Estimación de cantidad de personas por medio del CO_2 con modelo de clasificación k-NN

1ro Adrián Córdoba-Soto
Escuela de Ciencias de la Computación
Universidad de Costa Rica
San José, Costa Rica
adrian.cordobasoto@ucr.ac.cr

2do Celeste Balladares-Barboza
Escuela de Ciencias de la Computación
Universidad de Costa Rica
San José, Costa Rica
jenory.balladares@ucr.ac.cr

3ro Roy Padilla-Calderón
Escuela de Ciencias de la Computación
Universidad de Costa Rica
San José, Costa Rica
roy.padilla@ucr.ac.cr

4to Jose Pablo Mora-Villalobos
Escuela de Ciencias de la Computación
Universidad de Costa Rica
San José, Costa Rica
jopamovil@gmail.com

5to Juan José Valverde-Campos
Escuela de Ciencias de la Computación
Universidad de Costa Rica
San José, Costa Rica
juan.valverdecampos@ucr.ac.cr

***Resumen*—El establecer la cantidad de personas en una habitación es un mecanismo que brinda muchas aplicaciones en el uso cotidiano. No obstante, este presenta retos debido a estar propenso a errores y tener distintos acercamientos en la forma de lograr dicho objetivo. El presente trabajo busca analizar uno de estos métodos al utilizar el modelo clasificador k-NN para estimar de la cantidad de personas a partir de las concentraciones del CO_2 en una habitación.**

Al optimizar los hiperparámetros de entrenamiento del modelo k-NN se consiguió una predicción en promedio del 47.8% en el mejor de los casos. Estos resultados pudiere mostrar que este modelo para la tarea establecida puede tener problemas de escalamiento, debido a una contraposición de lo que se encontró en la literatura con respecto a este en otros escenarios.

A su vez, se destaca que los tratamientos a nivel de CO_2 pudieran tener un efecto para la consecución de las tareas deseadas, al encontrarse que estos afecten al modelo.

***Palabras claves*—k-NN, CO_2, Conteo de Personas, Edificios Inteligentes, Sensores, HVAC.**

Introducción

Los edificios inteligentes "son un tipo de edificio que tienen una inversión eficiente y razonable en el manejo de energía diseñado para aportar un ambiente conveniente y cómodo considerando la optimización de las relaciones entre estructura, sistema, servicios y administración dentro del mismo" [1]. Considerando que los edificios consumen alrededor del 40% de la energía total en el mundo para proporcionar un ambiente interior cómodo y saludable para los ocupantes [2] y que las crisis energéticas y la concisión del desarrollo sostenible, generan un reto tanto en costos como en impacto ambiental, se ha invertido cada vez más en la creación de edificios enérgicamente eficientes [3].

Según el programa de Agencia Internacional de Energía en Edificios y Comunidades (IEA-EBC, por sus siglas en inglés) existen varios factores que influyen en el rendimiento energético de los edificios; uno de los factores significativos es el comportamiento de los ocupantes [4]. Son muchos los beneficios que puede proporcionar obtener la cantidad de personas que están ocupando una habitación o edificio. Por ejemplo, puede ser útil para regular el consumo de la calefacción, ventilación y refrigeración (HVAC por sus siglas en inglés), sin embargo, obtener la estimación de la cantidad de personas en una habitación o edificio puede llegar a ser un proceso engorroso, propenso a errores y costoso [5]. Existen muchas formas de lograr este objetivo que varían desde los instrumentos utilizados hasta las técnicas y metodologías [6].

En su estudio, Tsou et al [7], proponen un método de conteo que consiste en identificar el número de personas que entra y sale de una habitación. Para esto utilizan un arreglo de 16 sensores pasivos infrarrojos ubicado sobre la entrada de una habitación. Las mediciones se realizan a intervalos iguales y se analizan con varios algoritmos de clasificación, como lo son la máquina restringida de Boltzmann y una red neuronal convolucional.

En cambio, en el estudio Zuraime et al [8] se utiliza 3 sensores de CO_2 repartidos a lo largo de un cuarto de $876\mathrm{m}^3$ y uno afuera para monitorear el nivel de CO_2 del ambiente, reuniendo datos cada minuto. Estos se promedian cada cinco minutos, y fueron utilizados para 3 algoritmos de aprendizaje automático: redes neuronales artificiales, métodos de error de predicción y máquinas de soporte vectorial, obteniendo valores similares con los cuales lograron obtener precisiones entre 70-76% en promedio, lo cual se consideró relativamente bueno tomando en cuenta que se estaba trabajando con un número elevado de ocupantes.

Por otra parte, Meyn et al [9] mencionan que la correlación confiable de los niveles de CO_2 con la ocupación real es difícil debido a la alta variabilidad, la cual se puede dar por las fluctuaciones en los niveles ambientales de CO_2 y factores como el estado de la puerta de la habitación, si se encuentra cerrado abierta. Para contrarrestar este efecto, estimaron la ocupación en edificios utilizando mediciones de sensores de diversas fuentes, como CO_2, infrarrojos pasivos (PIR), video, sonido, contadores de tarjetas, entre otros.

En cambio, Szczurek et al [10], tomaron un acercamiento distinto para estimar la cantidad de personas; Estos autores utilizaron series de tiempo y mediciones de concentración de CO_2, incluyendo adicionalmente la temperatura y humedad, reuniendo muestras de espacios controlados a lo largo de un periodo de tiempo determinado; estas series fueron utilizadas para entrenar dos modelos de clasificación de aprendizaje mecánico: el k-NN (vecino más cercano) y el LDA (análisis de discriminante lineal), donde el método de k-NN superó a LDA.

El método propuesto en este trabajo se basa en un modelo de clasificación k-NN para estimar la cantidad de personas en una habitación, usando sólo mediciones de CO_2 y controlando ciertos factores del entorno como el estado de las puertas, las ventanas, el lapso de cada muestra y la cantidad de personas en la habitación. El objetivo de este trabajo es determinar la precisión con la que se puede estimar la cantidad de personas en una habitación a partir de medidas de concentración de CO_2, utilizando un modelo de clasificación k-NN. Para ello buscamos responder las siguientes preguntas de investigación:

- ¿Con qué precisión se puede estimar la cantidad de personas en una habitación con un medidor de CO_2?
- Basado en el estudio "Occupancy determination based on time series of CO_2 concentration, temperature and relative humidity" [10], ¿Qué tan efectivo es el estudio para poder contabilizar personas a partir de mediciones de CO_2 en la Escuela de las Ciencias de la Computación e Informática (ECCI) de la Universidad de Costa Rica?

La estructura de este documento es la siguiente: Sección I describe el desarrollo metodológico de esta investigación. La sección II presenta los resultados obtenidos por el modelo k-NN. La sección III presenta la discusión de los resultados del modelo k-NN y finalmente la sección IV expone las conclusiones obtenidas de la investigación.

I. Metodología

Para lograr el objetivo de estimar la cantidad de personas en un aula se segmenta la investigación en cuatro fases principales que son:

- **Adquisición de datos:** Fase de recolección de las muestras de CO_2 en la habitación. En esta se define el conjunto de medidas para la toma de las muestras y cómo son recolectadas.
- **Análisis Exploratorio de Datos:** Fase de análisis del dataset recolectado a fin de entender cómo se comporta el CO_2 en las distintas muestras y aplicando algunas alteraciones.
- **Análisis Experimental/Realización del modelo:** Fase en la cual se construye a partir del dataset recolectado el modelo de clasificación k-NN con diferentes hiperparámetros.
- **Evaluación del modelo:** Fase final en la cual se analiza si el modelo es preciso para estimar la cantidad de personas.

A continuación se va a detallar de manera más profunda los procesos a realizar en cado una de estas fases de la investigación.

A. Adquisición de datos

Con el fin de aprovechar la infraestructura existente se toma como objeto de estudio la concentración de CO_2 de la habitación 4-19 del cuarto piso de la Escuela de Ciencias de la Computación e Informática (ECCI), la cual cuenta con un sensor Adafruit SCD-30 NDIR conectado a un Raspberry, este provee la información de las medidas requeridas para generar el conjunto de datos para la investigación. La habitación es un espacio de 4,42 m 3,23 m, con 4 asientos, por lo cual el espacio está limitado a 3 personas, representando ésta una limitante en la cantidad de ocupantes por muestra (Figura. 1).

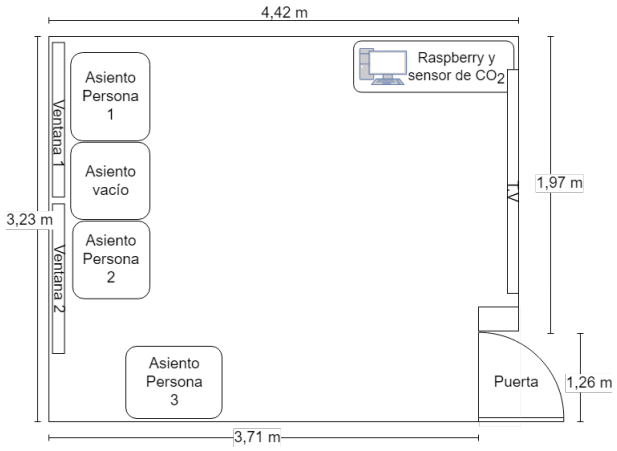


Fig. 1. Croquis de la habitación

*1) **Medidas para la ejecución de la toma de muestras:*** Para las muestras tomadas se realizó un control sobre la habitación, y las personas que estuvieron en la misma, para este se establecieron las siguientes medidas:

- Utilización obligatoria de la mascarilla a lo largo de la toma de la muestra, esto para seguir los lineamientos de seguridad por la pandemia COVID-19.
- La cantidad mínima y máxima es de 0 y 3 personas respectivamente, ya que el personal es limitado al contar con un grupo de 6 personas conformado por 5 investigadores y un profesor. Además, como se explicó la habitación tiene una limitante de máximo de personas debido a su tamaño.
- Las muestras son de 60 minutos.
- Tomando en cuenta que el cerrar o abrir las ventanas que dan al exterior tiene un efecto en los niveles de CO_2 [11], se establece que para tener un control adicional de esta medida las ventanas permanecerán cerradas a los largo del período de toma de la muestra.
- Durante la toma de una muestra la cantidad de personas no varía, es decir, durante los 60 minutos de toma de datos no pueden ingresar ni retirarse personas, esto para que no haya variabilidad en la cantidad de personas y que principalmente las medidas de CO_2 no varíen por su entrada y salida.
- Las personas pueden hablar durante la toma de datos, esto para hacer el experimento lo más realista posible a

un entorno natural donde las personas socializan entre ellas.

- Al surgir la necesidad de realizar muestras seguidas una de otra, en las cuales una misma persona puede formar parte de estas, se establece que antes de empezar a tomar una muestra sé de un periodo de preparación de 10 minutos antes de entrar a la habitación, esto para que haya tiempo para suplir las necesidades básicas de los participantes, como comer e ir al baño.

Debido a que los miembros del grupo tienen distintos horarios de estudio, coincidir para tomar muestras se vuelve un factor limitante en el tamaño del dataset esperado, en consecuencia se establece que este tenga un total de 40 muestras repartidas en grupos de 10.

*2) **Recolección de los datos de la muestra**:* La recolección de los datos sobre la concentración de CO_2 se llevará a cabo mediante un programa en lenguaje Python, dicho software se encarga de leer la información que reporta el sensor cada 60 segundos a lo largo del muestreo durante una hora, para un total de 60 mediciones de CO_2 por muestra. Como instrumento además del sensor y el dispositivo Raspberry, se utiliza el sistema de base de datos MongoDB [12]. Para la carga de los datos en la nube estos se transfieren con cada captura de forma paralela, este procedimiento se realiza de esta manera para eliminar el tiempo que toma realizar la transacción de subida de datos si se realiza de manera secuencial.

La base de datos está implementada por colecciones, la división de las muestras son realizadas por cantidad de personas, dentro de cada una de estas colecciones se guardan las muestras, finalmente dentro de cada muestra se almacenan las mediciones o tomas que se realizan cada minuto, almacenando así la información de CO_2, la fecha y hora y el número de toma para tener una relación de orden de la muestra en cuestión.

Finalmente, el último paso en la adquisición de datos se lleva a cabo en la extracción de la información almacenada por la base de datos a un formato apropiado para que el modelo k-NN lo pueda procesar. Para llevar a cabo la extracción de información se utiliza un programa en lenguaje Python independiente del que corre en el dispositivo de la Raspberry, este se encarga de conectar con la base de datos para extraer la información en archivos csv separados por muestras.

*3) **Análisis Exploratorio de los Datos**:* Para la realización del análisis exploratorio de las muestras estas se tratan de manera independiente una de la otra, no se toma en cuenta si una se hizo seguida de otra, ya que el principal objetivo es comprender si es posible determinar la cantidad de personas respecto a líneas de tiempo de 60 minutos. El CO_2 y el tiempo son las variables independientes de cada muestra. A partir de estas se busca si es posible determinar la variable de respuesta de cantidad de personas, para esto el análisis exploratorio de los datos se divide en dos secciones, la primera trata las muestras sin realizarles alguna alteración, para mostrarlas como fueron capturadas, y en la segunda se cambia la manera en la cuales son tratadas las muestras con el fin de tener una representación distinta que permita analizar los datos desde otra perspectiva.

Análisis de muestras sin alteraciones:

En primera instancia se analizan las muestras separadas en gráficos de línea de tendencia, en las cuales el eje y corresponde al nivel de CO_2 y el eje x al minuto en el cual se obtuvo dicha medida. Este tipo de gráfico permite observar por cada una de estas el patrón de crecimiento o decremento a lo largo del tiempo (60 minutos), lo cual muestra si existen similitudes en sus comportamientos dentro de la misma categoría de n personas.

De manera similar al gráfico descrito anteriormente, se toma la media de las muestras de una misma categoría por minuto, esto permite mostrar en la línea de tendencia el valor promedio de CO_2 a lo largo del tiempo por clasificación de n personas, con lo cual se puede dar una idea de la tendencia en general que poseen los datos.

Análisis de muestras con alteraciones:

El CO_2 puede fluctuar por variaciones en el ambiente [9]. Esto puede provocar que en algunos minutos se generen alteraciones las cuales pueden estar afectando el análisis. Para evitar esto se añade generar líneas de tendencias con saltos o tiempo entre medidas de 2, 5 y 10 minutos. Con esto se busca mostrar si realizar este tipo de variaciones genera que las tendencias tengan menos variabilidad entre un tiempo y otro generando así una línea más comprensible.

Por otra parte, dado que se recomienda que las muestras de concentraciones de CO_2 se tomen a partir de 30 minutos [13], se generan gráficos de línea de tendencia siguiendo esta recomendación, con esto se estaría tomando el tiempo en que se llena la habitación de CO_2 por las personas [13]. De esta manera se aísla el período final de la muestra de forma que quede únicamente la parte más significativa, esto es provechoso para el análisis, ya que permite ver las curvas de crecimiento y observar si existe alguna similitud entre muestras de una misma categoría.

Finalmente, para analizar de forma aislada el comportamiento del CO_2 por minuto, se supone que el anterior minuto es ruido, de manera que se modifican los datos de CO_2 dentro de las muestras restando el CO_2 del minuto siguiente con el anterior, esto provoca que únicamente exista el dato de crecimiento del CO_2 minuto a minuto.

Con esta información se realizan líneas de tendencia de su media para poder observar si los datos tienen un patrón que simplifique el análisis y por ende que pueda facilitar la estimación de la variable de respuesta más adelante en el modelo.

B. Análisis experimental

Continuando con el análisis experimental, a continuación, se procederá a explicar cómo ha sido el proceso de conversión de datos en información y la forma de cómo se interpretaron los mismos para las conclusiones del trabajo. El primer paso es explicar la unidad de muestreo y las características que se obtienen de la misma, creando así el concepto de instancia o vector de características. cuando se tenga una imagen clara de este elemento se procederá a explicar las transformaciones o ingeniería de características que son necesarias realizar sobre

TABLA I
UNIDAD DE MUESTREO

Unidad de Muestreo		
	Fecha y Hora	CO_2
Min 1		
Min 2		
...		
Min 60		
Etiqueta: # de Persona		

dicha unidad de muestreo. Posteriormente, se comenzará a discutir sobre el modelo que se implementó para estimar la cantidad de personas, en este caso k-vecinos más cercanos (por sus siglas en inglés k-NN). Finalmente, se explicará cómo se desea entrenar dicho modelo, buscando un refinamiento en hiperparámetros a partir de *grid-search*, para así finalmente contrastar los mejores modelos contra las métricas de desempeño que se establecieron para realizar una validación cruzada con K iteraciones sobre distintas unidades de testeo.

1) ***Unidad de muestreo****:* Para iniciar con el análisis experimental, es importante indicar qué entendemos por unidad de muestreo en el trabajo, para esto se muestra la Tabla I.

Como se observa, la unidad de muestreo funcionará como una matriz con dos características "fecha y hora" y "CO_2" bruto reportado que se extenderá por 60 filas representadas cada una por un minuto de tiempo que ha acontecido (que suman en total una hora), donde además se agregará una etiqueta que servirá para identificar a dicha matriz según la cantidad de personas.

2) ***Vector de Características y selección de características****:* Antes de presentar las transformaciones respectivas sobre la unidad de muestreo, es importante indicar que a partir de este momento, dicha unidad de muestreo debe de ser entendida como un conjunto de características, llamado de ahora en adelante vector de características o también llamado instancias [14]. Esta definición obedece a que los modelos de predicción ven estas como una sola entidad o instancia que permiten definir a qué clasificación pertenece.

En nuestro caso particular, las características que tendremos consisten en la *"n"* cantidad de CO_2 que se obtiene a partir de los minutos de reporte para unidades de muestra *"j"*, haciendo así que el CO_2, sea una variable *"Xi"* donde *"i"* consiste en el tiempo en que es tomado el CO_2 para una cantidad *"n"* de CO_2 de una instancia particular *"j"*. Haciendo lo mismo que se pueda obtener la representación de una instancia de la siguiente forma:

$$(Instancia/Vector\ de\ Caracterstica)_j = [X_0, X_1, X_2, ..., X_{n-1}]$$

Representación de Instancia / Vector de Características (1)

3) ***Ingeniería de Características****:* Al utilizar modelos de predicción pueden implementarse cambios en las variables para obtener mejores resultados en las predicciones. Estos cambios son conocidos como ingeniería de características, entre estas surgen lo que se conoce como transformación de datos [15], como resumen de este segmento se presenta la Fig. 2 la cual representa los 3 tipos de transformaciones y sus niveles respectivos que se procederán a realizar.

Dichas transformaciones fueron obtenidas a partir del análisis exploratorio realizado en el trabajo, en dicho procedimiento se observó que realizar este tipo de cambios podría cambiar en cómo las medidas de CO_2 muestran su comportamiento. Por lo tanto, se considera que podrían servir en el modelo para obtener una mejor predicción de la cantidad de personas.

Vector de características (Feature Vector)
A. Tiempos de duración d = [30,60]
B. Tiempo entre medidas m = [1,2,3,5,10,15]
C. Tipos de vectores V = [A,B]

Fig. 2. Transformaciones de vector de características y sus niveles

Tiempos de duración *d*:

Los tiempos de duración representan el plazo de tiempo de tomas de CO_2 para una instancia particular que se quiera utilizar. Dicho de otro modo, el plazo de tiempo significa la cantidad de tomas *"n"* realizadas durante una duración *"d"* de minutos antes de predecir la cantidad de personas. Para este estudio se han elegido 2 valores distintos de *"d"*, los cuales son 30 minutos y 60 minutos. La decisión de estos lapsos de tiempo surgen porque se ha detectado en estudios anteriores [10] [16] [17] que para determinar la cantidad de personas a partir de tomas de CO_2 se requiere el total de 60 minutos para predecirlas de forma exitosa, con casos particulares donde incluso con los últimos 30 minutos se alcanza la predicción deseada. En concreto, esto provoca que existan dos tipos distintos de instancias, siendo *k* un índice de valor 30 para obtener la última media hora de una muestra y 0 para las tomas de 60 minutos, lo cual se representa en eq. (2):

$$(Instancia/Vector\ de\ Caractersticas\ con\ d\ min) = [X_{k+0}, X_{k+1}, X_{k+2}, ..., X_{n-1}]\ .\ Donde\ k = \{30, 0\}$$

Instancia con tiempos de duración (2)

Tiempos entre medidas *m*:

El tiempo entre medidas representa cuántos minutos *"m"* pasan entre una toma de CO_2 (la toma de una característica) y la siguiente, su representación se muestra en la eq. (3).

$$(Instancia/Vector\ de\ Caracterstica\ con\ tiempo\ t) = \left[X_{(k+0)m}, X_{(k+1)m}, X_{(k+2)m}, ..., X_{n-1}\right] Donde\ m\ =\ \{1, 2, 3, 5, 10, 15\}\ y\ k = \{30, 0\}$$

Instancia de tipo A con tiempos de duración y tiempo entre medidas (3)

Para nuestro caso particular hemos determinado que lo valores *"m"* posible serán con los niveles 1, 2, 3, 5, 10 y 15 minutos de tiempo entre tomas. Estos han sido elegidos debido a que primeramente los mismos son divisibles entre los tiempos de duración de 30 y 60, volviendo así sencillas las conversiones necesarias, y segundo porque se considera que con tomas con una magnitud más elevada podría distinguirse y clasificarse mejor los vectores de características entre sí, ya que al existir menos características las mismas tienen mayor peso a nivel individual. No obstante, dicha elección podría a su vez perder un mayor número de características que permitirían una mejor clasificación, por lo que se busca obtener ambos enfoques.

Tipos de vectores (A y B):

Finalmente, la última conversión consistirá en una transformación directa sobre los valores reportados de CO_2. Hasta el momento, no se han realizado cambios de los valores de CO_2 respectivos, sino más bien se ha buscado eliminar o excluir por completo ciertos datos, reduciendo así la dimensionalidad que poseen los vectores de características [18]. Para esto es importante indicar que ciertos estudios [10] [16] [17] han identificado que los valores reportados por sensores ambientales, entre los que se destaca el CO_2, tienen como característica particular que para determinar la cantidad de personas puede no ser correcto tomar únicamente los valores brutos reportados de los mismos, sino también los cambios que ocurren entre una toma de una misma instancia y la que le sigue a la misma, obteniendo así diferencias de CO_2, de un tiempo a otro.

Dado esto, lo que hemos realizado consiste en generar un segundo vector B que busque obtener las diferencias entre una toma y su predecesora, es decir, por cada toma $x_{(i+1)m}$ se restaría x_{im}. Dado lo anterior el vector B puede definirse como:

$$(Instancia/Vector\ de\ Caracterstica\ B) = \left[X_{(k+1)m} - X_{(k+0)m}, ..., X_{n-1} - X_{n-2}\right].$$
$$Donde\ m\ =\ \{1, 2, 3, 5, 10, 15\}\ y\ k = \{30, 0\}$$

Instancia de tipo B con tiempos de duración y tiempo entre medidas (4)

Se tiene así que existirían en total 2 tipos de vectores, los cuales representan la forma en que vamos a ver los registros de CO_2, siendo definidos como:

- A: Toma cada registro de CO_2 como el valor bruto de partículas por millón (por sus siglas ppm) de CO_2 del minuto sin realizar ninguna transformación, como se muestra en eq. (3).
- B: Toma cada registro de CO_2 el valor bruto de ppm de CO_2 y calcula cuál es el cambio que hay con la medida del tiempo anterior, como se muestra en eq. (4).

*4) **Modelo K-NN**:* El modelo k-NN [19] puede ser entendido como aquel que clasifica medidas *x* (en el caso particular serán medidas brutas de CO_2 y cambios a través del tiempo de estas) a clases *C* (cantidad de personas) a partir de:

- Determinar los vectores *k* de entrenamiento más cercanos a las medidas *x*, usando una métrica de distancia.
- Clasificar las nuevas *x* a la clase con más representantes dentro del conjunto de *k* de vectores más cercano.

Por lo tanto, lo que se deberá de establecer para este modelo en su forma general consiste en:

- El número de vecinos *k* que se quieren utilizar.
- La métrica de distancia para definir quién es el vecino más cercano.
- Conjunto de datos de entrenamiento.

La elección de este modelo surge porque se desea conocer si es posible obtener una abstracción de lo que se presenta en la Fig. 3. Es decir, se buscará observar valores agrupados de CO_2 durante un período de tiempo y ver si el mismo es posible clasificarlo junto con otros valores agrupados con una misma clasificación (cantidad de personas), lo cual permitirá, por lo tanto, predecir la cantidad de personas que existen en una habitación con medidas de CO_2 completamente desconocidas.

A su vez, es importante recalcar que dicho modelo ya ha sido implementado para la estimación de personas a partir de los autores A. Szczurek et al [10] en donde los mismos reportan resultados de alrededor de 98.4% de exactitud a la hora de predecir la cantidad de personas. A partir de dicho estudio, hemos utilizado sus mismos hiperparámetros donde se utilizó una cantidad de vecinos k = 1 y una distancia métrica de tipo euclidiana.

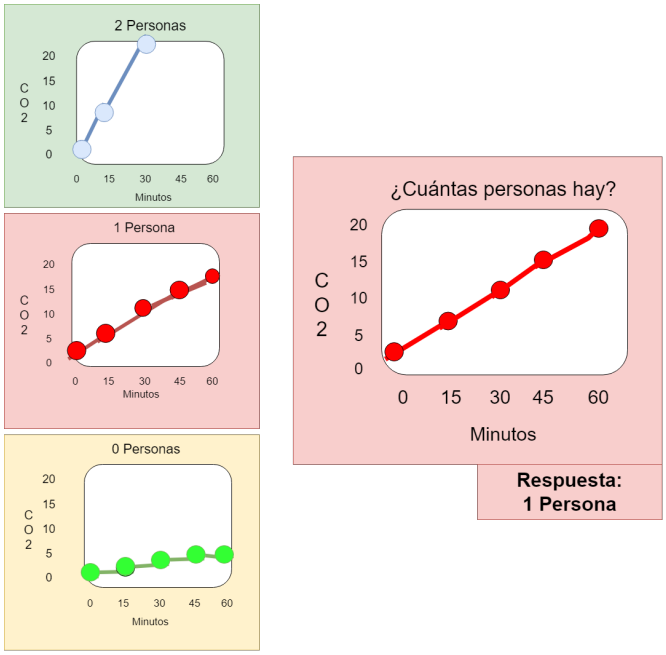


Fig. 3. Representación de k-NN en el ámbito de investigación

*5) **Refinamiento de Hiperparámetros**:* Primeramente, hay que definir un hiperparámetro [20] como aquel parámetro que no puede ser actualizado durante los entrenamientos de modelos, es decir, son aquellos que están relacionados con construir la estructura del modelo y así determinar la eficiencia y precisión en los entrenamientos del mismo. Junto a esto surge el proceso de refinamiento de los hiperparámetros, el cual consiste en un método para realizar la automatización de

estos en búsqueda de un mejor resultado en el momento de entrenamiento.

El modelo k-NN es un modelo instanciado, es decir, los datos de entrenamiento son parte integral del mismo [21], esto significa que la respuesta obtenida se verá directamente afectada sobre el tipo de instancias que utiliza. La situación que se presenta provoca que las mismas instancias y sus características particulares afecten de forma directa entre cada entrenamiento y, por lo tanto, las respuestas obtenidas. Esto, en suma con lo abarcado en la sección de ingeniería de características, provoca que esta misma sea parte esencial del experimento al requerirse como parte de la selección de modelos y no solamente una transformación de estas.

Dado el escenario planteado se pretende realizar una búsqueda en rejilla o *grid search* [20], la cual es definida como un método que efectúa una búsqueda exhaustiva en los hiperparámetros a partir de una medida de desempeño, esto se refiere a que se probarían todas sus combinaciones posibles respecto al tipo de vector, tiempo entre medidas y tiempos de duración, por lo que consistiría en la comprobación de un total de 24 modelos distintos.

Como parte de la comprobación y selección del mejor modelo se utilizará una validación cruzada con K iteraciones o también llamada *K-Fold Cross Validation* [22] a nivel estratificado.

El objetivo de esta técnica de validación es subdividir los datos disponibles en K grupos (en nuestro caso particular K=5) y obtener con K iteraciones un grupo de testeo y varios grupos de entrenamiento, de esta manera analizar qué tan bueno es el comportamiento que presenta el modelo no solo para un set de testeo, sino para los K conjuntos de testeo. Con este método se busca evitar un sesgo a la hora de mencionar que un modelo es incorrecto o correcto por un único conjunto de testeo particular.

A su vez, dado que contamos con una cantidad de 10 tomas por cada clasificación, se ha utilizado una estratificación para asegurarse que dichas clasificaciones sean iguales para cada uno de los conjuntos de testeo, siendo en total 2 muestras por cada una de las categorías, obteniendo así un *k-fold cross-validation* estratificado [22].

Por último, es importante indicar que hemos optado por elegir un total de 60 elecciones distintas de *k fold* de forma aleatoria, es decir, hemos realizado un proceso de *repeated k-fold cross-validation* [23] con el objetivo de reducir el sesgo a la hora de realizar la separación de datos.

Medidas de desempeño del modelo k-NN:

Como parte esencial de determinar el desempeño de los modelos de k-NN se han seleccionado dos medidas de distintas, donde una vez obtenidos los modelos se busca determinar qué tan bueno es su rendimiento en la clasificación de cantidad de personas utilizando los resultados que se obtuvieron en la etapa del refinamiento de hiperparámetros.

- Medida para identificación del mejor modelo - Exactitud (*Accuracy*):
  La exactitud es una métrica para los métodos de clasificación que identifican qué tan bueno es el modelo para predecir n_c clasificaciones correctas sobre un grupo total de *n* muestras.
  A partir de la eq. (5) se define una variable *c* que representa el índice de clase, en nuestro caso la cantidad de personas (c = 0,1,2,3), la variable *C* es el total de clases que existen, n_c son el número de vectores correctamente calificados para la clase *c*, y *n* el total de vectores que fueron comprobados [10]. Básicamente, se toman cuántas muestras de las que se evaluaron fueron acertadas y se divide entre el total de muestras evaluadas, para así obtener el porcentaje de predicciones correctas que obtuvo el modelo.

$$\frac{\sum_{c=1}^{C} n_c}{n}$$

Cálculo de medida de exactitud (5)

- Medida para rendimiento del mejor modelo - Precisión:
  La precisión es una medida de desempeño que busca obtener, como su nombre indica, la precisión con la que un modelo logra predecir de forma acertada una clasificación positiva o correcta una vez que este ha tomado su decisión. A partir de la eq. (6), se puede obtener el cálculo de esta medida, la cual C_p corresponde a las predicciones correctas para una variable *C* y C_f corresponde a las predicciones incorrectas sobre esta misma variable para uno particular. Básicamente, se toman cuántas muestras de las predichas fueron acertadas para una clase particular y se divide entre el total de estas, obteniendo así el porcentaje de predicciones correctas que tuvo el modelo.

$$\frac{C_p}{C_p + C_f}$$

Cálculo de medida de precisión (6)

II. Resultados

A continuación se exponen los resultados obtenidos con el análisis sobre el comportamiento de las muestras respecto a las variables independientes de CO_2 y tiempo de toma con respecto cantidad de personas, y las transformaciones realizadas sobre estas muestras. Después, se exponen los datos obtenidos al entrenar los modelos con el conjunto de datos y las distintas combinaciones de hiperparámetros. Finalmente se muestran los datos obtenidos para el modelo más representativo que se logró a partir del entrenamiento.

A. Muestras recolectadas y transformaciones de los datos

Las muestras recolectadas tienen un comportamiento muy esporádico incluso en las mismas características. La Fig. 4a y Fig. 4b permiten observar el comportamiento dentro de la categoría de cero personas. En la primera el CO_2 disminuye conforme pasa el tiempo, mientras que en la segunda sucede lo contrario, es decir, el CO_2 se mantiene constante y tiene un leve aumento al final de los 60 minutos. De igual forma, la Fig. 4c y Fig. 4d, refieren a la categoría de 3 personas, en donde se observa que el CO_2 de la primera aumenta en los primeros minutos, pero conforme pasa el tiempo, esta medida

se ve disminuida. Por otro lado, en la segunda se observa un aumento en los primeros minutos, sin embargo, disminuye para volver a subir con picos de CO_2 pasados los 30 minutos.

Al comparar la media de cada minuto de las muestras por categoría, se resalta que la clasificación de 2 personas posee un comportamiento similar al de 3 personas, con valores de CO_2 muy cercanos entre ambas clases, manteniendo un patrón que si bien tiene alteraciones donde la medida aumenta y disminuye no parece que varíe mucho a lo largo del tiempo. En la misma representación se puede notar que este comportamiento no ocurre así con la de 1 persona y 0 personas, las cuales se mantienen alejadas de las líneas de 3 y 2 personas. Además, se destaca que mientras que los valores de las muestras de una única persona tienden a subir, los escenarios donde no hay nadie tiende a disminuir la medida de CO_2 en lo que pasa el tiempo. Estos datos se pueden observar en la Fig. 5.

Además, se realizaron tres transformaciones a los datos y se utilizaron de manera conjunta para obtener un nuevo conjunto de entrada para el entrenamiento de los modelos. Estas modificaciones son: tiempos de duración, lapsos entre medidas y tipo de vectores.

Tiempos de duración corresponde al uso de los 60 minutos de las muestras o solo los últimos 30 minutos cada una. Esto con la intención de mostrar si el modelo tiene un mejor desempeño con solo los 30 minutos o si es mejor utilizar los 60 minutos totales.

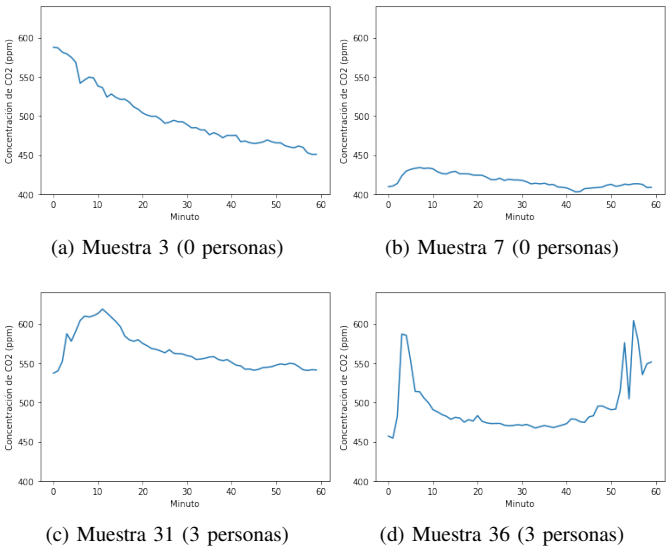
Lapsos entre medidas es utilizar la toma de CO_2 cada cierta cantidad de minutos. Se emplearon lapsos de 1, 2, 3, 5, 10 y 15 minutos. Al hacer esta transformación sobre las muestras se consigue un "suavizado" en su tendencia. Este proceso ayuda a reducir la dimensionalidad de las muestras y puede tener un efecto favorable en el entrenamiento de los modelos.

Por último, se crearon dos tipos distintos de vectores para el conjuntos de datos: uno con las muestras como se recolectaron (vector A) y otro tomando las diferencias por minutos como se muestra en la ecuación (4) (vector b).

El orden en que se aplicaron las transformaciónes para generar las entradas del modelo fue: tipo de vector, tiempos de duración y lapsos entre medidas. Cada set de datos resultante se utilizó para entrenar un modelo de k-NN distinto como se muestra a continuación.

B. Entrenamiento del modelo clasificador k-NN

Como se mencionó anteriormente, se utilizaron las técnicas de optimización de hiperparámetros por rejilla junto con validación cruzada con k iteraciones estratificadas (*stratified k-fold*). Se realizaron 5 particiones del set de datos para el algoritmo de k iteraciones, además de trabajar con 10 ordenamientos aleatorios de los datos, para asegurarse de que los datos estuvieran en orden distinto cada vez que se hicieran las particiones. Los hiperparámetros usados fueron la instancia de datos 1 o 2, los lapsos entre minutos (1, 2, 3, 5, 10 y 15) y la cantidad de minutos (30 o 60). Estas divisiones resultaron en 25 entrenamientos para cada combinación de hiperparámetros. Los datos de precisión por combinación de hiperparámetro se



(a) Muestra 3 (0 personas) (b) Muestra 7 (0 personas)

(c) Muestra 31 (3 personas) (d) Muestra 36 (3 personas)

Fig. 4. Selección de algunas muestras de 0 y 3 personas

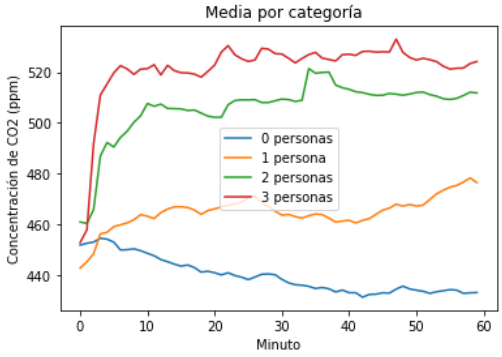


Fig. 5. Media por categoría de las muestras sin alterar los minutos

muestran en la Fig. 6, y el índice de cada combinación en la Tabla II.

Como conjunto, ninguna de las combinaciones logró sobrepasar el 50% de precisión y los valores mínimos y máximos son muy distantes en la mayoría de los casos. Las combinaciones 10 y 18 presentan un promedio muy bajo de precisión, por lo que esta combinación de parámetros no se considera adecuada. Con el proceso de refinamiento de hiperparámetros por rejilla y la validación cruzada, se considera que la combinación con la media más alta representa los parámetros del modelo más robusto. En este caso, la mejor combinación fue la 23, con un 49,75% de precisión. Los modelos 7, 17 y 19 están justo por debajo de esta precisión, con un comportamiento similar al 23, sus precisiones van desde el 10% hasta el 90% a excepción del 7, que tiene valores de 0% de precisión.

Con respecto a la precisión por hiperparámetro, los datos correspondientes a los vectores A y B son las únicas que muestran un comportamiento distinto. En vector A tiene una dispersión muy baja, pero presenta algunos valores atípicos y una asimetría positiva bastante marcada; por otro lado, el vector B tiene una dispersión alta y una mediana más baja

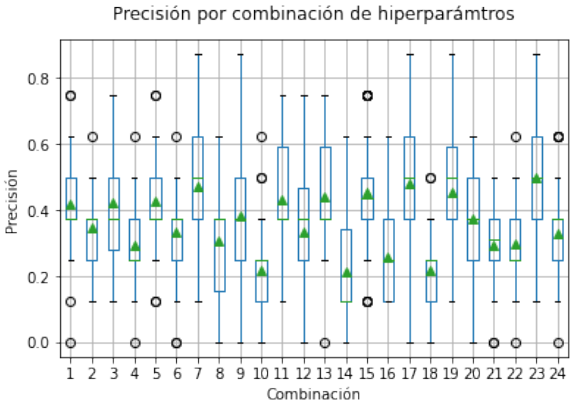


Fig. 6. Estimación de la precisión por combinación de hiperparámetro. La línea verde representa la mediana de los datos, el triángulo verde la media y los puntos son valores atípicos

TABLA II
VALORES DE PRECISIÓN POR COMBINACIÓN DE HIPERPARÁMETROS

Modelo	Hiperparámetro			Exactitud media
	Lapso entre medidas (min)	Tiempos de duración	Tipo de vector	
	m	d	V	
1	1	30	A	42%
2	1	30	B	34.75%
3	1	60	A	34.75%
4	1	60	B	29.5%
5	2	30	A	42.75%
6	2	30	B	33.25%
7	2	60	A	47.25%
8	2	60	B	30.75%
9	3	30	A	38.5%
10	3	30	B	22%
11	3	60	A	43.25%
12	3	60	B	33.5%
13	5	30	A	44.25%
14	5	30	B	21.25%
15	5	60	A	45.25%
16	5	60	B	25.75%
17	10	30	A	48%
18	10	30	B	21.75%
19	10	60	A	45.25%
20	10	60	B	37.5%
21	15	30	A	29.5%
22	15	30	B	29.75%
23	15	60	A	49.75%
24	15	60	B	32.75%

que el vector A. Estas diferencias sugieren que los resultados de precisión se comportan de manera distinta con respecto al tipo de vector. Los valores se pueden observar en la Fig. 6. Por su parte, tanto la transformación de tiempos de duración como lapso entre medidas no presentan una diferencia notable entre los niveles de cada factor.

C. Modelo más representativo

A partir de la optimización de hiperparámetros, con el uso de una validación cruzada con 5 iteraciones, junto con 10 ordenamientos aleatorios de los datos para las iteraciones, se logró conseguir 50 modelos distintos por combinación de

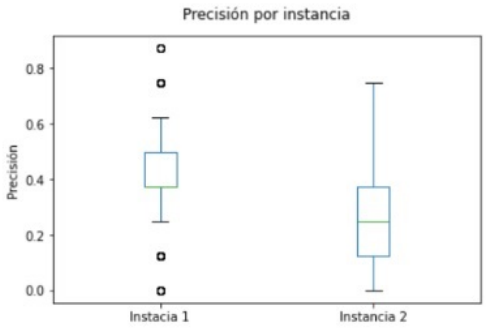


Fig. 7. Precisión de los modelos por tipo de vector. Instancia 1 corresponde al vector A e Instancia 2 al vector B.

TABLA III
PRECISIÓN DE LOS MODELOS PARA LA COMBINACIÓN DE HIPERPARÁMETROS CON MEJOR DESEMPEÑO

Clasificación	Presición (%)		
	Mejor modelo	Modelo intermedio	Peor modelo
0 Personas	100	100	0
1 Persona	100	0	0
2 Personas	50	50	0
3 Personas	100	50	50

hiperparámetros (step, length e instancia). En la sección anterior se menciona que la combinación con mejor rendimiento fue la de salto de 15, 60 muestras en total y la instancia de datos 1. Para mostrar los resultados de esta combinación, se presentan las matrices de confusión de los modelos que se desempeñaron mejor, peor y con un resultado en el intermedio, los cuales se encuentran en Tabla III.

El modelo con mejor desempeño alcanzó un 87.5%. Este modelo clasifica bien todas las categorías menos la de 2 personas, en la cual predice erróneamente una muestra de 2 como una de 1. El modelo de precisión mediana obtuvo un 50% de precisión. Clasificó bien las dos muestras de cero personas, pero solo pudo predecir una de dos personas y una de tres personas. El problema principal fue la predicción de una persona, ya que no logró acertar ninguna de las dos muestras. Otro problema importante es que clasifica una muestra de dos personas como una de cero. Por último, el modelo con el peor desempeño solo logró clasificar una muestra de tres personas correctamente, para una precisión de 12.5%. Lo más representativo de este modelo es que las demás predicciones están a una clasificación de distancia del valor real. Por ejemplo, las predicciones de una persona se clasificaron como cero o dos personas, con la excepción de una de las muestras de cero personas que fue clasificada como dos personas.

III. DISCUSIÓN

Este artículo buscó realizar un modelo de clasificación de tipo k-NN que permitiera a partir de las concentraciones del CO_2 de una habitación determinar la cantidad de personas que existen. Para realizar esto, se realizó una recolección de distintas muestras de mediciones de CO_2 por un período de una hora, donde posteriormente se realizó un tratamiento de

estos datos y la búsqueda de los mejores hiperparámetros para el modelo de clasificación que finalmente fue evaluado, los resultados alcanzados se analizan a continuación.

Primeramente, a nivel descriptivo se observa a partir de la Fig.5 como la clasificación de 0 personas tiene una media que va en decrecimiento, esta situación es resaltable, pues esto puede ser ocasionado ya que cuando se realizaron las tomas de muestras algunas se iniciaban posterior de otras que ya tenían personas adentro, lo cual genera que estas parten con un ruido inicial. Esta tendencia no es similar con las demás categorías donde las mismas muestran un mayor crecimiento durante los primeros 10 minutos para posteriormente mantenerse en valores cercanos.

También se observa como las categorías presentan una posible separación entre sí a la hora de ver sus promedios alrededor de las concentraciones en CO_2, lo que da lugar a pensar que la comprensión brindada en la Fig. 3 alrededor de lo que quiere alcanzar con el modelo k-NN puede ser posible.

Continuando con el análisis descriptivo, se puede distinguir a partir de la Fig. 4 como las muestras obtenidas presentan tendencias distintas entre sí para una misma categoría, lo que permite distinguir la complejidad que puede existir a la hora de lograr el objetivo deseado de estimar la cantidad de personas únicamente con los datos de CO_2, a partir de un modelo de k-NN, al poder existir vecinos que no son parecidos entre sí. Lo anterior, puede llegar a contraponer la teoría vista en estudios anteriores [10].

Por otra parte, en asociación directa al modelo implementado, cuando se analiza los hiperparámetros de la Tabla II parece ser que distintas combinaciones sí afectan la estimación del modelo, por lo que puede existir un campo de estudio con respecto a esta temática, debido a que según se realice el tratamiento del CO_2 se pueden tener resultados distintos como indicaron los estudios [10] [16] [17].

Así mismo, se destaca como la mejor selección de hiperparámetros se obtiene a partir de tiempos entre minutos de 15, con líneas de tiempo de 60 minutos y utilizando el vector con CO_2 bruto (vector A) para alcanzar una exactitud media de 49.75%. Conviene subrayar que los rangos de exactitud de estos hiperparámetros en particular están entre 12.5% y el 87.5%, como se muestra en la Fig 6 por lo que estos a pesar de ser los mejores poseen una variabilidad considerable.

Debido a la variabilidad existente en los modelos obtenidos, se ha procedido a obtener sus resultados de precisión por clasificación, a aquellos modelos con los hiperparámetros que obtuvieron el máximo, mínimo y mediana en exactitud a la hora de realizar la validación cruzada estratificada, obteniendo así modelos que obtuvieron una precisión de 10%, 49.75%, 90% respectivamente.

Además, se aprecia y se confirma que estos no son perfectos y resultan heterogéneos entre sí, no obstante, el que tuvo mejor desempeño es capaz de generar clasificaciones bastante concisas y donde los otros modelos obtenidos también tienen aspectos a destacar. El mejor modelo, mostrado en la Tabla III, solo erró una clasificación, mientras que otros dos modelos, aunque obtuvieron mucha menos precisión, obtienen clasificaciones que son cercanas al valor real que debieron clasificar, con muy pocas excepciones a esto último.

A partir de estos modelos obtenidos y al existir estudios que sí lograron obtener resultados deseados utilizando un modelo de k-NN con habitaciones con una mayor extensión y cantidad de personas, queda pendiente así analizar la escalabilidad de este modelo y ver si el mismo tiene una situación que logra obtener una mejor precisión bajo estas circunstancias y donde al reducir estas características el modelo se ve afectado.

Por otra parte, como otra futura investigación se denota que con los datos obtenidos podrían ayudar a identificar si existen diferencias de CO_2 al realizar tomas de diferentes épocas del año, como a su vez de sus efectos producidos cuando se toman con otro tipo de población, todo lo anterior con objetivo de identificar posibles sesgos o el descarte de estos para la implementación de futuros modelos de predicción.

Relacionado a futuros modelos de predicción, se destaca también que pudieron haberse utilizando otros con menor interpretabilidad, pero con mayor precisión, como redes neuronales y máquinas de soporte vectorial, para ver si estos realizaban una clasificación más consistente y exacta. A pesar de esto, la transformación de las variables se denota que sí tuvieron un efecto para aumentar la precisión del modelo de k-NN, esto, junto con técnicas de reducción de dimensionalidad podrían usarse para obtener mejores resultados en futuras investigaciones.

Como parte de las limitaciones de este experimento se destaca la poca cantidad de muestras que se pudieron recolectar para entrenar el modelo. Con 40 muestras es difícil dividir el conjunto de datos para realizar entrenamiento y pruebas, como es lo usual al entrenar modelos de aprendizaje mecánico.

Debido a esta limitación, no se pudo hacer una partición para validación, por lo que, a pesar de intentar reducir el ruido y sesgo sobre los modelo se consideran que estos no son tan robustos como lo hubiera sido si se hubiera contado con más muestras, y puede que una cantidad mayor de estas podría ayudar a obtener mejores resultados en cuanto a precisión.

IV. Conclusiones

A partir del desarrollo de la investigación se destaca que el modelo k-NN obtenido no alcanza un grado de certeza deseable, dado que a pesar de tener un mejor modelo con 87.5% de exactitud, el mismo con otros escenarios de testeo tiene otros resultados con una menor exactitud.

A pesar de que se considera que los resultados obtenidos puedan no ser robustos, debido a la cantidad de muestras, como lecciones aprendidas se destaca la implementación del uso de transformaciones sobre las concentraciones de CO_2 pudieran lograr un efecto distinto sobre la clasificación.

Finalmente, destacamos como un último aprendizaje el haber encontrado que el modelo k-NN a pesar de presentarse con buenos resultados para la estimación de personas en otros escenarios, para el escenario implementado no fue así, lo que permite preguntarse si este modelo puede padecer de problemas de escalabilidad para el problema de determinar la cantidad de personas a partir de mediciones de CO_2.

References

[1] K. Zhou and S. Yang, "5.11 smart energy management," in *Comprehensive Energy Systems*, pp. 423–456, Elsevier, 2018.

[2] S. D'Oca, T. Hong, and J. Langevin, "The human dimensions of energy use in buildings: A review," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 731–742, 2018.

[3] D. Yan, W. O'Brien, T. Hong, X. Feng, H. B. Gunay, F. Tahmasebi, and A. Mahdavi, "Occupant behavior modeling for building performance simulation: Current state and future challenges," *Energy and buildings*, vol. 107, pp. 264–278, 2015.

[4] H. Yoshino, T. Hong, and N. Nord, "Iea ebc annex 53: Total energy use in buildings—analysis and evaluation methods," *Energy and Buildings*, vol. 152, pp. 124–136, 2017.

[5] T. A. Nguyen and M. Aiello, "Energy intelligent buildings based on user activity: A survey," *Energy and buildings*, vol. 56, pp. 244–257, 2013.

[6] Z. Chen, C. Jiang, and L. Xie, "Building occupancy estimation and detection: A review," *Energy and Buildings*, vol. 169, pp. 260–270, June 2018.

[7] P.-R. Tsou, C.-E. Wu, Y.-R. Chen, Y.-T. Ho, J.-K. Chang, and H.-P. Tsai, "Counting people by using convolutional neural network and a PIR array," in *2020 21st IEEE International Conference on Mobile Data Management (MDM)*, IEEE, June 2020.

[8] M. Zuraimi, A. Pantazaras, K. Chaturvedi, J. Yang, K. Tham, and S. Lee, "Predicting occupancy counts using physical and statistical co2-based modeling methodologies," *Building and Environment*, vol. 123, pp. 517–528, Oct. 2017.

[9] S. Meyn, A. Surana, Y. Lin, S. M. Oggianu, S. Narayanan, and T. A. Frewen, "A sensor-utility-network method for estimation of occupancy in buildings," in *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, pp. 1494–1500, 2009.

[10] A. Szczurek, M. Maciejewska, and T. Pietrucha, "Occupancy determination based on time series of CO2 concentration, temperature and relative humidity," *Energy and Buildings*, vol. 147, pp. 142–154, July 2017.

[11] S. S. Korsavi, R. V. Jones, and A. Fuertes, "Operations on windows and external doors in uk primary schools and their effects on indoor environmental quality," *Building and Environment*, vol. 207, p. 108416, 2022.

[12] "Build faster. build smarter.."

[13] A. Szczurek, M. Maciejewska, A. Wyłomańska, R. Zimroz, G. Żak, and A. Dolega, "Detection of occupancy profile based on carbon dioxide concentration pattern matching," *Measurement*, vol. 93, pp. 265–271, 2016.

[14] Y. Fan, L. Raphael, and M. Kon, "Feature vector regularization in machine learning," *arXiv preprint arXiv:1212.4569*, 2012.

[15] A. Zheng and A. Casari, *Feature engineering for machine learning: principles and techniques for data scientists*. " O'Reilly Media, Inc.", 2018.

[16] A. Szczurek, M. Maciejewska, M. Teuerle, and A. Wyłomańska, "Method to characterize collective impact of factors on indoor air," *Physica A: Statistical Mechanics and its Applications*, vol. 420, pp. 190–199, 2015.

[17] A. Szczurek, M. Maciejewska, R. Połoczański, M. Teuerle, and A. Wyłomańska, "Dynamics of carbon dioxide concentration in indoor air," *Stochastic Environmental Research and Risk Assessment*, vol. 29, no. 8, pp. 2193–2199, 2015.

[18] L. Van Der Maaten, E. Postma, J. Van den Herik, *et al.*, "Dimensionality reduction: a comparative," *J Mach Learn Res*, vol. 10, no. 66-71, p. 13, 2009.

[19] B. V. Dasarathy, "Nearest neighbor (nn) norms: Nn pattern classification techniques," *IEEE Computer Society Tutorial*, 1991.

[20] T. Yu and H. Zhu, "Hyper-parameter optimization: A review of algorithms and applications," *arXiv preprint arXiv:2003.05689*, 2020.

[21] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine Learning*, vol. 6, no. 1, p. 37–66, 1991.

[22] D. Berrar, "Cross-validation.," 2019.

[23] M. Kuhn, K. Johnson, *et al.*, *Applied predictive modeling*, vol. 26. Springer, 2013.