Identification of risk zones for Road Safety through unsupervised learning algorithms

Abstract– ***The following work applies Machine Learning algorithms as a tool for a possible solution to the problem of citizen security in a South American city. This application aims to reduce the threat risk to the physical integrity of pedestrians through the geolocation, in real time, using safer places to walk. A database of free disposal for the user is the Open Data San Isidro, district of Lima, Peru, which has been used in the development of this work. This database keeps records of different accidents types (most of the automobile type) occurring in different places of this district, this data will be used to determine safe areas in the route from one place to another, decreasing the probability of suffering an accident.***

For this work, techniques of non-supervised learning algorithms of Clustering type: k-Means have been used. Likewise, a reduction of dimensions has previously been made using the Principal Component Analysis (PCA) technique.

Keywords- Machine learning, Smart City, Open Data PCA, k-Means, safe routes, Data Mining, citizen security

Digital Object Identifier (DOI):http://dx.doi.org/10.18687/LACCEI2018.1.1.413
ISBN: 978-0-9993443-1-6
ISSN: 2414-6390

Identificación de zonas de riesgo para la Seguridad Vial mediante algoritmos de aprendizaje no supervisado

Jesús Lovón-Melgarejo, Alonso Tenorio-Trigoso, Manuel Castillo-Cara, Daniel Miranda
Center of Information and Communication Technologies
Universidad Nacional de Ingeniería, Av. Tupac Amaru, 210, Rimac, Lima, 25, Perú
Email: {jlovonm, atenoriot, mcastillo}@uni.edu.pe, dmirandam@uni.pe

Resumen—El siguiente trabajo establece algoritmos de Machine Learning como herramienta para una posible solución al problema de seguridad vial en una ciudad a través de datos abiertos. El propósito del análisis desarrollado es reducir el riesgo de amenaza de la integridad física de los peatones mediante la geolocalización en tiempo real, tomando en cuenta los lugares más seguros para transitar. Se estudió una base de datos de disposición libre (portal de datos abiertos de San Isidro, distrito de Lima, Perú). Estos datos guardan registros de distintos tipos de accidentes (la mayoría del tipo automovilístico) ocurridos en diferentes lugares de este distrito, los cuales se analizarán para establecer un recorrido seguro, disminuyendo la probabilidad de que un usuario sufra un accidente.

Por tanto, para este trabajo se han usado técnicas de algoritmos de aprendizaje No Supervisado (Clustering): *k*-Means. Asimismo, previamente, se ha realizado un tratamiento de datos utilizando la técnica de Análisis de Componentes Principales (PCA).

Index Terms—Machine learning; Smart City; Open Data; PCA; k-Means; safe routes; Data Mining; citizen security.

1. INTRODUCCIÓN

En la actualidad todas las ciudades a nivel mundial buscan crear un entorno más ergonómico, eficiente y centrada en el ciudadano [1]. Para ello, el Internet of Things (IoT), los datos y la participación ciudadana son los pilares básico para una Smart City[2], [3]. Todos estos datos generados por las Tecnologías de la Información y Comunicación (TICs) desplegadas en la ciudad han supuesto un gran reto a la hora de poder analizar y sacar conclusiones centradas en el ciudadano [4]. Para ello, las infraestructuras de computación distribuidas han adquirido gran interés por la comunidad académica y tecnológica, sobre todo con el despliegue de IoT para las Smart Cities y Agricultura de precisión. Entre estas infraestructuras, actualmente se encuentra trabajando, en mayor medida, el paradigma Fog Computing, ya que permite una optimización de recursos en nivel edge (edge computing) [5], dejando libre para análisis de datos de una o más dimensiones al nivel core (cloud computing) [6], [7].

En este sentido, para el análisis en macrodatos en el core level, podemos encontrar numerosas técnicas algorítmicas que permiten analizar todos estos datos que tenemos a nuestra disposición, normalmente obtenidos a través de las bases de datos abiertas. Estos tipos de técnicas de análisis de datos pueden clasificarse en dos técnicas basadas en el aprendizaje: (i) algoritmos de aprendizaje supervisado (SLAs); y (ii) algoritmos de aprendizaje no supervisado (ULAs); que tienen el objetivo principal de realizar una toma de decisiones basadas en el aprendizaje.

Por un lado, las técnicas SLAs resuelven un gran ámbito de problemas en la actualidad, no solamente en Smart Cities, sino también en temas de seguridad para TICs como se expone en [8]. Más cercano a la temática de este trabajo, podemos observar como los autores de [9] establecen las rutas más seguras a los ciudadanos por geolocalización realizando un análisis a través de una base de datos abierta con los modelos Random Forest (RF) y Multiple Logistic Regression (LoR).

Por otro lado, los ULAs son utilizados en problemas donde los datos no se encuentran etiquetados (clasificados). Este tipo de algoritmos también se utilizan en problemas de muy diferente índole. Por ejemplo, para análisis sismológicos se logró obtener en Irán nuevos patrones de comportamiento que permitieron un comportamiento diferente de los sismos. Este proceso se realizó debido a la gran cantidad de datos que se tienen sobre la ocurrencia de sismos en diferentes partes del mundo [10], [11]. Otro trabajo importante se puede observar en los autores [12], [13], quienes desarrollan una aplicación para comprimir una imagen en formato PNG, es decir, una imagen de miles de colores es reducida a 16 colores, disminuyendo así considerablemente la cantidad de bits.

Como se observa en la literatura, los algoritmos basados en aprendizaje han tenido muy diferentes usos y aplicaciones. En este trabajo, se va a analizar un caso de uso en el cual se desea realizar una aplicación de Seguridad Vial como la que se ha desarrollado en [9]. Sin embargo, para una mejor predicción se utilizarán ULAs, ya que estos algoritmos son óptimos cuando no hay una etiqueta o clasificación previa y solo existen los datos; como son los datos de Seguridad Vial en el distrito de San Isidro en Lima (Perú) [14]. Para ello, se han evaluado dos algoritmos ULAs: Análisis de componentes

Digital Object Identifier (DOI): http://dx.doi.org/10.18687/LACCEI2018.1.1.413
ISBN: 978-0-9993443-1-6
ISSN: 2414-6390

principales (PCA) [15] y *k*-Means [16]; los cuales, PCA simplifica la complejidad de los datos manteniendo su relación y realizando *k*-Means la clasificación de estos.

Por tanto, al igual que en [9], se ha utilizado la base de datos que provee información de los accidentes ocurridos en el distrito, lugar, fecha, hora, etc. Todos estos datos serán nuestras dimensiones originales, transformándolos con PCA y agrupándolos con k-Means para así evaluar las zonas de riesgo.

En este contexto, el trabajo se encuentra estructurado como muestra la Figura 1. En la Sección 2 se especificarán la metodología de desarrollo del trabajo, es decir, algoritmos y el objetivo de estos. La Sección 3 muestra un primer análisis de las tablas y los datos seleccionados, realizando un primer tratamiento para, posteriormente, en la Sección 4 evaluar el resultado obtenido por PCA y *k*-Means; discutiendo las primeras conclusiones sobre la importancia de estas técnicas algorítmicas para este tipo de datos. Finalmente, en la Sección 5 se discutirán cuales son las conclusiones obtenidas de este trabajo, estableciendo posteriormente la línea de trabajo a futuro.

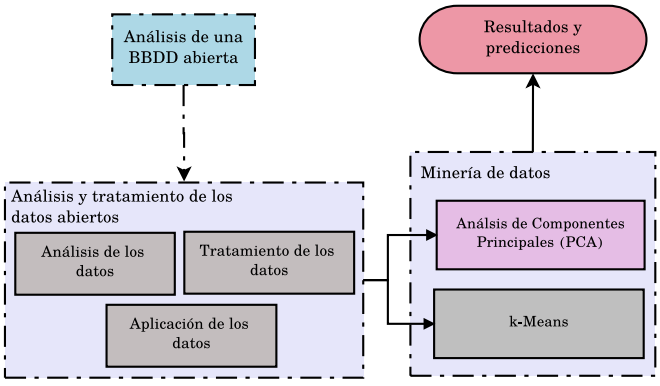


Figura 1. Esquema general.

2. METODOLOGÍA DESARROLLO

Los ULAs, a diferencia de los SLAs, no utilizan datos históricos, sino que ellos tienen que descubrir patrones y tendencias en los datos a analizar, por tanto, los datos no se encuentra etiquetados [17]. Por ello, una clasificación de los datos (Clustering) agrupándolos por características afines no directamente identificables, permitirá separarlos en clases. Las técnicas de clustering buscan agrupar un conjunto de datos en un número de clústers. Este agrupamiento se basa en la idea de distancia entre datos y la cantidad de clústers depende de la técnica utilizada[18].

Por otro lado, en muchas ocasiones los datos tienen muchos atributos, por lo que es necesario reducir el número de estos. Entre estas técnicas de reducción las más destacadas son:

- Selección de Atributos, que se seleccionan directamente los atributos más representativos y descarta el resto.
- Reducción de Dimensiones: que a partir de nuestros atributos se crean nuevos, los cuales son combinaciones lineales de los originales.

En nuestro estudio vamos a comparar la eficiencia de estos 2 procesos. Respecto al segundo proceso, se analizará con la técnica PCA.

2.1. Análisis de Componentes Principales (PCA)

PCA es un método de análisis multivariante, su objetivo es encontrar las dimensiones con máxima varianza en los datos[19]. Primero encuentra el componente (vector propio) con mayor varianza, este es el componente principal, luego se encuentra un componente con la segunda mayor varianza entre todos, de nuevo se busca otro componente ortogonal a los primeros con la tercera mayor varianza, y así sucesivamente. Después de este proceso, es posible reducir el número de dimensiones con un sistema de ejes modificados, donde generalmente se eliminan los ejes de los datos originales que no generan mucha varianza. Además, como propone [20], se realizó un estandarizado de los datos extraídos para así uniformizarlos con una preparación óptima para PCA.

2.2. *k*-Means

En la actualidad es la técnica de clustering más utilizada debido a su simplicidad y eficiencia. Lo primero que se debe hacer en esta técnica es definir los k centroides al azar (uno para cada clúster), luego se toma cada punto de la base de datos y se sitúa en el que tiene el centroide más cercano (distancia euclídea). Luego, se recalcula el centroide de cada clúster (la media de todos los datos que lo componen, teniendo en cuenta que se quiere minimizar) y se vuelve a distribuir el centroide más cercano y así sucesivamente hasta que no haya más datos [21].

Teniendo en cuenta estos algoritmos, en este artículo se ha utilizado lenguaje de programación Python para el análisis y tratamiento de los datos y la librería *scikit-learn* para el desarrollo algorítmico.

3. ANÁLISIS Y TRATAMIENTO DE DATOS

Como se ha comentado el siguiente trabajo busca aplicar PCA y *k*-Means a la base de datos de seguridad vial. En esta base de datos podemos observar cuantificados los datos de los accidentes de tránsito ocurridos, por lo que a través de estas técnicas se evaluarán los accidentes de tránsito que van a ocurrir en un día determinado. Una vez utilizada esa información, se determinará la ruta óptima, para que el ciudadano pueda movilizarse de manera más segura con el sistema expuesto en [9]. Antes de poder procesar las técnicas algorítmicas, se estudiará como se encuentran estructurados el conjunto de datos.

	ID	TIPO	CATEGORIA	N°	CASO	DIA	FECHA	HORA	MODALIDAD	MEDIO	DIRECCION	SECTOR	SUBSECTOR
0	36090	APOYO A BOMBEROS	Accidente de Transito Choque	1	NaN	MIERCOLES	31/08/2016	20:10	Exceso de velocidad	NaN	PQ. PARQUE SAN MARTIN DE PORRES N SN, SAN ISIDRO	2.0	7.0
1	36091	APOYO A BOMBEROS	Accidente de Transito Choque	2	NaN	MIERCOLES	31/08/2016	19:05	Exceso de velocidad	NaN	AV. CAMINO REAL, SAN ISIDRO	2.0	5.0
2	36092	APOYO A BOMBEROS	Accidente de Transito Choque	3	NaN	MIERCOLES	31/08/2016	18:40	Exceso de velocidad	NaN	AV. PRADO OESTE, JAVIER, SAN ISIDRO	1.0	3.0
3	36093	APOYO A BOMBEROS	Accidente de Transito Choque	4	NaN	MIERCOLES	31/08/2016	17:01	Exceso de velocidad	NaN	AV. RIVERA NAVARRETE, RICARDO, SAN ISIDRO	4.0	2.0
4	36094	APOYO A BOMBEROS	Accidente de Transito Choque	5	NaN	MIERCOLES	31/08/2016	04:40	Exceso de velocidad	NaN	AV. PEREZ ARANIBAR, AUGUSTO, SAN ISIDRO	2.0	7.0
5	36095	APOYO A BOMBEROS	Accidente de Transito Choque	6	NaN	MIERCOLES	31/08/2016	03:39	Exceso de velocidad	NaN	AV. REPUBLICA, PASEO DE LA N SN, SAN ISIDRO	4.0	4.0
6	36096	APOYO A BOMBEROS	Accidente de Transito Choque	7	NaN	MARTES	30/08/2016	18:25	Exceso de velocidad	NaN	CA. UGARTE Y MOSCOSO, MANUEL SEBASTIAN N SN, S...	2.0	1.0
7	36097	APOYO A BOMBEROS	Accidente de Transito Choque	8	NaN	MARTES	30/08/2016	18:00	Exceso de velocidad	NaN	AV. MIRO QUESADA, AURELIO N SN, SAN ISIDRO	1.0	5.0
8	36098	APOYO A BOMBEROS	Accidente de Transito Choque	9	NaN	MARTES	30/08/2016	14:20	Exceso de velocidad	NaN	AV. PRADO OESTE, JAVIER N SN, SAN ISIDRO	1.0	4.0
9	36099	APOYO A BOMBEROS	Accidente de Transito Choque	10	NaN	MARTES	30/08/2016	13:40	Exceso de velocidad	NaN	CA. LAS TORDILLAS N 0176, SAN ISIDRO	5.0	1.0

Figura 2. Distribución de los datos.

3.1. Análisis de los datos

Lo principal es poder observar como están distribuidos los datos y qué diferentes características tienen. En este sentido, la Figura 2 muestra la estructura de la tabla "*APOYO_BOMBEROS*".

Bajo este contexto, y teniendo en cuenta que los datos que interesan para nuestro trabajo son los de seguridad vial, podemos observar en la Figura 3(a) como están distribuidos según el tipo de accidente; y en la Figura 3(b) según la modalidad. Por ejemplo, puede verse *Accidente de Transito Choque* como la clase principal de accidentes y *Exceso de velocidad* e *Imprudencia del conductor* como la modalidad.

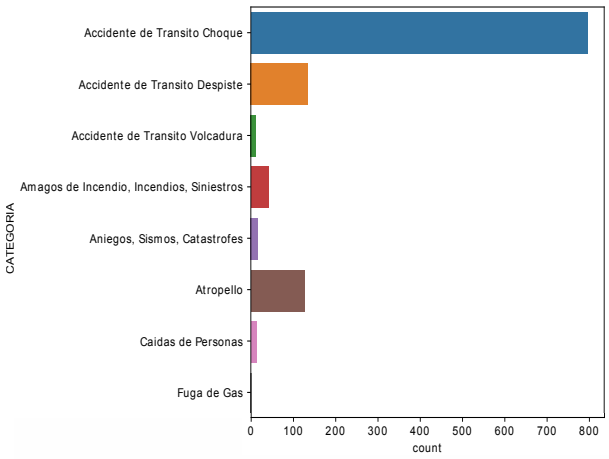
Por otro lado, podemos ver como es el comportamiento por sectores, subsectores o zonas del distrito. Asimismo, dónde ocurren más accidentes y cuándo ocurren estos accidentes, ver Figura 4.

Así, la Figura 4(a) muestra que el sector 1 es el más peligroso ya que tienen un número mayor de accidentes que los otros. Además, en los sectores $1-6$ y $3-1$, ver Figura 4(b), tienen mayor tasa de accidentes. Respecto al día de la semana, ver Figura 4(c), se tiene que los jueves son los días cuando ocurren más accidentes, y los domingos los que menos. Sin embargo, la varianza de estos datos es menor en comparación con las categorías de accidente, en donde *Accidente de Transito Choque* absorbe la gran mayoría de los datos.

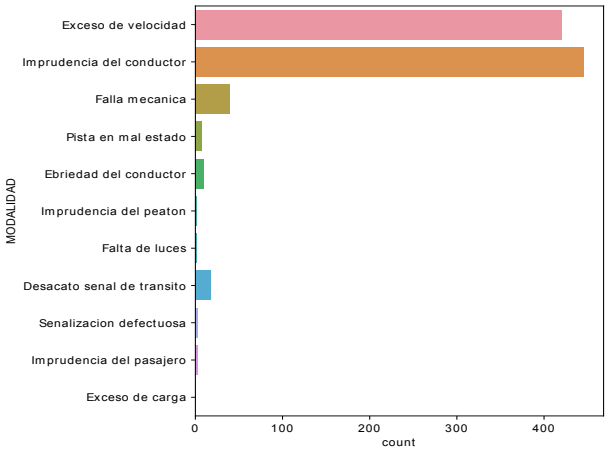
En resumen, se ha notado que hay una sobreabundancia de Categoría de Accidente, por *Accidente de Transito Choque* y modalidades, por *Exceso de velocidad* e *Imprudencia del conductor* a diferencia de los demás datos que son muy escasos, motivo principal por el que se evitará trabajar en base a estas columnas, además de tener que realizar un tratamiento previo.

3.2. Tratamiento de los datos

Como se puede observar en la Figura 2, se tienen datos faltantes (nulos) por columnas (variables). Además, la Figura 5(a) muestra resumidamente la distribución de estos tipos de datos por sectores. Por tanto, se mantendrán sólo las filas que tengan las columnas: *CATEGORIA*, *SECTOR* y *HORA* como no nulos, ya que no deseamos trabajar con datos que no indique dónde y a qué hora ocurrió el accidente. Así, la Figura 5(b) muestra las



(a) Distribución por tipo de accidente.



(b) Distribución por modalidad de accidente.

Figura 3. Distribución de los datos respecto al valor accidente.

columnas donde sí fueron permitidas los datos nulos y las que no.

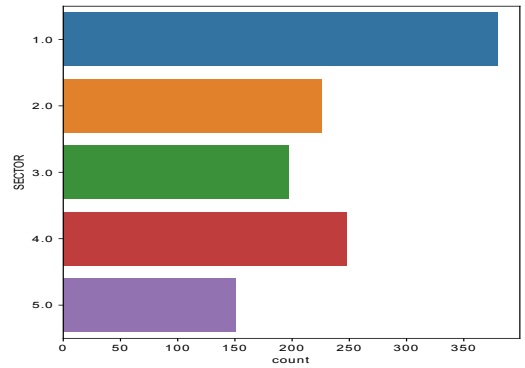
Seguidamente, se han realizado las siguientes operaciones (en este orden):

- Se eliminan las columnas innecesarias: *TIPO*, *N°*, *CASO*, *MEDIO*, *DIRECCION* e *ID*.
- Se convierte las columnas *SECTOR* y *SUBSECTOR* en tipo entero. Posteriormente, se crea la columna (variable) *SS* para comprimir *SECTOR* y *SUBSECTOR* en una sola columna, ver Figura 6(a), eliminando finalmente las columnas *SECTOR* Y *SUBSECTOR*.
- Se crea una columna T_M que es la hora convertida en minutos, es decir, se convierten las horas en enteros.

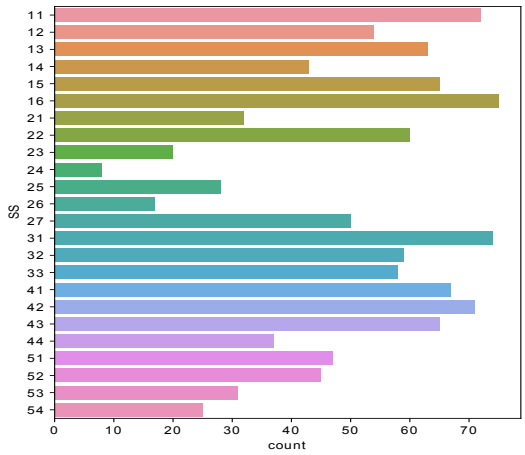
La Figura 6(b) muestra la tabla actualizada con las nuevas columnas después del tratamiento anterior.

3.3. Aplicación de Datos

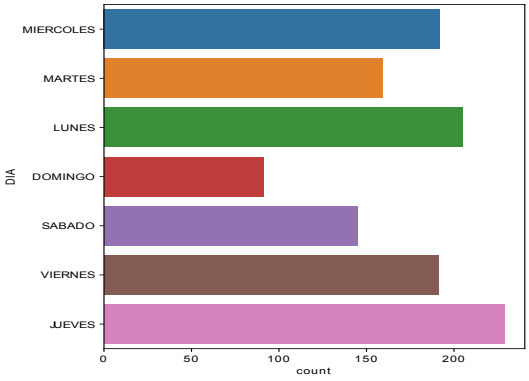
Respecto a la aplicabilidad de los datos, el presente trabajo realizó dos análisis: (i) en base a los días en los



(a) Por número de sector.



(b) Por número de subsector.



(c) Por día de la semana.

Figura 4. Distribución de datos.

```
ID              0
TIPO            0
CATEGORIA      36
N°              0
CASO         1106
DIA             0
FECHA           0
HORA            1
MODALIDAD     260
MEDIO        1027
DIRECCION       8
SECTOR         10
SUBSECTOR      10
dtype: int64
```

(a) Reconocimiento de datos nulos.

```
ID              0
TIPO            0
CATEGORIA       0
N°              0
CASO         1066
DIA             0
FECHA           0
HORA            0
MODALIDAD     223
MEDIO        1016
DIRECCION       0
SECTOR          0
SUBSECTOR       0
dtype: int64
```

(b) Eliminación de datos nulos.

Figura 5. Tratamiento de datos nulos.

	CATEGORIA	DIA	FECHA	HORA	MODALIDAD	SECTOR	SUBSECTOR	SS
0	Accidente de Transito Choque	MIERCOLES	31/08/2016	20:10	Exceso de velocidad	2	7	27
1	Accidente de Transito Choque	MIERCOLES	31/08/2016	19:05	Exceso de velocidad	2	5	25
2	Accidente de Transito Choque	MIERCOLES	31/08/2016	18:40	Exceso de velocidad	1	3	13
3	Accidente de Transito Choque	MIERCOLES	31/08/2016	17:01	Exceso de velocidad	4	2	42
4	Accidente de Transito Choque	MIERCOLES	31/08/2016	04:40	Exceso de velocidad	2	7	27
5	Accidente de Transito Choque	MIERCOLES	31/08/2016	03:39	Exceso de velocidad	4	4	44
6	Accidente de Transito Choque	MARTES	30/08/2016	18:25	Exceso de velocidad	2	1	21
7	Accidente de Transito Choque	MARTES	30/08/2016	18:00	Exceso de velocidad	1	5	15
8	Accidente de Transito Choque	MARTES	30/08/2016	14:20	Exceso de velocidad	1	4	14
9	Accidente de Transito Choque	MARTES	30/08/2016	13:40	Exceso de velocidad	5	1	51

(a) Compresión de sectores y subsectores.

	CATEGORIA	DIA	FECHA	HORA	MODALIDAD	SS	T_M
0	Accidente de Transito Choque	MIERCOLES	31/08/2016	20:10	Exceso de velocidad	27	1210
1	Accidente de Transito Choque	MIERCOLES	31/08/2016	19:05	Exceso de velocidad	25	1145
2	Accidente de Transito Choque	MIERCOLES	31/08/2016	18:40	Exceso de velocidad	13	1120
3	Accidente de Transito Choque	MIERCOLES	31/08/2016	17:01	Exceso de velocidad	42	1021
4	Accidente de Transito Choque	MIERCOLES	31/08/2016	04:40	Exceso de velocidad	27	280
5	Accidente de Transito Choque	MIERCOLES	31/08/2016	03:39	Exceso de velocidad	44	219
6	Accidente de Transito Choque	MARTES	30/08/2016	18:25	Exceso de velocidad	21	1105
7	Accidente de Transito Choque	MARTES	30/08/2016	18:00	Exceso de velocidad	15	1080
8	Accidente de Transito Choque	MARTES	30/08/2016	14:20	Exceso de velocidad	14	860
9	Accidente de Transito Choque	MARTES	30/08/2016	13:40	Exceso de velocidad	51	820

(b) Conversión de tiempo a minutos.

Figura 6. Tratamiento general de los datos.

que ocurrieron accidentes para predecirlos en un día y visualizar qué fechas tienen rasgos parecidos; y (ii) en base a los sectores en donde ocurrieron los accidentes, para predecir las características de un nuevo sector y visualizar qué sectores son más parecidos. Por tanto, en la agrupación por fechas se obtienen las siguientes columnas:

- *total_sectores*: El total de sectores en donde ocurrió un accidente en ese día.
- *total_unique_accidentes*: El total de tipos de accidentes.
- *total_accidentes*: Total de accidentes ocurridos en un día.
- *total_modalidades*: Total de modalidades de accidentes.

La Figura 7(a) muestra la tabla de los datos agrupados por fechas. Análogamente, en la agrupación por sectores colocamos las columnas:

- *total_dias*: El total de días en que hubo accidentes en ese sector.
- *total_unique_accidentes*: Total de tipos de accidente.
- *total_accidentes*: Total de accidentes en ese sector.
- *total_modalidades*: Total de modalidades de accidentes.

La Figura 7(b) muestra los datos agrupados por sectores.

Una vez realizado el análisis y tratamiento de datos y seleccionadas las variables que se van a evaluar, en la siguiente sección se procederá a discutir los resultados obtenidos de las técnicas algorítmicas.

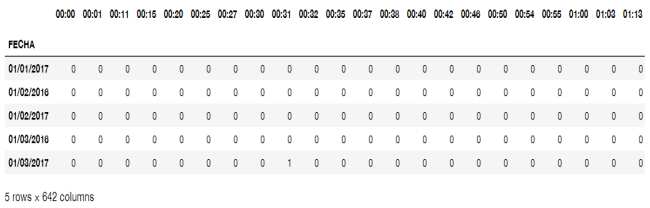
FECHA	total_sectores	total_unique_accidentes	total_accidentes	total_modalidades
01/01/2017	1	1	1	1
01/02/2016	1	1	1	1
01/02/2017	2	1	2	1
01/03/2016	3	2	4	1
01/03/2017	2	2	2	1

(a) Agrupación por fechas.

SS	total_dias	total_unique_accidentes	total_accidentes	total_modalidades
11	64	7	72	4
12	52	7	54	4
13	59	7	63	4
14	37	5	43	6
15	52	5	65	6

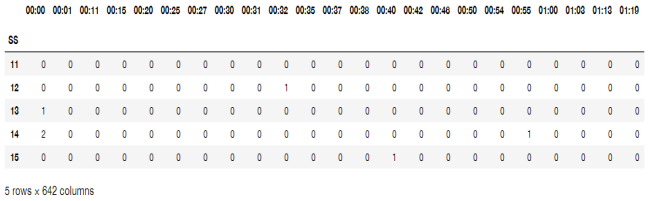
(b) Agrupación por sectores.

Figura 7. Agrupación de Datos.



5 rows x 642 columns

(a) Agrupación por fechas.



5 rows x 642 columns

(b) Agrupación por sectores.

Figura 8. Agrupación de Datos después de la reducción de variables.

4. MINERÍA DE DATOS

En esta sección se muestran los resultados después de utilizar los algoritmos PCA y *k*-Means, analizando los resultados obtenidos de las agrupaciones seleccionadas, es decir, fechas y sectores.

4.1. Análisis de Componentes Principales

Como se indicó en la Sección 2, existen dos procesos para reducir la complejidad de las variables: Selección de Atributos y Reducción de Dimensiones. El primero es directo, mientras que para el segundo se utilizará PCA.

Reducción de variables

Antes de analizar PCA, se añadirán las horas y cantidad en las que ocurrió un accidente con la finalidad de tener una mejor valoración de cada día en el que ocurrió el evento. Luego, se transforma la variable categórica *HORAS* en variable indicador, es decir, cada una de las horas que tenía un accidente es ahora una variable. Posteriormente se agrupará por fechas, ver Figura 8(a), y sectores, ver Figura 8(b).

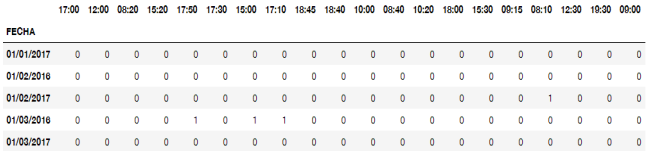
Finalmente, de esta forma se puede visualizar la cantidad de veces en que ocurrió un accidente a determinada hora. Por ejemplo, la Figura 9 muestra las horas en que ocurrieron mayor cantidad de accidentes.

```
17:00     6
12:00     6
08:20     6
15:20     6
17:50     6
17:30     6
15:00     6
17:10     6
18:45     6
18:40     7
10:00     7
08:40     7
10:20     7
18:00     7
15:30     7
09:15     8
08:10     8
12:30     9
19:30    10
09:00    11
dtype: int64
```

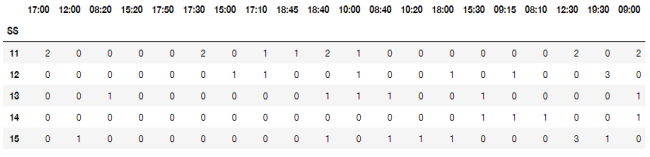
Figura 9. Ejemplo de cantidad de accidentes por hora.

Selección de atributos

Este proceso selecciona directamente los atributos que identificamos como relevantes. En nuestro caso de estudio, seleccionamos las columnas en donde hay más variación, es decir, las 20 horas en donde ocurren accidentes con más frecuencia y eliminamos el resto. La Figura 10(a) muestra los datos agrupados por fechas con las 20 horas de mayor accidente; y la Figura 10(b) muestra los datos agrupados por sectores con las 20 horas de mayor accidente.



(a) Por fechas.



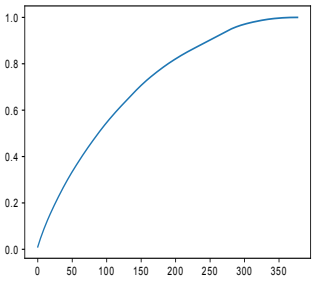
(b) Por sectores.

Figura 10. Resultado de Selección de Atributos.

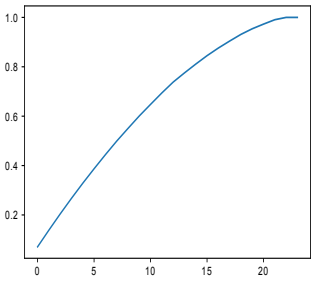
Reducción de dimensiones

Para esta técnica existen varios algoritmos, pero el más conocido es PCA. Antes de aplicar PCA, se realiza la estandarización de los datos[20]. Una vez estandarizados, se inicializa PCA y se ajusta a nuestros datos estandarizados, en este caso se utilizó los parámetros por defecto que tiene PCA en *scikit-learn*.

Finalmente, se visualiza la varianza acumulada para una mejor visualización de cómo cambia esta acorde a la cantidad de componentes, es decir, determinar el porcentaje de variación de cada nueva columna (variable) y realizando una suma acumulada. Las Figuras 11(a) y 11(b) muestran la varianza explicada acumulada tanto en los datos agrupados por fecha, como por sector, respectivamente.



(a) En agrupación por fecha



(b) En agrupación por sector

Figura 11. Varianza explicada acumulada para PCA.

En este sentido, se puede observar la cantidad de componentes con la cual existe al menos un $0{,}75$ de varianza acumulada. Se observa que en el caso de la agrupación por fechas, ver Figura 11(a), se necesitaron 168 componentes para conservar $0{,}75$ de la varianza acumulada, mientras que en la agrupación por sectores con 14 componentes se tuvo $0{,}77$ de la varianza acumulada. Por tanto, en el primero se observa un total de componentes de más de 350, mientras que en el segundo sólo son 24 componentes.

En este contexto, se realiza el algoritmo PCA conservando la cantidad designada para cada tipo de agrupación; número de componentes de 168 y 14. Es decir, se hace una reducción de dimensiones a 168 para, posteriormente, ajustar y transformar estos datos. La Figura 12(a) y 12(b) muestran los resultados de realizar la transformación PCA en las agrupaciones por fecha y por sector. El análisis de este resultado se detallará en la siguiente sección.

4.2. *k*-Means

En esta sección se discutirán los resultados de aplicar *k*-Means en los datos, procesados anteriormente. Por tanto, se aplicará este algoritmo para los datos originales, los transformados por el proceso de Selección de Atributos y los transformados por PCA, tanto para las agrupaciones por *Fecha* como por *Sector*.

Para esta finalidad, se utilizó la función $KMeans(n\_clusters = 2, random\_state = 123)$, donde

FECHA	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14
01/01/2017	-0.185373	-0.030123	-0.215046	-0.120995	-0.155666	-0.148969	-0.047481	0.015802	-0.119778	-0.227261	-0.148393	0.063500	-0.175897	-0.014960
01/02/2016	-0.183372	-0.027750	-0.216347	-0.127582	-0.156601	-0.153374	-0.045531	0.016483	-0.121555	-0.230690	-0.140312	0.057175	-0.173489	-0.024874
01/02/2017	-0.256609	-0.125561	-0.098078	-0.139386	-0.198800	-0.262299	-0.126831	-0.011069	-0.168746	-0.425438	-0.075351	0.120208	-0.226991	0.012228
01/03/2016	-0.124437	-0.138675	-0.405694	-0.366507	0.710038	-0.573535	-0.176042	-0.031682	-0.254895	-0.526039	-0.280722	0.188492	-0.502061	-0.233776
01/03/2017	-0.233712	-0.048507	-0.268741	-0.185771	-0.145784	-0.273688	-0.066232	0.006349	-0.230997	-0.395269	0.094980	0.080724	-0.112467	0.043606

5 rows × 168 columns

(a) Resultado en agrupación por fechas.

SS	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14
11	-1.102621	-2.693674	1.674671	22.668650	-9.852736	8.595804	11.267424	4.076612	4.207332	1.298908	-0.741488	1.404219	-4.118827	0.780747
12	-3.447970	-1.643476	-0.363519	-1.899406	2.985015	7.690558	-3.731925	-4.350842	-15.277197	-4.601964	-10.006165	-1.533503	-13.596376	4.529677
13	-2.707412	1.587541	-2.448790	-2.136521	19.592803	-8.092382	18.101631	4.930093	2.118480	2.445121	-1.512520	-2.937995	-4.094557	0.750564
14	-0.869588	0.396469	-0.117712	2.951985	-0.311525	-0.833402	0.717539	-0.199250	-1.043812	-2.161955	-2.286056	-3.501418	5.474675	-8.591356
15	-3.411989	-4.267176	-3.157723	-0.026558	0.343737	-6.462478	-9.421521	-3.155382	11.007997	15.627261	-8.054042	8.921811	-6.730038	-0.644347

(b) Resultado en agrupación por sectores.

Figura 12. Resultado de PCA.

$n\_clusters = 2$ es debido a que solo se desea 2 grupos de datos: peligrosos y seguros. Además, cuando se intentó con mayor número de clústers se obtuvo clústers con 1 elemento, es decir, se comprobó que el k óptimo para el trabajo realizado era 2. Luego, el parámetro $random\_state = 123$ se utiliza para que se queden almacenados los números aleatorios generados, y se utilicen los mismos en otros tipos de agrupamiento. Los demás hiperparámetros de *k*-Means fueron establecidos por defecto.

La Figura 13 muestra todo este proceso descrito.

Caso 1: Datos sin modificar

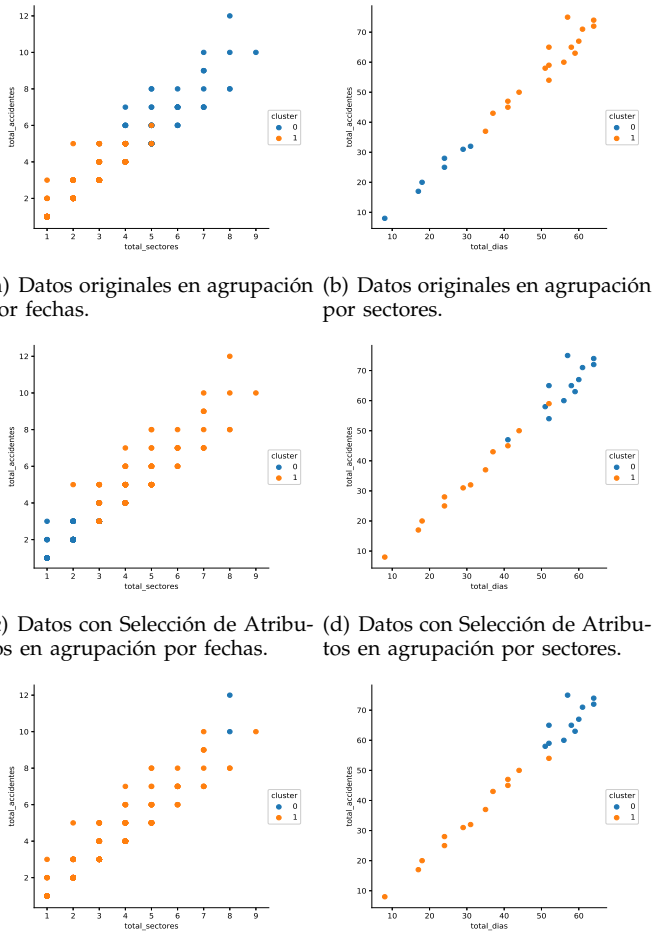
En la Figura 13(a) se obtuvieron bastantes fechas peligrosas, en total fueron 141 fechas, las cuáles fueron: el 28 de Abril, el 28 de Junio, el 28 de Octubre, el 29 de Enero, el 29 de Abril, el 29 de Setiembre, el 29 de Octubre, el 30 de Junio, el 31 de Agosto y el 31 de Octubre. Por otro lado, la Figura 13(b) muestra los sectores peligrosos, en total fueron 17 sectores: $1-1, 1-2, 1-3, 1-4, 1-5, 1-6, 2-2, 2-7, 3-1, 3-2, 3-3, 4-1, 4-2, 4-3, 4-4, 5-1$ y $5-2$.

Caso 2: Datos con Selección de atributos

En este caso, la Figura 13(c) se muestra menos fechas peligrosas que con los datos originales, en total fueron 127 fechas peligrosas, entre ellas están: el 22 de Enero, el 22 de Abril, el 22 de Agosto, el 23 de Febrero, el 23 de Junio, el 24 de Abril, el 24 de Agosto, el 24 de octubre, el 25 de Enero y el 26 de Febrero. Respecto a los sectores, la Figura 13(d) muestra menos sectores peligrosos que con los datos originales, en total fueron 12: $1-1, 1-2, 1-3, 1-5, 1-6, 2-2, 3-1, 3-3, 4-1, 4-2, 4-3$ y $5-1$.

Caso 3: Datos transformados con PCA

Finalmente, con PCA, ver Figura 13(e), se obtuvieron una cantidad sumamente inferior para las fechas peligrosas que con los datos originales y con los datos modificados con selección de atributos, en total fueron 2: el 18 de Julio y el 21 de Enero. Respecto a los sectores, ver Figura 13(f), se obtuvieron una cantidad inferior que con los datos



(a) Datos originales en agrupación por fechas.

(b) Datos originales en agrupación por sectores.

(c) Datos con Selección de Atributos en agrupación por fechas.

(d) Datos con Selección de Atributos en agrupación por sectores.

(e) Datos transformados con PCA en agrupación por fechas.

(f) Datos transformados con PCA en agrupación por sectores.

Figura 13. Resultado de *k*-Means.

anteriores, en total fueron 11: $1-1, 1-3, 1-5, 1-6, 2-2, 3-1, 3-2, 3-3, 4-1, 4-2$ y $4-3$.

4.3. Análisis de los resultados

Una vez obtenidos los resultados derivados de PCA y *k*-Means, en esta sección se va a discutir los resultados obtenidos. Para ello realizaremos un análisis de los resultados por sectores y fechas para, posteriormente, evaluar qué tratamiento de datos es el óptimo en este tipo de problemas.

Análisis por Sectores y fechas

Por un lado, en las gráficas de los datos agrupados por fechas, ver Figuras 13, se obtiene la percepción que se han agrupado más fechas peligrosas. Sin embargo, al examinar la cantidad de datos de cada clúster, observamos que son menor cantidad. La razón por la que hay pocos datos dentro del clúster de las fechas seguras se justifica porque muchas fechas repiten la cantidad de sectores en los que hubieron accidentes y el total de accidentes, aunque en realidad la cantidad de fechas seguras es mayor.

Por otro lado, en las gráficas de los datos agrupados por sectores los datos no se sobreponen, esto se debe a que son sólo 24 datos (sectores) y en el eje X se encuentran el total de días en vez del total de sectores. El total de días, puede variar de 1 a más de 60, mientras que el total de sectores solo varía de 1 a 9 (es el máximo total de sectores en donde ocurrieron accidentes el mismo día).

En este último caso, se puede ver que tanto en los datos originales como en los que se realizaron selección de atributos, más de la mitad son sectores peligrosos. Sin embargo, en el caso de los datos con transformación PCA se obtuvo un clúster más refinado, con los 11 sectores más peligrosos. En este contexto, la Figura 14 se muestra el mapa de San Isidro con los sectores más peligrosos señalizados.

Igualmente en el caso de los datos agrupados por fecha, se prefiere el resultado obtenido por los datos transformados por PCA, ya que muestran con claridad las dos fechas más peligrosas que se tiene registradas: 21 de Enero y 18 de Julio del año 2016.

Análisis de los clústers

Utilizando la función $adjusted\_rand\_score$ de *scikit-learn* podemos mostrar el parecido que hay entre los datos de un conjunto y de otro. Con esta función observamos el resultado mostrado en la Tabla 1. En este caso podemos ver que si el resultado está más cerca este del 0 indica que sus clústers son aleatoriamente independientes y, por tanto, es bueno ya que justamente se quiere que cada clúster tenga sus propias características. Sin embargo, más a 1 indica que los clústers son casi idénticos. La Tabla 1 muestra los resultados tanto en el que se agruparon por fechas como en el de sectores.

Similaridad entre	Por fechas	Por sectores
Originales y Sel. de Atr.	0.4664	0.3157
Originales y PCA	0.0340	0.2213

Tabla 1
Similaridad entre los Datos Originales, con Selección de Atributos y transformados con PCA.

En esta Tabla se puede apreciar que en la agrupación por fechas, la similaridad entre los datos originales y los datos transformados por PCA es cercano a cero; mientras que la similaridad entre los datos originales y los datos con selección de atributos es casi la mitad. Esto significa que los clústers del PCA tienen más independencia que los clústers de la selección de atributos.

Por otro lado, en la agrupación por sectores, la similaridad entre los datos originales y los datos transformados por PCA es más de 0,2. Esto quiere decir que los clústers no son independientes, se puede apreciar cuando se mostró la cantidad de sectores peligrosos en cada uno, donde la diferencia era sólo de uno. En este caso, es casi lo mismo que utilizar selección de atributos en donde el resultado fue un poco más de 0,3. Sin

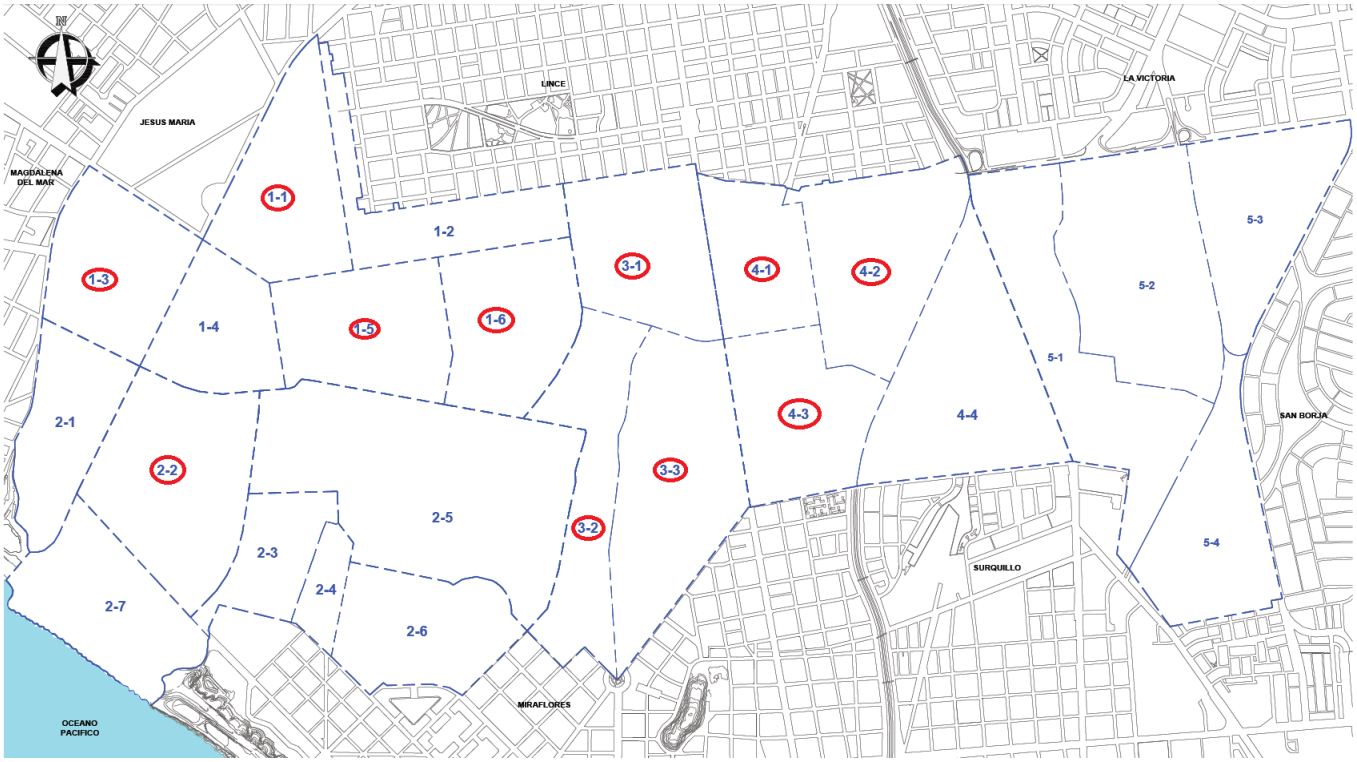


Figura 14. Mapa de San Isidro con los sectores más peligrosos señalados

embargo, se sigue prefiriendo la transformación PCA por ser más óptimo.

5. Conclusiones y trabajo futuro

Como se ha visto a lo largo de este trabajo el análisis y tratamiento de datos son una herramienta fundamental a la hora de poder realizar técnicas de minería de datos para sacar nuestras propias conclusiones. Este trabajo se ha evaluado junto a [9] para mejorar la aplicación móvil y así poder identificar mejor sectores y fechas peligrosas a la hora de crear rutas óptimas a los ciudadanos.

En cuanto a las técnicas utilizadas en este trabajo hemos podido observar que es de suma importancia realizar la estandarización de los datos antes de aplicar PCA. En cuanto a PCA, una mayor cantidad de datos ayudaría a mejorar los resultados en los clústers, teniendo mayor independencia entre ellos.

Sin embargo, en situaciones en donde no se cuente con muchos datos, como el caso donde se agruparon por sectores, es casi igual de efectivo utilizar el proceso de selección de atributos que PCA. De lo anterior se extrapola que la selección de atributos es efectivo cuando se tienen muy pocos datos.

En cuanto al algoritmo k-Means, fue efectivo utilizar $k = 2$ ya que se debía dividir en 2 clases de grupos: peligrosos y seguros. Sin embargo, no se puede generalizar que este k siempre se deba utilizar. Es necesario realizar pruebas con distintas cantidad de centroides.

Comparando estos resultados con [9], podemos ver como los resultados obtenidos con los ULAs fueron más significativos que los Supervisados, ya que se obtuvo una mejor visión del comportamiento de los datos.

Agradecimientos

Este trabajo ha sido parcialmente financiado por "Cienciactiva – CONCYTEC" del gobierno peruano, bajo el número de proyecto 128-2015-FONDECYT y por el "Programa Nacional de Innovación para la Competitividad y Productividad, Innóvate - Perúçon número de contrato FINCyT 363-PNICP-PIAP-2014.

Referencias

[1] D. López-de Ipiña, L. Chen, A. Jara, E. Mannens, and Y. Li, "Internet of things, linked data, and citizen participation as enablers of smarter cities," 2016.

[2] I. A. T. Hashem, V. Chang, N. B. Anuar, K. Adewole, I. Yaqoob, A. Gani, E. Ahmed, and H. Chiroma, "The role of big data in smart city," *International Journal of Information Management*, vol. 36, no. 5, pp. 748 – 758, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0268401216302778

[3] M. M. Rathore, A. Ahmad, A. Paul, and S. Rho, "Urban planning and building smart cities based on the internet of things using big data analytics," *Computer Networks*, vol. 101, pp. 63 – 80, 2016, industrial Technologies and Applications for the Internet of Things.

[4] I. A. T. Hashem, V. Chang, N. B. Anuar, K. Adewole, I. Yaqoob, A. Gani, E. Ahmed, and H. Chiroma, "The role of big data in smart city," *International Journal of Information Management*, vol. 36, no. 5, pp. 748–758, 2016.

[5] M. Castillo-Cara, E. Huaranga-Junco, M. Quispe-Montesinos, L. Orozco-Barbosa, and E. A. Antúnez, "Frog: A robust and green wireless sensor node for fog computing platforms," *Journal of Sensors*, vol. 2018, 2018.

[6] S. Yi, C. Li, and Q. Li, "A survey of fog computing: Concepts, applications and issues," in *Proceedings of the 2015 Workshop on Mobile Big Data*, ser. Mobidata '15. ACM, 2015, pp. 37–42. [Online]. Available: http://doi.acm.org/10.1145/2757384.2757397

[7] B. Tang, Z. Chen, G. Hefferman, T. Wei, H. He, and Q. Yang, "A hierarchical distributed fog computing architecture for big data sis in smart cities," in *Proceedings of the ASE BigData & SocialInformatics 2015*, ser. ASE BD&SI '15. ACM, 2015, pp. 28:1–28:6. [Online]. Available: http://doi.acm.org/10.1145/2818869.2818898

[8] F. Camastra, A. Ciaramella, and A. Staiano, "Machine learning and soft computing for ict security: an overview of current trends," *Journal of Ambient Intelligence and Humanized Computing*, vol. 4, no. 2, pp. 235–247, Apr 2013.

[9] G. B. Rocca, M. Castillo-Cara, R. A. Levano, J. V. Herrera, and L. Orozco-Barbosa, "Citizen security using machine learning algorithms through open data," in *2016 8th IEEE Latin-American Conference on Communications (LATINCOM)*, Nov 2016, pp. 1–6.

[10] A. V. Calcines *et al.*, "Algoritmos de aprendizaje automático: aplicación en la solución a problemas medioambientales," *Cuadernos de Educación y Desarrollo*, no. 49, 2014.

[11] Y. R. Sarabia, X. C. Bermúdez, R. F. Martinez, Z. H. Rodríguez, A. M. C. Moya, and M. M. G. Lorenzo, "Cbr-ann hybrid model to optimize the sequence of wastewater treatments." in *ITEE*, 2005, pp. 711–720.

[12] J. P. Theiler and G. Gisler, "Contiguity-enhanced k-means clustering algorithm for unsupervised multispectral image segmentation," in *Algorithms, Devices, and Systems for Optical Information Processing*, vol. 3159. International Society for Optics and Photonics, 1997, pp. 108–119.

[13] K. Krishna, K. Ramakrishnan, and M. Thathachar, "Vector quantization using genetic k-means algorithm for image compression," in *Information, Communications and Signal Processing, 1997. ICICS., Proceedings of 1997 International Conference on*, vol. 3. IEEE, 1997, pp. 1585–1587.

[14] O. G. D. Abiertos, "Municipalidad de san isidro." http://datosabiertos.msi.gob.pe/home/, accessed: 24/01/2018.

[15] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[16] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 881–892, 2002.

[17] J. Tuya, I. R. Román, and J. J. D. Cosín, *Técnicas cuantitativas para la gestión en la ingeniería del software*. NetBiblo, 2007.

[18] I. Benítez, "Técnicas de agrupamiento para el análisis de datos cuantitativos y cualitativos," *Universidad Politécnica de Valencia*, 2005.

[19] S. Muñoz Romero, "Análisis multivariante: soluciones eficientes e interpretables," 2015.

[20] S. Muñoz Armayones, "Técnicas multivariantes para el análisis de datos ómicos," 2016.

[21] D. Pascual, F. Pla, and S. Sánchez, "Algoritmos de agrupamiento," *Método Informáticos Avanzados*, pp. 164–174, 2007.