ISSN: 0798-1015

REVISTA ESPACIOS

Vol. 41 (41) 2020 • Art. 15

Recibido/Received: 17/07/2020 • **Aprobado**/Approved: 29/09/2020 • **Publicado**/Published: 29/10/2020

Cuantificación de la contaminación por dióxido de carbono producida por empresas courier usando minería de datos: una mirada a las provincias de Lima y Constitucional del Callao (Colombia)

Quantification of Carbon Dioxide Pollution produced by Courier Companies using Data Mining: A look at the Province of Lima and the Constitutional Province of Callao

GARCIA-OJEDA, Juan C.[1]
ALVITES, John A.[2]
PUELLO, Plinio [3]

Resumen
Este trabajo describe un método, basado en minería de datos, que determina el impacto ambiental de una empresa Courier en la provincia de Lima y la provincia Constitucional del Callao. El método permite: (i) determinar la frecuencia de despachos y clientes, (ii) identificar los distritos con mayor concentración de despachos; y, (iii) calcular la contaminación total producida por las diferentes operaciones de despacho. Para el caso de estudio, las emisiones de CO_2 son aproximadamente de 10,5 toneladas.
Palabras clave: contaminación ambiental, tráfico vehicular, minería de datos.

Abstract
This work describes a method, based on data mining, that determines the environmental impact of a courier company in the Province of Lima and the Constitutional Province of Callao. The method allows: (i) to determine the frequency of deliveries, (ii) to identify the districts with the highest concentration of deliveries; and, (iii) to calculate the total pollution produced by the different delivery operations. For the case study, the emissions of CO_2 is approximately 10.5 tons.
Keywords: environmental pollution, vehicular traffic, data mining

1. Introducción

La situación ambiental de Perú en el periodo comprendido entre 1980 y 2000 no fue nada alentadora, sobre todo en su ciudad capital: Lima. Una de las causas fue la desmedida entrada de vehículos particulares y la ausencia de políticas públicas de transporte para ofrecer servicios de movilidad a un creciente número de habitantes en el país (Martínez Espinal, 2017). Sin embargo, el aumento del parque automotor, con la finalidad de ofrecer transporte público, era la solución para llevar el sustento diario a miles de hogares que vivían en un país con una fuerte crisis económica: Hiperinflación de casi 7600% (Martínez Espinal, 2017; Tavera, 2001). Crisis que fue

[1] Docente, Programa de Ingeniería de Sistemas, Facultad de Ingeniería, Universidad de Cartagena. Colombia. Email: jcgarciao@unicartagena.edu.co
[2] Alumno, Carrera de Ingeniería de Sistemas, Facultad de Ingeniería y Arquitectura, Universidad de Lima. Peru. Email: johnalvitescastillo@gmail.com
[3] Docente, Programa de Ingeniería de Sistemas, Facultad de Ingeniería, Universidad de Cartagena. Colombia. Email: ppuellom@unicartagena.edu.co

manejada entre los años 1990-2000, cuando se privatizó un gran número de empresas y se implementaron medidas económicas para controlar tal crisis (Tavera, 2001). Conforme el panorama económico fue mejorando en Perú, aún persistía el problema vehicular. Lo cual continuó ocasionando gran congestión en zonas pequeñas de Lima y trajo un incremento de Dióxido de Carbono, i.e., CO_2, que afectó la salud de muchas personas dentro de Lima y, también, Callao (Gonzales et al., 2014; Martínez Espinal, 2017). Lo anterior, durante el final del gobierno del presidente Fujimori e inicios de la década del 2000, desencadenó en regulaciones y planes que tenían como objetivo la reducción de los vehículos públicos, la reducción de ciertas sustancias contaminantes y el control progresivo de algunas zonas en la provincia de Lima, como Puente Piedra, el Centro de Lima, y en la provincia constitucional del Callao; no obstante, ninguna tuvo éxito en reducir de manera considerable estos problemas (Chavhan & Venkataram, 2019; Romero et al., 2020; Tavera, 2001).

Así mismo, otros negocios basados en servicios de transporte, e.g., mensajería, habían empezado a establecerse en Lima, lo cual contribuyó al incremento desmesurado de vehículos tanto públicos (dados de baja, vida útil alcanzada) como privados (Martínez Espinal, 2017). En Perú, los vehículos privados (e.g., autos, camiones, motos, entre otros), luego de registrados a una empresa *Courier*, son equipados y utilizados para realizar tareas de carga y descarga de encomiendas; tanto a clientes como a la misma empresa. En cada recorrido, estos vehículos arrojan al ambiente cantidades importantes de CO_2 que contribuyen a la contaminación ambiental. Esto último, de interés para el Perú en conexión con el ODS (Objetivo de Desarrollo Sostenible) N°13 de la ONU respecto al cambio climático y como este afecta al ecosistema humano (personas y comunidades) y su impacto negativo en la economía (*Objetivos y metas de desarrollo sostenible – Desarrollo Sostenible*, 2015). Empero, muchas empresas *Courier*, en Perú, desconocen el nivel de contaminación de sus operaciones (i.e., despacho de las encomiendas); sin embargo, la información de interés subyace en las bases de datos que almacenan la información transaccional de cada operación.

Es importante indicar que la contaminación ambiental puede ser vista como una combinación de sustancias químicas con el aire y la humedad del lugar que afectan al clima del entorno, y que afecta, en el tiempo, la salud de las personas (Gonzales et al., 2014; Romero et al., 2020). En Perú, uno de los actores principales que colabora a la contaminación ambiental es el transporte público que contribuye con aproximadamente 40% del total de emisiones nacionales de energía (Romero et al., 2020). De igual forma un 27% del total de vehículos despachados excede los 20 años de uso; asimismo se encontró, que por año, se registran 150.000 a 200.000 nuevos vehículos (Romero et al., 2020). Como consecuencia de las unidades antiguas presentes y del aumento de vehiculos, los emboltellamientos, o congestión vehicular, aumentaron considerablemente en la provincia de Lima y en la provincia constitucional del Callao (Chavhan & Venkataram, 2019; Romero et al., 2020). Lo anterior reviste importancia, porque en un área metropolitana la congestión vehicular “es un evento inaguantable para viajeros y presenta una pérdida de tiempo y dinero en forma de gasolina”, además de producir estrés, problemas respiratorios y problemas crónicos de salud den las personas (Chavhan et al.,2019; Gonzales et al., 2014). Entre las principales sustancias que contribuyen a la polución está el Plomo en las gasolinas, el Dióxido de Carbono (CO_2), el Dióxido de Azufre, los Gases de Efecto Invernadero, Material Particulado tipo PM2.5 y PM10 y el Dióxido de Nitrógeno (Gonzales et al., 2014; Romero et al., 2020).

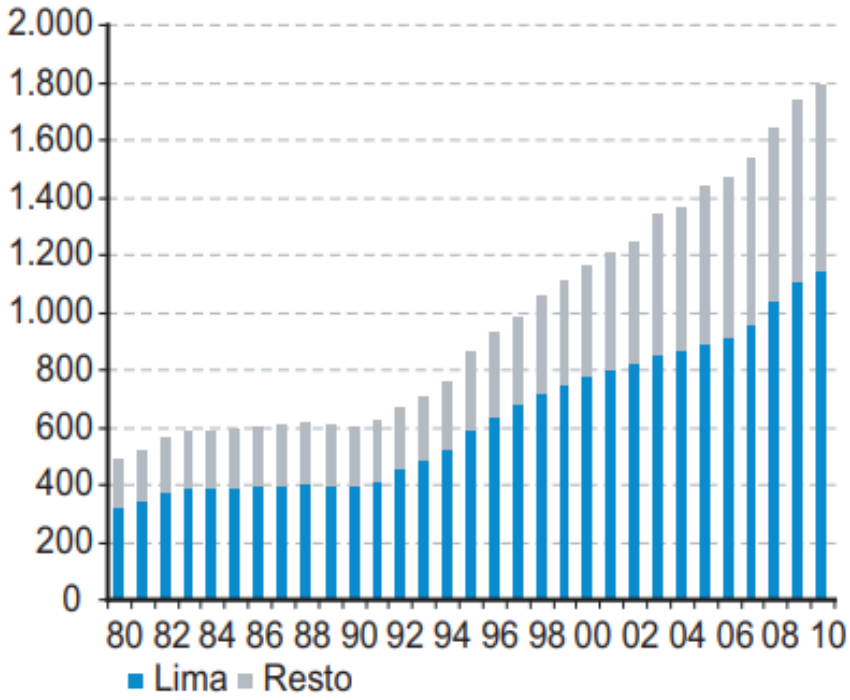
1.1. Punto de Partida

En Perú, la contaminación del aire se hace notable entre los años 1980-2000. En gran parte por la entrada de vehículos particulares y la falta de políticas respecto al transporte público (Martínez Espinal, 2017). Sin embargo es importante mencionar que en la década de los 80’s no era común debido a los precios elevados que tenían los autos, mantenimiento e importación, además de la poca popularidad (BBVAResearch, 2010). El mercado automotriz, según el informe del grupo BBVA Research (2010), fue creciendo exponencialmente después de las reformas del gobierno de Fujimori (1990-2000), otorgando facilidades a la importación de vehículos nuevos y

usados por igual (vér Gráfica 1). Estas facilidades dieron píe para que las personas, dueños de vehículos particulares y ante las dificultades económicas de la época, prestaran servicios de transporte. Lo anterior dió lugar a la expedición de muchas regulaciones y derogaciones de las mismas en beneficio de la economía. Originando, estos desacuerdos, el aumento del parque automotor tanto particular como público; lo cual desencadenó en un aumento de la congestión vehicular, por ende la polución ambiental durante los últimos tres lustros (Martínez Espinal, 2017).

Después del gobierno de Fujimori (2000-2014), con la consolidación de la economía, el número de vehículos siguió aumentando significatimente en Lima; de igual forma la contaminación hasta un 65% (Romero et al., 2020). En respuesta al creciente problema, el Ministerio de Transportes y Comunicaciones presentó el Plan Maestro de Transporte Urbano cuyo propósito es reducir el número de transporte público para reducir la tasa de congestión en 20 años (Yachiyo Engineering Co., 2005). Sin embargo, de acuerdo con el análisis realizado por Martínez Espinal (2017) al Plan Maestro de Transporte Urbano, se ha determinado de que, si no se toman medidas específicas como el cambio a combustibles más limpios o la eliminación de buses anticuados, la contaminación ambiental seguiría aumentando de manera considerable. La mayoría de análisis para reducir el índice de $KgCO_2$ toman como base el transporte público. Sin embargo, es importante recordar que el transporte privado aporta un número significativo de unidades de transporte (Yachiyo Engineering Co., 2005). Estos elementos son importantes porque tanto las empresas de transporte y distribución públicas y privadas deben respetar una serie de reglas específicas para cada tipo de vehículo implementadas en el 2012. Estas reglas no solo aplican para las empresas, sino también todas las municipalidades del Perú (Torres Trujillo, 2012).

Gráfica 1
Crecimiento del número de autos durante
30 años (1980-2010) en Perú, en miles de unidades



Fuente: BBVA Research

De acuerdo con un estudio realizado por la Revista Peruana de Medicina Experimental y Salud Pública (Tapia et al., 2018), el 80% de la contaminación del aire pertenece a emanaciones de los vehículos públicos y privados. Para contrarrestar ese 80% presente, como se había mencionado anteriormente, se planificaron planes de ordenamiento que complementaron el Plan Maestro de Transporte Urbano (Yachiyo Engineering Co., 2005) y

posteriormente la Regulación de Transporte en el Perú (Torres Trujillo, 2012). En primer lugar, tenemos la reducción o eliminación de ciertas sustancias dentro de los distintos tipos de combustibles, sin embargo, se reportó en 2014 por parte de la OMS (Organización Mundial de la Salud) que Lima era uno de los países con mayor contaminación en América Latina lo cual reflejó que esta estrategia no funcionó (Tapia et al., 2018). En segundo lugar, tenemos el plan de ordenamiento del tráfico vehicular de la provincia de Lima y la provincia constitucional del Callao cuyo propósito es "ordenar la circulación de vehículos en función de la estructura urbana para mejorar la fluidez del tránsito vehicular" (Tapia et al., 2018). Este segundo plan propuesto ayudó a reducir de manera considerable cuatro contaminantes ambientales: Material Particulado 2.5, Material Particulado 10, Dióxido de Azufre y Dióxido de Nitrógeno; no obstante, no soluciona de manera directa el problema de la contaminación. Es decir, sólo reduce el índice respectivo (Tapia et al., 2018). En tercer lugar, las Municipalidades de Lima y Callao implementaron rutas privadas que solo carros, taxis y/o un cierto tipo de vehículo pueden acceder, en esta sección entran los llamados alimentadores (Ministerio de Transportes y Comunicaciones, 2016). Con la llegada de los alimentadores, se redujo la demanda de transporte público considerablemente, aunque también provocó la abundancia de vehículos en una ruta específica (Ministerio de Transportes y Comunicaciones, 2016). Todos estos estudios han tenido en consideración los transportes públicos, las rutas designadas en Lima de acuerdo a su estructura urbana, los tipos de sustancias que contribuyen a la contaminación y la población presente en Lima (Ministerio de Transportes y Comunicaciones, 2016; Tapia et al., 2018; Torres Trujillo, 2012; Yachiyo Engineering Co., 2005).

1.2. Minería de Datos

Cuando se habla de minería de datos, esta posee múltiples definiciones de acuerdo con distintas épocas. "La Minería de Datos es la extracción no trivial de información implícita, previamente desconocida y potencialmente útil de los datos (Frawley et al., 1992)" (Nisbet et al., 2018b). Este concepto lo complementa Fayyad et al. (1996), referenciado de Nisbet et al. (2018b), adicionando "la aplicación de varios algoritmos que permitan encontrar relaciones o patrones en un conjunto de datos", lo cual implica la introducción del concepto de descubrimiento de conocimiento. La anterior definición se extiende para incluir el término de análisis predictivo como el "el proceso de identificar patrones válidos, novedosos y fáciles de entender dentro del conjunto de datos con el fin de sacar nuevos resultados con pequeño margen de error" (Ozaydin et al., 2016). Para ello, también es necesario saber el preprocesamiento necesario de datos, la interpretación de los resultados y el suministro de la información extraída en una forma útil en la toma de decisiones (Nisbet et al., 2018b). En términos generales, la Minería de Datos es el estudio de recopilar, limpiar, procesar, analizar y obtener información útil de los datos (Aggarwal, 2015) que permita la visualización de soluciones posibles a problemas planteados para proceder a la toma de decisiones correspondiente.

La Minería de Datos tiene un proceso respectivo explicado de muchas maneras diferentes, de las cuales solo se hará referencia a dos. El proceso tradicional consiste en tres fases importantes: recopilación de datos, extracción de características y limpieza de datos, y análisis de los datos (Aggarwal, 2015). Según Aggarwal (2015),

> *Los datos se recogen, por lo general, mediante software, los cuales son almancenados en bases de datos. Después, se limpian los datos, i.e., se eliminan datos nulos o incompletos; posteriormente se extraen las características, o atributos de interés. Por último, se analiza el conjunto de datos de interés mediante algoritmos especializado (p. 3).*

El otro proceso es denominado CRISP - DM (Cross Industry Standard Process for Data Mining) considerado el formato más completo de entre todos los procesos de Minería de Datos (Nisbet et al., 2018b). Este proceso está descrito en términos de un modelo de proceso jerárquico, que comprende cuatro niveles de abstracción: Fases, Tareas Genéricas, Tareas Especializadas e Instancias de Proceso (Azevedo & Santos, 2008). Corresponde a un número de seis fases cada uno con sus partes respectivas: Business Understanding, Data Understanding, Data

Preparation, Modeling, Evaluation y Deployment (Azevedo & Santos, 2008; Nisbet et al., 2018b; Wirth & Hipp, 2000). De acuerdo con Azevedo et al. (2008),

1.1.1. Aprendizaje Supervisado

El aprendizaje supervisado permite al usuario estimar el comportamiento de los atributos presentes en una base de datos de acuerdo con un conjunto de instancias que organiza y designa el usuario (Aggarwal, 2015; Roiger, 2017). Se busca, al emplear el aprendizaje supervisado, la reducción de las discrepancias entre los valores esperados y observados usando un proceso llamado capacitación. La finalidad de este proceso de capacitación es formar una descripción que pueda usarse para predecir ejemplos nunca vistos anteriormente. De acuerdo con Roiger (2017), el propósito del supervisado es construir modelos de clasificación a partir de un conjunto de datos que contienen ejemplos o no ejemplos de los conceptos a conocer para poder clasificar cada uno de ellos en su determinada clase. De entre todas las técnicas que incluye el conocimiento supervisado, e.g., regresión logística, algoritmos genéticos, red neuronal, entre otros, los árboles de decisión son una de las formas más populares de representación de conocimiento pues estas son estructuras jerárquicas (Kretowski, 2019). Al tratar de encontrar una solución, este sistema adopta una estrategia de búsqueda descendente dividiendo el espacio de la instancia en dos o más subespacios de acuerdo con una determinada función de los valores de los atributos de entrada, es decir, el algoritmo correspondiente examina todos los atributos y todos los valores de cada atributo lo cual ocasiona una prueba que considera un solo atributo, de modo que el espacio de la instancia se divide de acuerdo con el valor del atributo (Aggarwal, 2015; Daham et al., 2014). El proceso de búsqueda se repite hasta que no sea posible una separación adicional o se cumpla la condición de aplicar una clasificación única a cada miembro de los subgrupos derivados, lo cual queda como resultado que los nodos finales representan las diferentes clases (Aggarwal, 2015).

Los árboles de decisiones utilizan diferentes tipos de algoritmos, entre los más destacados se encuentran el algoritmo ID3, el algoritmo C4.5, el algoritmo CART y el algoritmo CHAID (Lee & Siau, 2001; Nisbet et al., 2018a; Quinlan, 1986; Urso et al., 2018). Conforme a Urso et al. (2018),

> *ID3 es un algoritmo diseñado por J. R. Quinlan a finales de los años 70 con una estructura iterativa basada en el Sistema de Aprendizaje Conceptual de Hunt, el cual durante el proceso de aprendizaje se selecciona aleatoriamente un conjunto de datos, este conjunto de datos según Quinlan (1986) es denominado ventana, para formar un primer árbol de decisión. Si el árbol clasifica bien todas las iteraciones correspondientes entonces este representa el árbol correcto y el proceso finaliza, caso contrario se hace una selección que contiene todos los datos que fueron clasificados incorrectamente para formar otro árbol y el proceso continua. El proceso termina cuando el algoritmo clasifique todo el conjunto de datos correspondiente (p. 4).*

La principal ventaja del algoritmo ID3 es el poder trabajar con grandes cantidades de datos y muchos atributos, pero donde se requiere un árbol de decisión razonablemente bueno sin muchos cálculos, es decir, que permita identificar el árbol de decisión correcto sin hacer muchas iteraciones aun con bases de datos abundantes (Quinlan, 1986; Urso et al., 2018). Por otra parte, el algoritmo C4.5 es una improvisación del algoritmo ID3. Según Urso et al. (2018),

> *Es formulado igualmente por J. R. Quinlan en 1993 cuyas ventajas sobre su predecesor son: el manejo de valores de atributos faltantes, capacidad de poder trabajar con atributos numéricos y categóricos, usar los criterios basados en información como medida de selección de atributos y presenta una fase de poda que previene el problema de ajuste excesivo (p. 4).*

En este sentido, el algoritmo CART (Classifier and Regression Tree) es un algoritmo de árbol de decisión que puede aplicar ambas tareas de clasificación y regresión respectivamente (Urso et al., 2018). CART fue

especificado por los investigadores Leo Breiman, Jerome Friedman, Richard Olshen y Charles Stone, está estructurado como una secuencia de preguntas simples el cual considerando las respuestas anteriores denomina la siguiente pregunta a plantear para formar un árbol de decisión que contiene como resultado la red de preguntas correspondiente (Nisbet et al., 2018a). Por último, el Algoritmo CHAID (Chi-squared Automatic Interaction Detector) depende de la prueba de chi-cuadrado para sacar las mejores divisiones a cada paso (Nisbet et al., 2018a). Este algoritmo necesita más preparación de los datos a diferencia del algoritmo CART pues este permite realizar múltiples divisiones en una variable (Lee & Siau, 2001; Nisbet et al., 2018a).

Otra técnica empleada en este trabajo es el *Rule-based Classification*. A diferencia de los árboles de decisión, el *Rule-Based Classification* propone reglas convenientes en una forma de condición *If-then* que permite al usuario almacenar y manipular información fácilmente con el propósito de extraer datos útiles para determinar una decisión beneficiosa para el usuario (Bhatnagar & Kumar, 2018; Mahmood & Lei, 2009). En ciertos casos se puede afirmar de manera general que se determinó dicha conclusión de acuerdo a los datos presentes en la respectiva base de datos, es decir, esta clasificación depende de la información presente y de las reglas generales que se asumen al querer analizar la base de datos (Bhatnagar & Kumar, 2018; Christopher, 2019; Mahmood & Lei, 2009). Por consiguiente, una clara diferencia con el árbol de decisión el cual busca una clasificación única a los datos usando una estrategia de búsqueda. Sin embargo, según Christopher (2019) su eficiencia depende de factores como la calidad de las reglas en el conjunto de reglas, el ordenamiento de las reglas y la cardinalidad del conjunto de reglas.

1.1.2. Aprendizaje No Supervisado

Aunque los datos se encuentran almacenados físicamente, i.e., base de datos local, en la nube, archivos planos, entre otros, en el aprendizaje no supervisado no existe una definición inicial de cómo agrupar los datos (Ozaydin et al., 2016). Por esta razón, una de las ventajas de este tipo de estrategias es que las similitudes entre los datos se determinan según el número de clústeres o conjuntos que se detecta en el conjunto de patrones que empieza a surgir (Ozaydin et al., 2016). Entre las técnicas presentes en el conocimiento no supervisado, se detalla la más importante que es la clusterización (K-means clustering). Acerca de clusterización, este no presenta una definición en específico, sin embargo, en términos generales es un algoritmo que permite encontrar agrupaciones naturales o agrupaciones basados en criterios entre los elementos del grupo usando un conjunto de datos multidimensional (Aggarwal, 2015; Ozaydin et al., 2016). En el proceso de clusterización, el conjunto de datos se divide en grupos de datos el cual es denominado Clúster que corresponde a una colección de datos en un clúster que son similares entre sí y que se diferencian de otros clústeres (Raval & Jani, 2016). El proceso utiliza para su división la fórmula de distancia euclidiana (ver Ecuación 1) el cual cierto dato se agrupa en un clúster si el dato está cerca de dicho centroide, el proceso termina cuando no se requiere un cambio en los clústeres (Raval & Jani, 2016; Wu, 2012).

Ecuación 1

Clusterización (Distancia Euclideana)

$$D_e(\mathrm{P},\mathrm{Q}) = \sqrt{(\mathrm{p}_1 - q_1)^2 + (\mathrm{p}_2 - q_2)^2 + \ldots + (\mathrm{p}_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (\mathrm{p}_i - q_i)^2}$$

Fuente: (Raval & Jani, 2016)

El proceso de clusterización es una de las técnicas pioneras más populares cuyos propósitos son el entendimiento (detectar grupos de datos similares que comparten características) y la utilidad (permitir la diferenciación entre grupos de datos para mejor visualización de estos) (Raval & Jani, 2016; Wu, 2012). De acuerdo con Wu (2012), el proceso de clusterización tiene distintos tipos de acuerdo a los casos que se quieren realizar empleado

clusterización: *Prototype-Based Algorithms*, *Density-Based Algorithms*, *Graph-Based Algorithms*, *Hybrid Algorithmsm y Algorithm-Independent Methods*.

2. Metodología

2.1. Generalidades

En esta sección se describe una metodología para emplear técnicas de minería de datos, y análisis geoespacial para determinar los trayectos más contaminantes; así como, los sectores con mayor índice de polución por servicio de una empresa Courier en la provincia de Lima y la provincia constitucional del Callao; y, cuanta contaminación en $KgCO_2$. En el análisis la información fue segmentada en semanas para una mejor compresión de los resultados. De igual forma, en el análisis se tiene en cuenta la Norma Europea de contaminación de autos (Comunidad Europea, 2007); la cual es empleada en Perú desde el año 2018.

El proyecto se divide en dos fases, siendo la primera el análisis de los datos usando las técnicas de minería de datos para encontrar relaciones entre ellos. Luego de relacionados los datos, en la segunda parte, se asociaron con información georeferenciadas para detectar en qué lugares se produce mayor polución debido a las diferentes entregas realizadas por la empresa denominada *Courier* A (se asoció este nombre a la empresa por cuestiones de privacidad, cuya sede principal se ubica en la ciudad de Lima-Perú), dedicada a servicios de mensajería. Dicha información espacial se emplea para calcular las distancias respectivas de los despachos y estimar de esta manera la contaminación de $KgCO_2$ producida por trayecto.

En el presente estudio se emplearon los registros del último trimestre del 2019 de la empresa *Courier* A. Cabe destacar que el análisis realizado se concentró en la provincia de Lima y la provincia constitucional del Callao. El número de instancias, del conjunto inicial de datos, es 3557 registros, con 30 atributos. Luego de un proceso de depuración (eliminación de valores homónimos o datos vacíos) (Nisbet et al., 2018b), resultaron 2798 registros, con 13 atributos. Estos 13 atributos se presentan en el Cuadro 1, donde se describen cada uno de ellos.

Cuadro 1
Atributos de la base de datos a analizar

Nombre del atributo	Tipo de Dato	Descripción
Fecha Creacion	*Date*	La fecha de creación del pedido
Cliente	*String*	Nombre del cliente que hace el pedido
Des. Prod	*String*	Descripción del producto a enviar
Tipo Serv.	*String*	El tipo de servicio que escoge el cliente para enviar el producto
Imp. Cobrado COD	*Integer*	El impuesto total que se le cobra al cliente destino
Valor total	*Integer*	El impuesto total que ya fue cobrado por el sistema
Origen	*String*	El punto de partida del producto
Departamento	*String*	Departamento en donde se hace la entrega del pedido
Provincia	*String*	Provincia en donde se hace la entrega del pedido
Distrito	*String*	El distrito en donde se hace la entrega del pedido
Zona	*String*	Tipo de zona en el que hace la entrega del pedido
Evento	*String*	Situación actual del paquete
Fecha	*Date*	La fecha de llegada del producto

Fuente: Propia

Estos 13 atributos describen de manera concisa la información general de la base de datos de interés; principalmente la relación entre los atributos *Cliente*, *Origen* (ciudad de despacho), *Departamento*, *Provincia*, *Distrito,* y *Evento* (estado del paquete, por ejemplo, enviado o entregado). Lo anterior permite establecer las interacciones entre despachos y clientes específicos. Cabe mencionar, entonces, que el atributo *Evento* se empleó como la clase de interés en nuestro trabajo. Lo anterior debido a la importancia que este representa en

todo el proceso de mensajería. Por tal motivo, en la fase inicial de este trabajo se dividió en dos partes: tomando el conjunto inicial de valores para la clase *Evento*; la cual incluye entre otros, los siguientes valores: *Entregado*, *Entregado* a *Remitente*, *Falso Flete*, *Creado*, *Rechazado*, *Dirección Inubicable*, *Dirección Equivocada*, por citar algunos. Y una segunda parte que incluye solamente los valores de: *Creado*, *Entregado*, y *No entregado*; por ser más específicos en el proceso de envío (ver Cuadro 2). Cabe anotar, sin embargo, que el valor *Creado* representa la creación del evento en el sistema más no implica que el paquete haya salido de las instalaciones de la empresa.

Las instancias que se visualizan son una parte de la clasificación total que usa la empresa Courier *A* para determinar la situación de los paquetes, es decir, las instancias en el gráfico no son todas las clasificaciones impuestas por la empresa; sin embargo, son pertenecientes a la base de datos proporcionada por ella. La principal razón de tomar esta consideración, es por el número de datos que las instancias tienen y, de acuerdo con la gráfica, la instancia *Entregado* supera en gran cantidad a la mayoría de las instancias presentes en la base de datos, además de que este atributo posee muchas instancias diferentes pero que corresponden a situaciones similares como *Entregado a Remitente* que fácilmente se puede relacionar con *Entregado*. En este caso todo lo referente con la entrega de productos se encuentra clasificado en la nueva instancia "*Entregado*" mientras que todo lo referente con la creación de la transacción se encuentra en "*Creados*" y todo lo referente con la particularidad en la cual no se pueda entregar el paquete o que el paquete sea falso o rechazado se encuentra en "*No entregado*" (ver Cuadro 3).

Cuadro 2
Datos del atributo evento

Evento	Lima	Callao	Total
Creado	185	7	192
Entregado	2451	61	2512
Cambio de Domicilio	1	-	1
Contacto Desconocido	6	-	6
Coordinado con Cliente	3	-	3
Dirección Inubicable	8	-	8
Dirección Incorrecta	4	-	4
Entregado a Remitente	27	1	28
Falso Flete	10	-	10
HUB	2	-	2
Incidente	1	-	1
Nadie en Casa	5	-	5
Nota	7	-	7
Oficina Cerrada	1	-	1
Rechazado	4	1	5
Recojo	1	-	1
Retorno a HUB	11	1	12
Total	2727	71	2798

Fuente: Propia

Cuadro 3
Datos, resumidos, del atributo evento

Evento	Lima	Callao	Total
Creado	185	7	192
Entregado	2478	62	2540
No Entregado	64	2	66
Total	**2727**	**71**	**2798**

Fuente: Propia

De igual forma, y con el objetivo de agrupar información de manera georefenciada se procedió a incluir la latitud y longitud de todos los distritos pertenecientes a la provincia de Lima y la provincia constitucional del Callao (cabe mencionar que la provincia constitucional del Callao sólo tiene un destino, i.e., distrito el Callao) a los cuales se envían los pedidos. En el Cuadro 4 se ilustra la latitud y longitud general de los distritos pertenecientes a la provincia de Lima y la provincia constitucional del Callao; así como la distancia desde la empresa *Courier* A hasta dicho distritos. El propósito de la clusterización en este trabajo es organizar en grupos las localidades con mayor concentración de entregas. Lo cual se describe en la siguiente sección.

Cuadro 4
Datos Latitud, Longitud de los distritos a evaluar; así como,
la distancia, promedio en kilómetros desde la empresa *Courier* A

Distrito	Latitud	Longitud	Distancia
ANCÓN	-11.69655	-77.11165	39.9 Km
ATE	-12.03873	-76.89687	18.1 Km
BARRANCO	-12.14396	-77.02027	11.5 Km
BREÑA	-12.05970	-77.05012	1.6 Km
CALLAO	-12.05195	-77.12578	7.5 Km
CARABAYLLO	-11.79499	-76.98929	29.6 Km
CHACLACAYO	-11.99248	-76.77618	32 Km
CHORRILLOS	-12.19235	-77.00896	16.9 Km
CIENEGUILLA	-12.07317	-76.77707	31.6 Km
COMAS	-11.93286	-77.04067	14.2 Km
EL AGUSTINO	-12.04205	-76.99571	7.9 Km
INDEPENDENCIA	-11.98931	-77.04733	5.9 Km
JESÚS MARÍA	-12.07819	-77.04641	3.4 Km
LA MOLINA	-12.09018	-76.92234	15.4 Km
LA VICTORIA	-12.07336	-77.01642	5.5 Km
LIMA	-12.06211	-77.03653	3 Km
LINCE	-12.08657	-77.03665	4.95 Km
LOS OLIVOS	-11.96599	-77.07307	10.1 Km
LURIGANCHO-CHOSICA	-12.00188	-76.91889	17 Km
LURIN	-12.23805	-76.78386	38 Km
MAGDALENA DEL MAR	-12.09237	-77.07331	5 Km
MIRAFLORES	-12.12150	-77.02591	8.7 Km
PACHACAMAC	-12.22338	-76.84771	28.8 Km
PUEBLO LIBRE	-12.07664	-77.06786	3.4 Km
PUENTE PIEDRA	-11.87683	-77.07448	20 Km
PUNTA HERMOSA	-12.33268	-76.82570	40.8 Km
RÍMAC	-12.02030	-77.03546	4.4 Km
SAN BORJA	-12.09645	-76.99569	8.9 Km
SAN ISIDRO	-12.09790	-77.03537	6.5 Km
SAN JUAN DE LURIGANCHO	-11.94875	-76.97791	14.3 Km
SAN JUAN DE MIRAFLORES	-12.15585	-76.97213	16.5 Km
SAN LUIS	-12.07236	-76.99589	7.8 Km
SAN MARTÍN DE PORRES	-11.98676	-77.09766	8.5 Km
SAN MIGUEL	-12.07866	-77.09528	5.05 Km
SANTA ANITA	-12.04293	-76.94567	10 Km
SANTIAGO DE SURCO	-12.12510	-76.98192	11.6 Km
SURQUILLO	-12.11420	-77.01047	8.3 Km
VILLA EL SALVADOR	-12.21350	-76.93703	22.4 Km
VILLA MARÍA DEL TRIUNFO	-12.17664	-76.91897	21 Km

Fuente: Propia

2.2. Desarrollo

2.2.1. Fase 1: Mineria de Datos

Una vez seleccionado y depurado el conjunto de datos, se procedió al análisis de los mismos. Para ello se emplearon técnicas de árboles de decisión, de clasificación, y de clusterización soportadas en la herramienta WEKA (*Waikato Environment Knowledge Acquisition*, https://www.cs.waikato.ac.nz/ml/weka/). Dichas técnicas son: *J48* (árbol de decisión), *OneRule* (clasificación), y *K-means* (clusterización), respectivamente. Antes de utilizar WEKA, los datos fueron preparados el un archivo plano de formato *CSV* (*Comma Separated Values*). Primero se analizaron los datos aplicando la técnica *J48* y *OneRule*. Como resultado se pone de manifiesto que el atributo *Cliente* está relacionado con el atributo *Evento*. Por ejemplo, la relación entre el *Cliente H* y el *Evento Entregado* tiene una probabilidad del 98.25% que el paquete sea entregado. Pero si evaluamos la probabilidad que un paquete sea no sea entregado al *Cliente J* es del 30.0%. A partir de este resultado, por ejemplo, la empresa *Courier* A puede iniciar un proceso de revisión interna y determinar porque 3 de cada 10 despachos al *Cliente J* no son entregados (ver Cuadro 5). Cabe resaltar que un 95% de las reglas fue clasificada correctamente (i.e., 2632/2798).

Cuadro 5
Análisis realizado en WEKA (Arbol de Decisión)

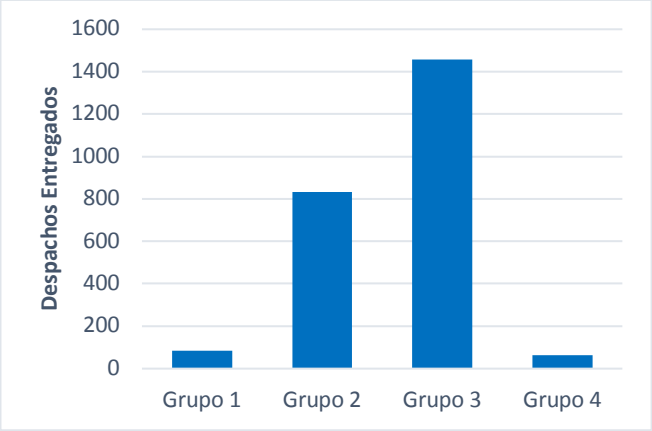
Cliente	Evento	Instancias (Observaciones/Clasificación Incorrecta)
A	ENTREGADO	(768.0/65.0)
B	ENTREGADO	(430.0/5.0)
C	CREADO	(27.0/2.0)
D	NO ENTREGADO	(1)
E	NO ENTREGADO	(1)
F	ENTREGADO	(103.0/20.0)
G	CREADO	(65.0/1.0)
H	ENTREGADO	(1091.0/19.0)
I	CREADO	(2.0/1.0)
J	ENTREGADO	(30.0/9.0)
K	ENTREGADO	(9.0/2.0)
L	CREADO	(1)
M	CREADO	(11)
N	CREADO	(6)
O	CREADO	(6)
P	ENTREGADO	(27)
Q	ENTREGADO	(3)
R	ENTREGADO	(112.0/31.0)
S	CREADO	(6)
T	ENTREGADO	(20.0/1.0)
U	ENTREGADO	(45.0/4.0)
V	ENTREGADO	(21.0/1.0)
W	NO ENTREGADO	(1)
X	ENTREGADO	(3)
Y	ENTREGADO	(5)
Z	CREADO	(3)
AA	ENTREGADO	(1)

Fuente: Propia

Luego de aplicados los métodos de clasificación y clusterización se procedió a evaluar la relación entre los atributos *CLIENTE* y *CIUDAD DESTINO*. Este análisis permitió detectar los distritos con mayor número de

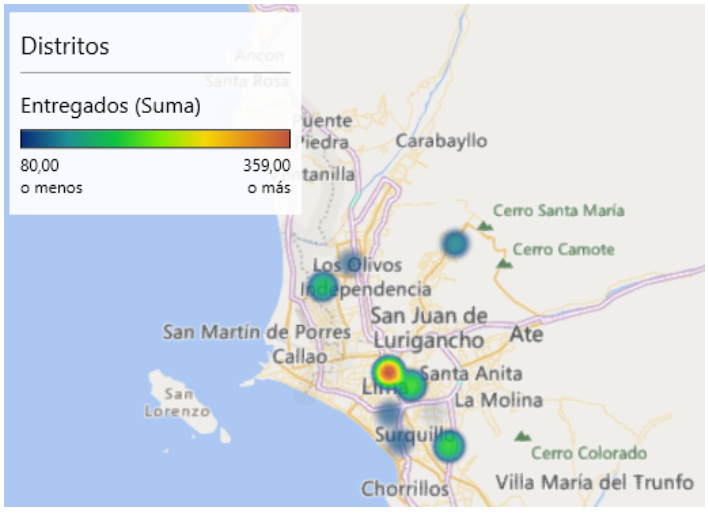
despachos; así como, el agrupamiento de estos. Como resultado, se determinaron 4 grupos. El grupo 1 corresponde a los distritos de Chaclacayo, Santa Anita, Cieneguilla, Punta Hermosa y Lurin; el grupo 2 corresponde a Pachacamac, Villa Maria del Triunfo, Villa el Salvador, Ate, La Molina, Lurigancho-Chosica, San Juan de Miraflores, Chorrillos, Santiago de Surco, Miraflores, Surquillo, San Borja, Barranco, San Luis y El Agustino; y el grupo 3 corresponde a Lima, Ancon, Breña, Comas, Independencia, Jesus Maria, La Victoria, Lince, Los Olivos, Madgalena de Mar, Pueblo Libre, Puente Piedra, Rimac, San Isidro, San Juan de Lurigancho, San Martin de Porres, San Miguel y Carabayllo. Finalmente, el grupo 4 incluye al distrito de El Callao (ver Gráfica 3). Como se aprecia, la empresa *Courier* A se enfoca mayoritariamente en los distritos del grupo 3 y en el grupo 2. En la Gráfica 4 se visualizan aquellos distritos en los cuales el número de entregas fue mayor a 75, a saber: Lima (359), La Victoria (194), San Martín de Porres (164), San Juan de Lurigancho (123), San Isidro (104), y Los Olivos (103); del grupo 3; y, Santiago de Surco (187), Miraflores (104), San Borja (85), Chorrillos (82), y Surquillo (80); del grupo 2. Cabe mencionar que estos distritos representan un 62% de los despachos totales de la empresa durante el periodo de estudio. De igual forma es relevante indicar que solamente se entregaron 62 encomiendas en el distrito de El Callao, grupo 4.

Gráfica 3
Grupos designados por Clusterización



Fuente: Propia

Gráfica 4
Distribución de entregas en la Provincia de Lima



Fuente: Propia

2.2.2. Fase 2: Descubrimiento de la Polucion Total de la Empresa *Courier* A

En esta segunda parte, usando la base de datos depurada (i.e., los datos ajustados del atributo *Evento*) y los datos de ls atributos *Latitud* y *Longitud* de cada uno de los distritos, se analiza el total de polución por dióxido de carbono que produce la empresa *Courier* A por semana. Para ello se tuvo en cuenta los viajes realizados, la información técnica de los autos pertenecientes a la empresa, y la información de la norma europea de vehículos (Comunidad Europea, 2007). La contaminación, mencionado anteriormente, se evaluó por semanas; las cuales corresponden al ultimo trimestre del año 2019, i.e., octubre 2019 – diciembre 2019. El total de semanas de los los 3 meses son 13, los cuales se trabajaron en una relación de 4-4-5. Siendo 4 semanas para octubre empezando el miércoles 2 de octubre del 2019 hasta el martes 29 de octubre del 2019, 4 semanas para noviembre empezando el miércoles 30 de octubre del 2019 hasta el martes 26 de novimebre del 2019, y 5 semanas para diciembre empezando el 27 de noviembre del 2019 hasta el martes 31 de diciembre del 2019. Cabe destacar que la empresa *Courier* A tiene algunas condiciones que aplica en el despacho de sus pedidos de manera efectiva. Dichas condiciones son las siguientes: (i) la realización de pedidos siempre se hace un dia o varios días antes, razón por la cual no se considera dentro del rango a evaluar el 1 de octubre del 2019; (ii) al inicio del dia todos los vehiculos comienzan con el tanque combustible lleno; (iii) de acuerdo con la empresa, solamente sale un vehiculo por pedido respectivamente; (iv) el día en que salen los vehiculos, se realizan todos los despachos correspondientes al día; y, (v) la empresa tiene 7 vehiculos disponibles, según la condición anterior, en el caso que se presente más de 7 despachos, algunos de ellos saldrán más de una vez.

A continuación se presenta la información técnica de los vehículos (ver Cuadro 6):

Cuadro 6
Información básica de los vehículos

Datos de los vehículos	Cantidad	Norma Europea	Tipo de Combustible	Consumo de Combustible
NEW VAN-CHANGAN http://www.globalchangan.com	1	EURO IV	Gasolina	10.1 Km/L
V21 MINI TRUCK-DFSK	2	EURO III	Gasolina	11.2 Km/L
V25L-DFSK www.dfdongfeng.com.cn	1	EURO V	Gasolina	10.1 Km/L
H100 TRUCK-HYUNDAI www.hyundai.com	1	EURO II	Diesel	10.2 Km/L
NJ1020DF-YUEJIN www.yuejin.cl	1	EURO II	Diesel	10.3 Km/L
N-P 15V-MITSUBISHI www.mitsubishi-motors.com	1	EURO IV	Diesel	10.8 Km/L

Fuente: Propia

Es importante mencionar, que los atributos *Fecha Creacion* y *Fecha* que especifican la creación del pedido y la entrega del mismo, respectivamente, son de gran importancia en el análisis de polución. Debido, a que los vehículos vuelven a quedar disponibles, una vez han despachado (entregado) todos los pedidos. Respecto al proceso principal, de esta segunda fase, se tienen en consideración 4 datos importantes: la latitud y longitud de todos los distritos pertenecientes a la provincia de Lima y la provincia constitucional del Callao, la latitud y longitud de la sede principal de la empresa *Courier* A, la información básica de los vehículos de la empresa, y la polución generada en promedio por litro de gasolina y diesel, los cuales son $2.3\ kg\ de\ CO_2/L$ y $2.6\ kg\ de\ CO_2/L$, respectivamente. Usando la aplicación Google Maps, se calcula la distancia entre la empresa *Courier* A respecto a cada uno de los distritos, todos las distancias se miden en Kilómetros (ver Cuadro 4). Finalmente, por cada vehículo, se procede a relacionar tipo de combustible, i.e., el Euro correspondiente, el consumo de

combustible, la distancia entre la empresa y el distrito destino, y litros consumidos de combustible. De igual forma, se tiene en cuenta si el despacho fue o no entregado. Ya que, en cualquier caso, genera consumo de combustible; por ende contaminación ambiental. La ecuación 2 describe la cantidad de contaminación emitida al ambiente de la empresa *Courier A*, para el periodo de estudio. Donde, *DR,* distancia recorrida (valor en Kilómetros, i.e., Km), es el trayecto recorrido entre la empresa, el destino, y de vuelta. *CC,* consumo de combustible (valor en kilómetros por litro, i.e., Km/L), valor que captura el consumo de combustible de un vehículo por kilómetro recorrido (ver Cuadro 6), y por último *PLC*, polución por litro de combustible (valor en kilogramo CO_2 por litro, i.e., $KgCO_2/L$), captura el dióxido de carbono, en kilogramos, que se produce por litro de combustible, ya sea gasolina o diesel, consumido.

Ecuación 2
Fórmula de Polución por $\boldsymbol{CO_2}$

$$PCO_2 = \frac{DR}{CC} * PLC$$

Fuente: Propia

3. Resultados

A continuación se presentan los principales resultados del análisis emisiones de $KgCO_2$ durante el periodo de estudio.

3.1. Análisis de Polución Ambiental

En los cuadros 7 – 12 se presenta el análisis de polución de las operación de la empresa *Courier A.* Cabe acotar que el promedio de consumo de combustible de cada automóvil registrado en la empresa observa la norma europea de vehículos (Comunidad Europea, 2007). En los cuadros 7, 9, y 11 se presentan las emisiones diarias de $KgCO_2$; mientras que en los cuadros 8, 10, y 12 discriminan las emisiones de $KgCO_2$ producidas por cada tipo de vehículo descritos en el Cuadro 6. Durante el periodo de estudio, i.e., Octubre, Noviembre, y Diciembre de 2019, las emisiones totales de $KgCO_2$ fueron apróximandamente 10 toneladas y media de CO_2. De los cuales, y en base a los resultados discutidos en la sección 2.2.1, unos 6.652,62 $KgCO_2$ son emitidos a la atmósfera en los distritos listados en lo grupos 2 y 3. De igual forma, la distribución de emisión por $KgCO_2$ por vehículo se lista a continuación, de mayor a menor: (i) V21 Mini Truck DFSK, 2698.70; (ii) NJ1020DF Yuejin, 1847.51; H100 Truck Hyundai, 1792.95 (iii), N-P 15V Mitsubishi, 1452.82; (iv) New Van Changan, 1384.87; y (v),V25L DFSK, 1343.50. De igual forma mencionar que en promedio por cada viaje o despacho se emiten al ambiente un aproximado de 4.33 $KgCO_2$.

Cuadro 7
Emisiones diarias de $KgCO_2$ correspondientes
al mes de Octubre, semanas 1 – 4. Caso 1

Mes 1: Octubre	Semana 1		Semana 2		Semana 3		Semana 4	
	Pedidos	Emisión ($KgCO_2$)	Pedidos	Emisión ($KgCO_2$)	Pedidos	Emisión ($KgCO_2$)	Pedidos	Emisión ($KgCO_2$)
Día 1	3	25.94	6	21.26	3	20.43	19	65.77
Día 2	17	87.78	13	52.98	3	19.44	5	24.82
Día 3	9	46.85	4	17.77	3	10.50	42	197.64
Día 4	7	20.85	9	60.17	8	48.41	79	379.22
Día 5	-	-	1	4.54	-	-	-	-
Día 6	2	15.62	-	-	25	163.67	32	214.34
Día 7	-	-	17	75.75	2	8.62	33	155.10
Total	**38**	**197.04**	**50**	**232.47**	**44**	**271.07**	**210**	**1036.89**

Fuente: Propia

Cuadro 8

Emisiones de $KgCO_2$ correspondientes al mes de Octubre según tipo de vehículo, semanas 1 – 4. Caso 1

Tipo Vehículo	Total Viajes	Distancia Total Recorrida (Km)	Consumo Promedio Combustible Vehículo (L/Km)	Consumo Total Combustible Recorrido (L)	Total Emisiones ($KgCO_2$)
H100 TRUCK-HYUNDAI	60	1312.00	10.20	128.63	334.43
NEW VAN-CHANGAN	44	999.20	10.10	98.93	227.54
NJ1020DF-YUEJIN	58	1360.20	10.30	132.06	343.35
N-P 15V-MITSUBISHI	41	923.80	10.80	85.54	222.40
V21 MINI TRUCK-DFSK	100	1972.10	11.20	176.08	405.44
V25L-DFSK	39	897.20	10.10	88.83	204.31
Total General	**342**	**7464.50**	**-**	**710.07**	**1737.47**

Fuente: Propia

Cuadro 9

Emisiones diarias de $KgCO_2$ correspondientes al mes de Noviembre, semanas 5 – 8. Caso 1

Mes 2: Noviembre	Semana 5		Semana 6		Semana 7		Semana 8	
	Pedidos	Emisión ($KgCO_2$)	Pedidos	Emisión ($KgCO_2$)	Pedidos	Emisión ($KgCO_2$)	Pedidos	Emisión ($KgCO_2$)
Dia 1	18	119.16	14	79.72	77	235.24	78	310.56
Dia 2	33	85.14	17	75.13	137	508.99	27	127.50
Dia 3	-	-	21	114.64	117	392.55	18	54.56
Dia 4	12	62.07	31	145.15	54	191.68	40	148.22
Dia 5	-	-	-	-	94	331.91	-	-
Dia 6	21	58.93	12	87.49	189	767.69	11	39.31
Dia 7	25	149.63	57	224.14	66	205.95	6	22.45
Total	**109**	**474.93**	**152**	**726.27**	**734**	**2634.01**	**180**	**702.6**

Fuente: Propia

Cuadro 10

Emisiones de $KgCO_2$ correspondientes al mes de Noviembre según tipo de vehículo, semanas 5 – 8. Caso 1

Vehículo	Total Viajes	Distancia Total Recorrida (Km)	Consumo Promedio Combustible Vehículo (L/Km)	Consumo Total Combustible Recorrido (L)	Total Emisiones ($KgCO_2$)
H100 TRUCK-HYUNDAI	176	2977.30	10.20	291.89	758.92
NEW VAN-CHANGAN	165	2617.60	10.10	259.17	596.09
NJ1020DF-YUEJIN	174	3061.00	10.30	297.18	772.68
N-P 15V-MITSUBISHI	159	2523.20	10.80	233.63	607.44
V21 MINI TRUCK-DFSK	344	5756.70	11.20	513.99	1182.18
V25L-DFSK	157	2724.90	10.10	269.79	620.52
Total General	**1175**	**19660.70**	**-**	**1865.66**	**4537.82**

Fuente: Propia

Cuadro 11
Emisiones diarias de $KgCO_2$ correspondientes
al mes de Diciembre, semanas 9 – 13. Caso 1

Mes 3: Diciembre	Semana 9		Semana 10		Semana 11		Semana 12		Semana 13	
	Pedidos	Emisión $(KgCO_2)$	Pedidos	Emisión $(KgCO_2)$	Pedidos	Emisión $(KgCO_2)$	Pedidos	Emisión $(KgCO_2)$	Pedidos	Emisión $(KgCO_2)$
Dia 1	15	86.87	83	397.38	17	100.43	36	183.21		
Dia 2	8	34.61	25	131.11	5	24.35	24	109.92	49	177.52
Dia 3	39	171.69	25	160.27	28	102.21	43	210.04	82	331.13
Dia 4	11	50.62	25	85.27	59	297.05	30	161.62	57	285.33
Dia 5	-	-	-	-	-	-	-	-	-	-
Dia 6	18	61.49	7	28.03	15	86.15	8	52.44	19	111.31
Dia 7	17	45.66	27	137.99	41	176.58	51	173.87	48	270.90
Total	**108**	**450.94**	**192**	**940.05**	**165**	**786.77**	**192**	**891.1**	**255**	**1176.19**

Fuente: Propia

Cuadro 12
Emisiones de $KgCO_2$ correspondientes al mes de Noviembre
según tipo de vehículo, semanas 5 – 8. Caso 1

Vehículo	Total Viajes	Distancia Total Recorrida (Km)	Consumo Promedio Combustible Vehículo (L/Km)	Consumo Total Combustible Recorrido (L)	Total Emisiones $(KgCO_2)$
H100 TRUCK-HYUNDAI	144	2744.60	10.20	269.08	699.60
NEW VAN-CHANGAN	125	2464.60	10.10	244.02	561.25
NJ1020DF-YUEJIN	137	2897.80	10.30	281.34	731.48
N-P 15V-MITSUBISHI	122	2587.80	10.80	239.61	622.99
V21 MINI TRUCK-DFSK	266	5410.50	11.20	483.08	1111.08
V25L-DFSK	118	2277.60	10.10	225.50	518.66
Total General	**912**	**18382.90**	**-**	**1742.63**	**4245.07**

Fuente: Propia

4. Conclusiones

Las técnicas de minería de datos empleadas en el presente trabajo permitieron establecer importantes descubrimientos respecto a la operación de una empresa *Courier*, a partir de los datos recopilados en los meses de octubre a diciembre de 2019. El primer hallazgo refiere al hecho de poder identificar la relación entre despachos y clientes. Dicha relación permite establecer el grado de cumplimiento en la entrega de un pedido, lo cual es un factor importante al momento de establecer familiaridad y reconocimiento entre una empresa *Courier* y un cliente (Zhang et. al, 2016). Estas reglas fueron identificadas de manera correcta con una confianza del 95%.

El segundo hallazgo está asociado con los distritos con mayor concentración de despachos. Primero, el 97.55% de la operación de la empresa Courier A se encuentra en la provincia de Lima, y sólo el 2.45% en la provincia del Callao. Respecto a la provincia de Lima, los distritos de Lima, La Victoria, Santiago de Surco, San Martín de Porres, San Juan de Lurigancho, San Isidro, Miraflores, Los Olivos, San Borja, Chorrillos, y Surquillo representan un 62% de las operaciones de despacho de la empresa *Courier* A.

Finalmente, el tercer hallazgo establece que el total de emisones de CO_2 son aproximadamente 10.5 toneladas. De ellos, 10,24 toneladas son arrojadas a la atmósfera en la provincia de Lima. Igualmente, se determinó que en promedio por cada viaje o despacho se emite al ambiente un aproximado de 4.33 $KgCO_2$.

A partir del conocimiento descubierto luego del análisis de los datos, la empresa *Courier* A pudiera revisar dos aspectos de su proceso de despacho, los cuáles listamos a continuación: (i) revisar la pertinencia de continuar con el despacho de pedidos a la provincia constitucional del Callao (sólo 2.45% de la operación total, i.e., 64 pedidos de 2606); y, (ii) revisar el proceso actual de despacho de pedidos (el vehículo sale a despacho con un solo pedido). Esto último, algo ineficiente en términos de gastos por combustible, a razón que el presente estudio identifica los distritos con mayor demanda de despachos. Como trabajo a futuro se planea implementar un sistema de mineria de datos adjunto al sistema de pedidos de la empresa *Courier* A (Mettler & Raber, 2011).

Referencias bibliográficas

Aggarwal, C. C. (2015). An Introduction to Data Mining. In *Data Mining: The Textbook* (p. 746). https://doi.org/10.1007/978-3-319-14142-8

Azevedo, A., & Santos, M. F. (2008). KDD , SEMMA AND CRISP-DM : A PARALLEL OVERVIEW Ana Azevedo and M . F . Santos. *IADIS European Conference Data Mining*, 182–185.

Bhatnagar, S., & Kumar, A. (2018). A rule-based classification of short message service type. *Proceedings of the 2nd International Conference on Inventive Systems and Control, ICISC 2018*, *Icisc*, 1139–1142. https://doi.org/10.1109/ICISC.2018.8398982

Chavhan, S., & Venkataram, P. (2019). Prediction based traffic management in a metropolitan area. *Journal of Traffic and Transportation Engineering (English Edition)*, *xxx*. https://doi.org/10.1016/j.jtte.2018.05.003

Christopher, J. (2019). The Science of Rule-based Classifiers. *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 299–303.

Daham, H., Cohen, S., Rokach, L., & Maimon, O. (2014). *Proactive Data Mining with Decision Trees*. Springer. https://doi.org/10.1007/978-1-4939-0539-3

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge Discovery in Databases: An Overview. *AI Magazine*, *13*(3), 57–70. https://doi.org/10.1007/3-540-57253-8_65

Gonzales, G. F., Zevallos, A., Gonzales-Castañeda, C., Nuñez, D., Gastañaga, C., Cabezas, C., Naeher, L., Levy, K., & Steenland, K. (2014). Environmental pollution, climate variability and climate change: A review of health impacts on the peruvian population. *Revista Peruana de Medicina Experimental y Salud Publica*, *31*(3), 547–556.

Kretowski, M. (2019). *Evolutionary Decision Trees in Large-Scale Data Mining* (J. Kacprzyk (ed.); First Edit). Springer. https://doi.org/10.1007/978-3-030-21851-5

Lee, S. J., & Siau, K. (2001). A review of data mining techniques. *Industrial Management and Data Systems*, *101*(1), 41–46. https://doi.org/10.1108/02635570110365989

Mahmood, A., & Lei, W. (2009). *One-RM : An Improved One -Rule Classifier*. *44*(2), 171–180.

Martínez Espinal, M. (2017). Transporte público de buses versus congestión y contaminación en Lima y Callao. *Economía*, *40*(79), 47–86. https://doi.org/10.18800/economia.201701.002

Mettler, T. & Raber, D. (2011). Developing a collaborative business intelligence system for improving delivery reliability in business networks. *Proceedings of the 17th International Conference on Concurrent Enterprising*, Aachen, 1–7.

Nisbet, R., Miner, G., & Yale, K. (2018a). Basic Algorithms for Data Mining: A Brief Overview. *Handbook of Statistical Analysis and Data Mining Applications*, 121–147. https://doi.org/10.1016/b978-0-12-416632-5.00007-4

Nisbet, R., Miner, G., & Yale, K. (2018b). The Data Mining and Predictive Analytic Process. *Handbook of Statistical Analysis and Data Mining Applications*, 39–54. https://doi.org/10.1016/b978-0-12-416632-5.00003-7

Objetivos y metas de desarrollo sostenible – Desarrollo Sostenible. (2015). Naciones Unidas. https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/

Ozaydin, B., Hardin, J. M., & Chhieng, D. C. (2016). Data Mining and Clinical Decision Support Systems. *Clinical Decision Support Systems: Theory and Practice*, 45–68. https://doi.org/10.1007/978-3-319-31913-1_3

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. https://doi.org/10.1007/bf00116251

Raval, U. R., & Jani, C. (2016). Implementing & Improvisation of K-means Clustering Algorithm. *International Journal of Computer Science and Mobile Computing*, *55*(5), 191–203.

Roiger, R. (2017). *Data mining: a tutorial-based primer* (Second Edi). Pearson Education, Inc.

Romero, Y., Chicchon, N., Duarte, F., Noel, J., Ratti, C., & Nyhan, M. (2020). Quantifying and spatial disaggregation of air pollution emissions from ground transportation in a developing country context: Case study for the Lima Metropolitan Area in Peru. *Science of the Total Environment*, *698*, 134313. https://doi.org/10.1016/j.scitotenv.2019.134313

Tavera, J. A. (2001). After privatization: Regulation of Peruvian public utilities. *Quarterly Review of Economics and Finance*, *41*(5), 713–725. https://doi.org/10.1016/S1062-9769(01)00099-0

Torres Trujillo, R. (2012). Introducción a la Regulación del Transporte en el Perú. *Círculo de Derecho Administrativo*, 8.

Urso, A., Fiannaca, A., La Rosa, M., Ravì, V., & Rizzo, R. (2018). Data mining: Classification and prediction. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, *1–3*, 384–402. https://doi.org/10.1016/B978-0-12-809633-8.20461-5

Wirth, R., & Hipp, J. (2000). CRISP-DM : Towards a Standard Process Model for Data Mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, *24959*, 29–39. https://doi.org/10.1.1.198.5133

Wu, J. (2012). Advances in K-means Clustering. In J. Chen (Ed.), *Springer Theses*. Springer. https://doi.org/10.1007/978-3-642-29807-3

Yachiyo Engineering Co., L. (2005). Plan Maestro de Transporte Urbano para el Area Metropolitana de Lima y Callao en la Republica del Peru. In *MINISTERIO DE TRANSPORTES Y COMUNICACIONES DE LA REPÚBLICA DEL PERÚ* (Issue Fase 1).

Zhang, S., Qin, L., Zheng, Y., & Cheng, H. (2016). Effective and Efficient: Large-Scale Dynamic City Express. *IEEE Transactions on Knowledge and Data Engineering*, *28(12),* 3203–3217. https://doi: 10.1109/TKDE.2016.2604806.