. de M. (2009). Introducción al Análisis de series temporales. In *Revista Ingeniería Industrial* (Vol. 8).

Métodos de selección dividida para árboles de clasificación en JSTOR. (n.d.). Retrieved June 23,

2021, from https://www.jstor.org/stable/24306157

Molina Bustamante, M. E. (2017). *Modelo para el descubrimiento de patrones en series temporales simbólicas* (Escuela Técnica Superior de Ingenieros Informáticos, Universidad Politécnica de Madrid). Retrieved from http://oa.upm.es/47809/1/MARCO_EDUARDO_MOLINA_BUSTAMANTE.pdf

Molina, M. E., Perez, A., & Valente, J. P. (2016). Classification of auditory brainstem responses through symbolic pattern discovery. *Artificial Intelligence in Medicine*, *70*, 12–30. https://doi.org/10.1016/j.artmed.2016.05.001

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Nguyen Duc, H., Kamwa, I., Dessaint, L.-A., & Cao-Duc, H. (2017). A novel approach for early detection of impending voltage collapse events based on the support vector machine. *International Transactions on Electrical Energy Systems*, *27*(9), e2375. https://doi.org/10.1002/ETEP.2375

Ordóñez, P., DesJardins, M., Feltes, C., Lehmann, C. U., & Fackler, J. (2008). Visualizing multivariate time series data to detect specific medical conditions. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, *2008*, 530–534. Retrieved from /pmc/articles/PMC2656052/

Ordoñez, P., Schwarz, N., Figueroa-Jiménez, A., Garcia-Lebron, L. A., & Roche-Lima, A. (2016). Learning stochastic finite-state transducer to predict individual patient outcomes. *Health and Technology*, *6*(3), 239–245. https://doi.org/10.1007/s12553-016-0146-2

Parra Rodríguez, F. J. (2017). Métodos de clasificación | Estadística y Machine Learning con R. In *Estadística y Machine Learning con R.* (p. 288). Retrieved from https://bookdown.org/content/2274/metodos-de-clasificacion.html#algoritmo-k-vecinos-mas-cercanos

Patel, S. P., & Upadhyay, S. H. (2020). Euclidean distance based feature ranking and subset selection for bearing fault diagnosis. *Expert Systems with Applications*, *154*, 113400. https://doi.org/10.1016/j.eswa.2020.113400

Pham, N. D., Le, Q. L., & Dang, T. K. (2010). Two novel adaptive symbolic representations for similarity search in time series databases. *Advances in Web Technologies and Applications - Proceedings of the 12th Asia-Pacific Web Conference, APWeb 2010*, 181–187. https://doi.org/10.1109/APWeb.2010.23

Provost, F., & Kohavi, R. (1998). Guest editors' introduction: On applied research in machine learning. *Machine Learning*, Vol. 30, pp. 127–132. https://doi.org/10.1023/A:1007442505281

Quinlan, J. R. (1983). Learning Efficient Classification Procedures and Their Application to Chess End Games. In *Machine Learning* (pp. 463–482). https://doi.org/10.1007/978-3-662-12405-5_15

Rajaei, A., Dallalzadeh, E., & Rangarajan, L. (2015). Symbolic representation and classification of medical X-ray images. *Signal, Image and Video Processing*, *9*(3), 715–725. https://doi.org/10.1007/s11760-013-0486-6

Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C., & Kording, K. P. (2017). The need to approximate the use-case in clinical machine learning. *GigaScience*, *6*(5), 1–9. https://doi.org/10.1093/gigascience/gix019

San Segundo, E., Tsanas, A., & Gómez-Vilda, P. (2017). Euclidean Distances as measures of speaker similarity including identical twin pairs: A forensic investigation using source and filter voice characteristics. *Forensic Science International*, *270*, 25–38. https://doi.org/10.1016/j.forsciint.2016.11.020

Santamaría Falcón, A. (2011). *Modelo de descubrimiento de conocimiento para series temporales numéricas aplicando métodos simbólicos*. Facultad de Informática, Escuela Politécnica de Madrid.

Senin, P., & Malinchik, S. (2013). SAX-VSM: Interpretable time series classification using sax and vector space model. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 1175–1180. https://doi.org/10.1109/ICDM.2013.52

Sevcech, J., & Bielikova, M. (2015). Symbolic Time Series Representation for Stream Data Processing. *2015 IEEE Trustcom/BigDataSE/ISPA*, 217–222. https://doi.org/10.1109/Trustcom.2015.586

Shabtai, A. (2016). Anomaly Detection Using the Knowledge-based Temporal Abstraction Method. *ArXiv Preprint ArXiv:1612.04804*. Retrieved from http://arxiv.org/abs/1612.04804

Sharabiani, A., Sharabiani, A., & Darabi, H. (2016). A novel Bayesian and Chain Rule Model on symbolic representation for time series classification. *2016 IEEE International Conference on Automation Science and Engineering (CASE)*, 1014–1019. https://doi.org/10.1109/COASE.2016.7743515

Shknevsky, A., Shahar, Y., & Moskovitch, R. (2017). Consistent discovery of frequent interval-based temporal patterns in chronic patients' data. *Journal of Biomedical Informatics*, *75*, 83–95. https://doi.org/10.1016/j.jbi.2017.10.002

Shmueli, G., Bruce, P. C., Gedeck, P., & Patel, N. R. (2020). *Data Mining for Business Analytics_ Concepts, Techniques and Applications in Python* (FIRTS, Vol. 6; I. John Wiley & Sons, Ed.). USA: editorial Offices.

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, *147*(1), 195–197. https://doi.org/10.1016/0022-2836(81)90087-5

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Taktak, M., Triki, S., & Kamoun, A. (2017). SAX-Based Representation with Longest Common Subsequence Dissimilarity Measure for Time Series Data Classification. *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, 821–828. https://doi.org/10.1109/AICCSA.2017.29

Tamura, K., & Ichimura, T. (2017). Classifying of time series using local sequence alignment and its performance evaluation. *IAENG International Journal of Computer Science*, *44*(4), 462–470. Retrieved from http://www.iaeng.org/IJCS/issues_v44/issue_4/IJCS_44_4_07.pdf

Tejedor, C. (2016). POTENCIALES EVOCADOS AUDITIVOS – Audiología Ondas. Retrieved July 21,

2021, from audiología Ondas website: https://audiologiaondas.com/1911-2/

Tipos de series temporales - Qué es, definición y concepto | 2021 | Economipedia. (n.d.). Retrieved April 9, 2021, from https://economipedia.com/definiciones/tipos-de-series-temporales.html

Trinidad, G., Trinidad, G., & De La Cruz, E. (2008). Potenciales evocados auditivos. *Anales de Pediatria Continuada*, *6*(5), 296–301. https://doi.org/10.1016/S1696-2818(08)74884-4

Van Den Burg, G. J. J., & Groenen, P. J. F. (2016). GenSVM: A Generalized Multiclass Support Vector Machine. In *Journal of Machine Learning Research* (Vol. 17). Retrieved from http://www.stat.osu.edu/

Wang, X., Lin, J., Senin, P., Alamos, L., Oates, T., Gandhi, S., … Frankenstein, S. (2016). RPM : Representative Pattern Mining for Efficient Time Series Classification. *Proceedings of the 19th International Conference on Extending Database Technology*, 185–196.

Yahyaoui, H., & Al-Daihani, R. (2019). A novel trend based SAX reduction technique for time series. *Expert Systems with Applications*, *130*, 113–123. https://doi.org/10.1016/J.ESWA.2019.04.026

Ye, L., & Keogh, E. (2009). Time series shapelets: A new primitive for data mining. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 947–955. https://doi.org/10.1145/1557019.1557122

Yi, B. K., & Faloutsost, C. (2000). Fast time sequence indexing for arbitrary 4 norms. *Proceedings of the 26th International Conference on Very Large Data Bases, VLDB'00*, 385–394. Retrieved from https://kilthub.cmu.edu/articles/journal_contribution/Fast_Time_Sequence_Indexing_for_Arbitrary_Lp_Norms/6605618/files/12096092.pdf

Yin, H., Yang, S., Zhu, X., Ma, S., & Zhang, L. (2015). Symbolic representation based on trend features for knowledge discovery in long time series. *Frontiers of Information Technology & Electronic Engineering*, *16*(9), 744–758. https://doi.org/10.1631/FITEE.1400376

Yin, J., Xu, M., & Zheng, H. (2019). Fault diagnosis of bearing based on Symbolic Aggregate approXimation and Lempel-Ziv. *Measurement*, *138*, 206–216. https://doi.org/10.1016/J.MEASUREMENT.2019.02.011

Yu, Y., Zhu, Y., Wan, D., Liu, H., & Zhao, Q. (2019). A novel symbolic aggregate approximation for time series. *Advances in Intelligent Systems and Computing*, *935*, 805–822. https://doi.org/10.1007/978-3-030-19063-7_65

Zalewski, W., Silva, F., Maletzke, A. G., & Ferrero, C. A. (2016). Exploring shapelet transformation for time series classification in decision trees. *Knowledge-Based Systems*, *112*, 80–91. https://doi.org/10.1016/J.KNOSYS.2016.08.028

Žižka, J., Dařena, F., & Svoboda, A. (2019). Text Mining with Machine Learning. In *Text Mining with Machine Learning*. https://doi.org/10.1201/9780429469275