International Journal of Innovation and Scientific Research
ISSN 2351-8014 Vol. 44 No. 2 Sep. 2019, pp. 123-130

http://www.ijisr.issr-journals.org/

# Clasificación y Clustering de Series de Tiempo para Predicción del Comportamiento de Voltaje de Vehículos almacenados en Patios Automotrices

# [ Time Series Classification and Clustering for Voltage Behaviour Prediction from Vehicles stored inside Automotive Yards ]

***Carolina Flores Peralta, Perfecto M. Quintero F., and Rodolfo Eleazar Pérez Loaiza***

Departamento de Estudios de Posgrado,
Tecnológico Nacional de México/I.T. Apizaco,
Apizaco, Tlaxcala, México



**ABSTRACT:** An Automotive Yard stores a big number of vehicles that stay in the site for different intervals of time (days, weeks or months), requering preventive maintenance during their stay. Through Internet of Things Sensors, the battery voltage of each vehicle is recorded every day generating a large data transformed into time series to analyze it. Classification and Clustering Algorithms based on K-Nearest Neighbors were developed using scikit-learn tool, for the extraction of knowledge from the IoT Data, specifically battery voltaje behavior patterns according to certain vehicle models. The performance of the algorithms was obtained making a comparison between them. The information founded will be of help for the planning of preventive maintenance carried out in the logistics processes of the automotive yard, minimizing the replacement of batteries and along with this the economic and ecological cost.

**KEYWORDS:** Time Series, k-Nearest Neighbors, IoT Sensors, Clustering, Classification.

## 1 INTRODUCCIÓN

El Internet de las Cosas (IoT, Internet of Things) es un concepto que ha tenido un crecimiento en los últimos años con el uso de sistemas embebidos que obtienen información del entorno y la envían a la World Wide Web. Obtener información de los sensores de objetos ubicados en cualquier parte del mundo brinda una gran oportunidad para identificar, rastrear, controlar, evaluar, reparar y planificar objetos o "cosas" dentro de sistemas complejos [1]. El conocimiento que se obtiene de entornos de IoT permite crear nuevos e inteligentes sistemas de apoyo a la toma de decisiones para la optimización de procesos de diferentes campos, por ejemplo, la logística del transporte.

Existen diferentes procesos logísticos aplicados en la industria automotriz, uno de ellos comienza cuando los vehículos de una planta ensambladora o de importación son transportados a un Patio Automotriz donde son administrados y cuidados para después enviarlos a Concesionarias o exportarlos a mercados internacionales. La implementación de Internet de las Cosas en la logística del transporte comienza con un Dispositivo de Monitoreo OBD (On Board Diagnostics), que es instalado en cada uno de los vehículos almacenados, enviando información del estado del vehículo en tiempo real. Estos datos deben ser analizados e interpretados para la toma de decisiones por parte del personal logístico que ayuden a planificar y mejorar los procesos de llegada, mantenimiento y entrega de vehículos.

A través del enfoque de Análisis de Datos de Sensores IoT y la aplicación de técnicas de Aprendizaje Maquina se pretende encontrar información valiosa en los Datos de Registros de Voltaje de Baterías de Vehículos almacenados en un Patio Automotriz que mejore los procesos de una Empresa Logística.

**Corresponding Author:** Carolina Flores Peralta 

El objetivo de este trabajo es desarrollar Algoritmos de Aprendizaje Maquina que sean capaces de encontrar patrones y obtener conocimiento del comportamiento del nivel de voltaje de baterías de vehículos para predecir las descargas de las baterías y optimizar la planeación de mantenimiento preventivo de los vehículos almacenados, minimizando el remplazo de baterías y junto con esto el costo económico y ecológico.

La sección 2 presenta un Análisis Exploratorio y el Pre-procesamiento de Datos, la sección 3 presenta el desarrollo de los algoritmos de Aprendizaje Maquina basados en Vecinos Cercanos. La sección 4 muestra la evaluación de los algoritmos. Finalmente, las conclusiones son dadas en la sección 5.

## 2 ANÁLISIS EXPLORATORIO Y PRE-PROCESAMIENTO DE DATOS

El conjunto de datos utilizado en este trabajo contiene un mes de registros de vehículos, el conjunto de datos está conformado por 3168 registros con 4 parámetros: Día del Mes, VIN (Identificador de Vehículo), Nivel de Voltaje de la Batería y Modelo del Vehículo. Un extracto de los datos se muestra en la Tabla 1.

***Tabla 1. Conjunto de Datos: Registros de Voltaje de Vehículos***

| Día | VIN | Voltaje | Modelo |
|---|---|---|---|
| 1 | VF3CC5FS7HT003052 | 1202 | 5 |
| 1 | VF3CC5FS9HT002985 | 1256 | 7 |
| … | | | |
| 24 | VF3CC5FS7HT003052 | 664 | 5 |
| 24 | VF3CC5FS9HT002985 | 698 | 7 |

### 2.1 SERIES DE TIEMPO

Una Serie de tiempo $\boldsymbol{x} = \{\boldsymbol{x_1}, .., \boldsymbol{x_T}\}$ es un conjunto de valores reales ordenados donde $\boldsymbol{x_1}$es el n-ésimo elemento de $\boldsymbol{x}$ y donde su longitud es denotada por $\boldsymbol{T}$. Dado que una serie de tiempo está conformada por valores consecutivos registrados en un periodo de tiempo, los voltajes de Día 1 al Día 24 de cada vehículo fueron colocados sucesivamente añadiendo el modelo del vehículo como su Clase, obteniendo así 132 series de tiempo etiquetadas, que representan el comportamiento del voltaje de cada vehículo de acuerdo a su modelo. La Tabla 2 presenta un extracto de las series de tiempo obtenidas y la Figura 1 muestra el comportamiento del voltaje de la batería durante un mes agrupando los vehículos de acuerdo a su modelo.

***Tabla 2. Conjunto de Datos: Series de Tiempo***

| Día 1 | Día 2 | … | Día 24 | Modelo |
|---|---|---|---|---|
| 1202 | 998 | … | 745 | 5 |
| … | … | … | … | |
| 1244 | 1171 | … | 651 | 2 |

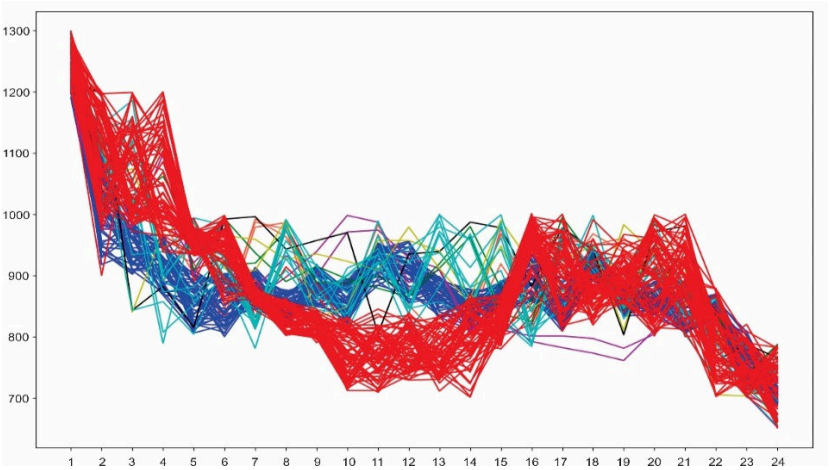


***Fig. 1. Comportamiento de Voltaje por Modelos***

## 3 DESARROLLO DE ALGORITMOS

### 3.1 CLASIFICACIÓN BASADA EN K-VECINOS MÁS CERCANOS

Mediante la aplicación de técnicas de clasificación para el análisis de datos es posible establecer relaciones entre una Clase y los atributos que la describen. La clasificación de datos es un proceso de dos etapas. El primero es la Etapa de Aprendizaje, donde un algoritmo de clasificación construye un Clasificador aprendiendo de un conjunto de datos de entrenamiento que asocia atributos a Clases. En la segunda etapa, llamada Etapa de Clasificación, el Clasificador es usado para predecir la Clase de un conjunto de datos de prueba. La exactitud del Clasificador es estimada por medio de la comparación de la Clase Real con la Clase Predicha por el Clasificador para cada registro del conjunto de datos [2].

La técnica de clasificación basada en k-Vecinos Más Cercanos asume que las instancias que están más cerca, basándose en alguna medida de similitud, tienden a pertenecer a la misma Clase [3]. El conjunto de datos de entrenamiento está formado por *n* atributos, cada registro representa un punto en un espacio n-dimensional. Todos los registros del conjunto de entrenamiento son almacenados en un esquema de espacio n-dimensional. Dado un registro desconocido, el clasificador busca un esquema de espacio de los *k* registros del conjunto de entrenamiento que este más cercano al registro desconocido. Entonces, al dar un registro sin clasificar, los k vecinos más cercanos que estén más cercanos al Registro Objetivo tienen más peso y son seleccionados, combinando sus Clases, y prediciendo así la Clase del Registro Objetivo.

En este trabajo, cada registro es una Serie de Tiempo, donde su Clase es el Modelo de Vehículo y los atributos que describen dicha clase son los valores de voltaje. Para predecir el Modelo de Vehículo de una Serie de Tiempo sin clasificar, el clasificador mide la similaridad del Registro Objetivo con cada registro del conjunto de datos de entrenamiento basado en una medida de distancia entre Series de Tiempo. Después, de acuerdo a los k vecinos más cercanos al Registro Objetivo, las Clases de los k vecinos son combinadas para derivar la predicción de la Clase (Modelo de Vehículo) del Registro Objetivo.

El cálculo de la distancia entre Series de Tiempo es clave para su Clasificación y Clustering. La distancia entre Series de Tiempo sigue la definición siguiente: Dado $\boldsymbol{x} = \{\boldsymbol{x_1}, \dots, \boldsymbol{x_T}\}$ y $\boldsymbol{y} = \{\boldsymbol{y_1}, \dots, \boldsymbol{y_T}\}$ como series de tiempo de longitud T, si la distancia entre dos series de tiempo es definida a través de todos sus puntos, entonces, $\boldsymbol{dist}(\boldsymbol{x}, \boldsymbol{y})$ es la suma de la distancia entre los puntos individuales (Ver Ecuación 1).

$$\boldsymbol{dist}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{T} \boldsymbol{dist}(\boldsymbol{x_t}, \boldsymbol{y_t}) \quad (1)$$

#### 3.1.1 CLASIFICADOR K-NN CON DISTANCIA EUCLIDIANA

Existen diferentes familias de medidas de distancia, una de ellas es la familia o norma $\boldsymbol{L^p}$, donde al cambiar el valor del parámetro $\boldsymbol{p}$ se definen distintas medidas de distancia [4]. A cambiar el valor de $\boldsymbol{p}$ a 2, la distancia Euclidiana (norma $\boldsymbol{L^2}$) se obtiene de acuerdo a la Ecuación 2, que corresponde a una línea recta directa entre dos puntos. La Figura 2 muestra la representación gráfica del cálculo de la distancia Euclidiana entre cada punto de dos Series de Tiempo.

$$\boldsymbol{dist_{Euclidiana}(x,y) = \sqrt{\sum_{i=1}^{T}(x_i - y_i)^2}} \quad (2)$$

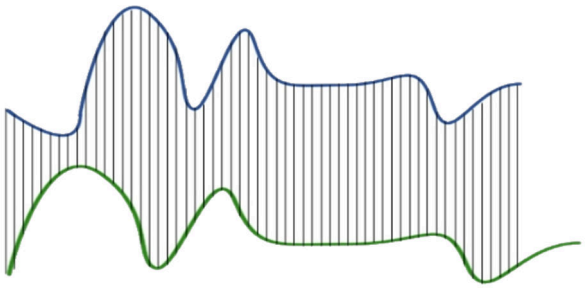

***Fig. 2. Representación gráfica del Cálculo de Distancia Euclidiana entre dos Series de Tiempo***

Un primer clasificador fue desarrollado basado en la técnica de k-Vecinos más Cercanos y utilizando como medida de similaridad entre Series de Tiempo la Distancia Euclidiana.

### 3.1.2 CLASIFICADOR K-NN CON ALINEAMIENTO TEMPORAL DINÁMICO (DTW)

Existe otra medida de distancia llamada Alineamiento Temporal Dinámico (DTW, Dynamic Time Warping) utilizada para encontrar la alineación no lineal óptima entre dos series de tiempo. DTW es un algoritmo para la medición de similitud entre dos secuencias temporales que pueden varias en tiempo o velocidad. Este método encuentra una coincidencia óptima entre dos secuencias dadas y permite una o varias asignaciones, por lo tanto, permite que un punto se asigne a múltiples puntos en la otra secuencia [5].

DTW comienza con la construcción de una Matriz de Costo Local (MCL o mcl), con $\boldsymbol{n} \times \boldsymbol{m}$ dimensiones. Considerando $\boldsymbol{x}$ y $\boldsymbol{y}$ como series de entrada, para cada elemento $(\boldsymbol{i},\boldsymbol{j})$ de la MCL, la norma $\boldsymbol{l_p}$ entre $\boldsymbol{x_i}$ y $\boldsymbol{y_j}$ debe ser calculada con la Ecuación 3.

$$\boldsymbol{mcl(i,j) = \left(\sum_v |x_i^v - y_j^v|^p\right)^{1/p}} \quad (3)$$

Definimos una distancia $\boldsymbol{DTW_p}$, donde $\boldsymbol{p}$ corresponde a la norma $\boldsymbol{l_p}$, la cual es usada en la construcción del MCL. El algoritmo DTW busca la trayectoria que minimice el alineamiento entre $\boldsymbol{x}$ y $\boldsymbol{y}$ por paso iterativo a través de la MCL desde $\boldsymbol{mcl(1,1)}$ hasta $\boldsymbol{mcl(n,m)}$ y agregando el costo. En cada paso, el algoritmo encuentra una dirección en la cual el costo aumenta menos de acuerdo a las restricciones elegidas. Si definimos $\boldsymbol{\phi} = \{(\boldsymbol{1},\boldsymbol{1}), \ldots, (\boldsymbol{n},\boldsymbol{m})\}$ como un conjunto que contiene todos los puntos que caen en la ruta óptima, la distancia final se calculará con la Ecuación 4, donde $\boldsymbol{m_\phi}$ es un coeficiente de ponderación por paso y $\boldsymbol{M_\phi}$ es la constante de normalización correspondiente [6]. La Figura 3 muestra la representación gráfica del cálculo de la distancia entre cada punto de dos Series de Tiempo.

$$\boldsymbol{DTW_p(x,y) = \left(\sum \frac{m_\phi lcm(k)^p}{M_\phi}\right)^{1/p}, \forall k \in \phi} \quad (4)$$

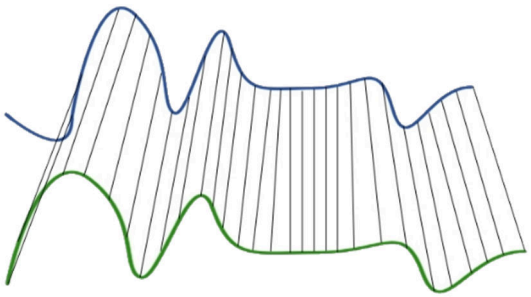

***Fig. 3. Representación gráfica del cálculo de Distancia DTW entre dos Series de Tiempo***

### 3.2 Clustering para Series de Tiempo basado en k-Means y Alineamiento Temporal Dinámico

Clustering es un proceso de partición de un conjunto de datos de objetos u observaciones en subconjuntos. Cada subconjunto es un Cluster, por lo que los objetos de un cluster son similares entre sí, y distintos a los objetos en otros clusters[2]. Como método de clustering, se utilizó un enfoque de k-vecinos más cercanos llamado k-Means utilizando como medida de Distancia el Alineamiento Temporal Dinámico DTW.

Formalmente, dado un conjunto de datos D de n objetos y k, el número de clusters a crear, el algoritmo k-Means organiza los objetos en k particiones ($k \leq n$), donde cada partición representa un cluster. Los clusters son formados para optimizar un criterio objetivo, en este caso, la función de distinción de dos series de tiempo basados en su distancia DTW, de modo que los objetos dentro de un grupo son similares entre sí y diferentes a los objetos en otros grupos en términos de los atributos del conjunto de datos, en este caso, los valores del voltaje de cada vehículo. La Figura 4 presenta los clusters encontrados dentro del conjunto de series de tiempo.

#### 3.2.1 Algoritmo de Predicción de Series de Tiempo por Clustering

Del previo trabajo de clustering, los 3 clusters encontrados serán tomados como Modelos de Clusters k, donde cada *k* está asociado a un Modelo de Vehículo. Para identificar el Modelo de Vehículo de una serie de tiempo, se encontrará el cluster al que pertenece, asociándole entonces el Modelo de Vehículo del *k* Modelo. La Figura 5 muestra las series de tiempo asociadas a cada cluster.

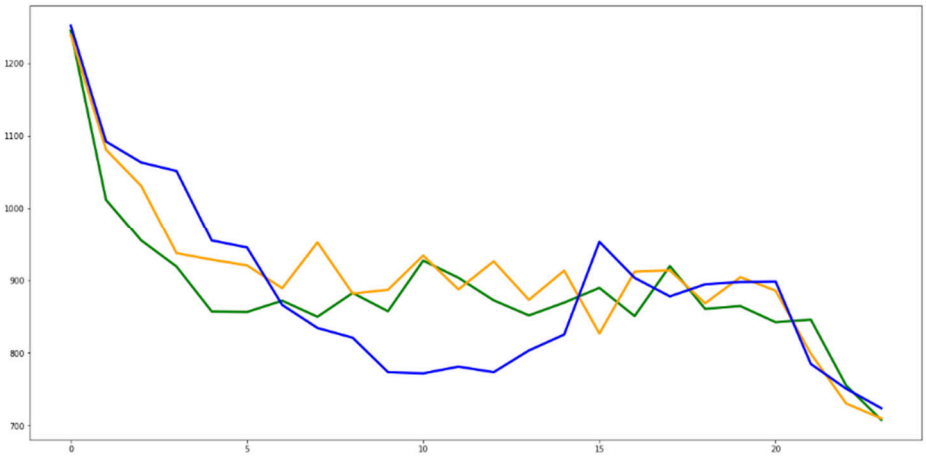


***Fig. 4. Clusters***

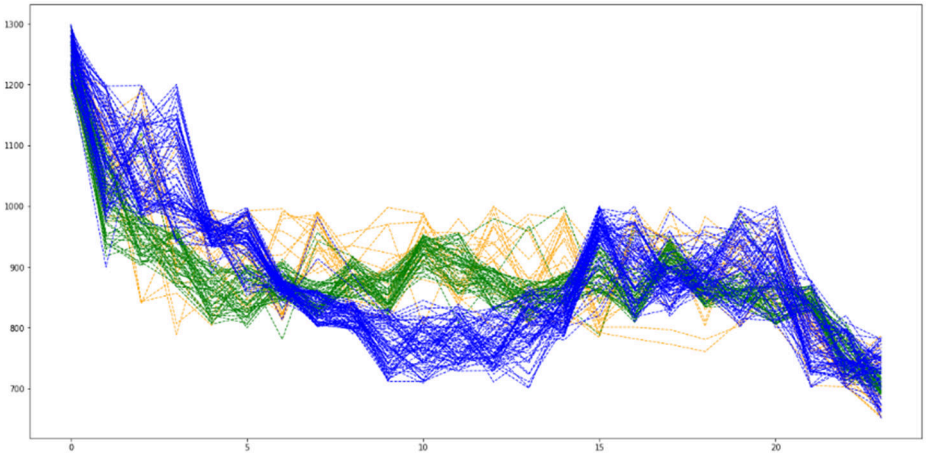


***Fig. 5. Miembros de cada Cluster***

## 4 EVALUACIÓN

Para evaluar los tres modelos construidos, evaluaremos la exactitud de cada uno, es decir, la capacidad predictiva de cada clasificador. La exactitud de un clasificador es obtenida por el porcentaje de registros de un conjunto de prueba que son correctamente clasificadas. Además, para garantizar una evaluación correcta de la exactitud de cada modelo, se utilizó la técnica Validación Cruzada k-fold.

### 4.1 EXACTITUD

Para evaluar los clasificadores, se utiliza un conjunto de pruebas de tuplas o registros etiquetados con su Clase, con tuplas positivas P, y tuplas negativas N. En cada tupla, se compara la Etiqueta de clase obtenida por del clasificador con la Etiqueta de clase Conocida de la tupla. Existen 4 términos adicionales conocidos como "bloques de construcción" utilizado para el cálculo de la exactitud [2]:

- TP Verdaderos positivos: Sea TP número de verdaderos positivos, es decir, tuplas positivas clasificadas correctamente.
- TN Negativos verdaderos: Sea N el número de verdaderos negativos, es decir, tuplas negativas clasificadas correctamente.
- FP Falsos positivos: Sea FP el número de falsos positivos, es decir, tuplas negativas etiquetadas incorrectamente como positivas.
- FN Falsos negativos: Sea FN, el número de falsos negativos, es decir, tuplas positivas etiquetadas incorretamente como negativas.

La matriz de confusión presentada en la Tabla 3 sirve como herramienta para analizar el desempeño de los clasificadores en el reconocimiento de las diferentes clases del conjunto de datos. Los términos TP y TN determinan un buen desempeño del clasificador, y los términos FP y FN el caso contrario, cuando el clasificador asigna una clase errónea

***Tabla 3. Matriz de Confusión***

| | **Clase Predicha** | | | |
|---|---|---|---|---|
| **Clase Actual** | | Yes | No | Total |
| | Yes | TP | FN | P |
| | No | FP | TN | N |
| | Total | P' | N' | P+N |

La exactitud de un clasificador para un conjunto de datos de prueba es calculada con la Ecuación 5, que obtiene el porcentaje de tuplas que son clasificadas correctamente.

$$Exactitud = \frac{TP + TN}{P+N} \tag{5}$$

### 4.2 VALIDACIÓN CRUZADA K-FOLD

En el cálculo de la exactitud de los clasificadores, los datos a clasificar se dividen aleatoriamente en un conjunto de entrenamiento y un conjunto de prueba, la estimación suele ser pesimista porque solo una parte de los datos es usada para la comprobación de la exactitud. En la validación cruzada, dos tercios del conjunto de datos son usados para el conjunto de entrenamiento, y el tercio sobrante como conjunto de prueba.

La Validación Cruzada k-Fold divide aleatoriamente el conjunto de datos en k "pliegues" (folds) exclusivos $D_1, D_2, \dots, D_k$, aproximadamente del mismo tamaño. El entrenamiento del clasificador y su evaluación es realizada n veces, cuando la iteración es $i$, el pliegue $D_i$ se retiene como conjunto de prueba, entrenando al clasificador con los pliegues sobrantes $k - 1$.

Cada pliegue se usa el mismo número de veces para entrenar el clasificador y una sola vez como conjunto de prueba, para calcular la exactitud en la clasificación se divide el total de clasificaciones correctas de las $k$ iteraciones entre el total de tuplas del conjunto de datos.

### 4.3 RESULTADOS

La evaluación de los diferentes algoritmos desarrollados fue realizada a partir del cálculo de la Exactitud y Validación Cruzada de cada uno. La Tabla presenta las puntuaciones de la Exactitud y la Validación Cruzada obtenidas por cada algoritmo. Al usar el Alineamiento Temporal Dinámico (DTW) como medida de distancia entre series de tiempo, el clasificador tuvo una mejora significativa.

***Tabla 4. Evaluación de Modelos***

| Clasificador | Puntuación de Exactitud | Puntuación de Validación Cruzada |
|---|---|---|
| Clasificador k-NN con Distancia Euclidiana | 0.61198 | 0.4678 |
| Clasificador k-NN con Alineamiento Temporal Dinámico (DTW) | 0.8888 | 0.9037 |
| Algoritmo de Predicción de Series de Tiempo por Clustering | 0.96 | 0.7884 |

## 5 CONCLUSIONES

En este artículo se presentó el trabajo inicial del desarrollo de modelos de clasificación y clustering para la predicción del comportamiento de baterías de vehículos almacenados en un Patio Automotriz. El conjunto de datos utilizado es pequeño, para algunos modelos de vehículos existen pocos registros por lo que debilitaron un poco el aprendizaje de los algoritmos para ciertos modelos.

Sin embargo, mediante el uso de series de tiempo que representan el comportamiento de las baterías de los vehículos y la aplicación de una medida de distancia adecuada, tal como DTW, fue posible obtener patrones del comportamiento de la batería de acuerdo al Modelo del Vehículo.

Utilizando enfoques de clasificación y clustering, fue posible obtener una predicción del comportamiento de la batería de los vehículos de diferentes modelos que ayuden a la planificación del mantenimiento correctivo y predictivo llevado a cabo en los procesos logísticos de un Patio Automotriz.

### AGRADECIMIENTO

El trabajo de investigación fue realizado bajo la supervisión de Pascal Poncelet, Ph.D., por lo que expresamos sinceros agradecimientos. Al Laboratorio de Informática, Robótica, Electrónica y Microelectrónica de Montpellier por abrir sus puertas para realizar una estancia de colaboración científica. Agradecemos también el apoyo económico del Consejo Nacional de Ciencia y Tecnología (CONACYT, México).

### REFERENCIAS

[1] Bibri, S. E., *The shaping of Ambient Intelligence and the Internet of Things: histórico epistemic, socio–cultural, politico–institutional and eco–environmental dimensions*, Berlin, Heidelberg: Springer–Verlag, 2015.
[2] Jiawei Han, Jian Pei, Micheline Kamber, *Data Mining: Concepts and Techniques*, Edition 3, Elsevier, 2011.
[3] T. Cover and P. Hart, "Nearest neighbor pattern classification", *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21-27, Enero 1967.
[4] Mirco Kocher, Jacques Savoy, "Distance measures in author profiling",*Information Processing & Management*, vol. 53, Issue 5,pp. 1103-1119, 2017.
[5] R. Ma and R. Angryk, "Distance and Density Clustering for Time Series Data", *IEEE International Conference on Data Mining Workshops (ICDMW)*, New Orleans, LA, 2017, pp. 25-32, 2017.
[6] Sarda-Espinosa, Alexis. "Comparing Time-Series Clustering Algorithms in R Using the dtwclust Package." 2017.
[7] Simon Elias Bibri, "The IoT for smart sustainable cities of the future: An analytical framework for sensor-based big data applications for environmental sustainability", *Sustainable Cities and Society*, vol 38, pp. 230-253, ISSN 2210-6707, 2018.
[8] Grazia Speranza, M., "Trends in transportation and logistics", *European Journal of Operational Research*, vol. 264, ISSN 0377-2217, 2018.

[9] Rahat Iqbal, Faiyaz Doctor, Brian More, Shahid Mahmud, Usman Yousuf, "Big Data analytics and Computational Intelligence for Cyber–Physical Systems: Recent trends and state of the art applications", *Future Generation Computer Systems*, ISSN 0167-739X, 2017.

[10] Yen-Hsien Lee, Chih-Ping Wei, Tsang-Hsiang Cheng, Ching-Ting Yang, "Nearest-neighbor-based approach to time-series classification", *Decision Support Systems*, vol. 53, Issue 1, pp. 207-217, 2012.

[11] J. Yin, D. Zhou and Q. Xie, "A Clustering Algorithm for Time Series Data", *Seventh International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT 06)*, Taipei, 2006, pp. 119-122, 2006.

[12] V. Prema, K. Uma Rao, "Development of statistical time series models for solar power prediction", *Renewable Energy*, vol. 83, pp. 100-109, 2015.

[13] H. Kremer, S. Gunnemann and T. Seidl, "Detecting Climate Change in Multivariate Time Series Data by Novel Clustering and Cluster Tracing Techniques", *IEEE International Conference on Data Mining Workshops*, Sydney, NSW, 2010, pp. 96-97, 2010.