**Metodología y Materiales**

The current research involved recollecting environmental data from two different locations to generate two datasets that can be analyzed using machine learning techniques to predict the occupancy levels. The datasets obtained corresponds to a gym and a home

**Dispositivo de Recolección**

Para la recolección de datos se diseñó un dispositivo conformado por un sensor ambiental BME280 [1] de la marca Bosh y un controlador ESP32 de la marca Espressif Systems [2]. Los datos recolectados se enviaron a una base datos en la nube para su almacenamiento y posterior análisis. En la imagen [1] se presenta el diagrama de conexiones del dispositivo.

El sensor BME280[1] es un sensor de bajo costo, bajo consumo energético y diseñado para el monitoreo meteorológico, el cual permite medir la humedad relativa (%), la temperatura (C), presión atmosférica (hPa) y altitud () del espacio en el que se encuentra. Estas características lo hacen adecuado para los fines de esta investigación. Los rangos y exactitud del sensor se listan a continuación: presión 300 hPa a 1100 hPa (+- 1 hPa exactitud), temperatura de -40 C a 85 C (+-1 C exactitud), humedad relativa de 0% a 100% (+- 3% exactitud), y altitud de 0 ft a 30000 ft (+- 1 metro exactitud). Para más información acerca de la sensibilidad del sensor y su error, se recomienda revisar su hora técnica el cual se puede consultar en [1].

**Data Recollection**

La recolección de datos se llevo acabo en dos ubicaciones: un gimnasio del ITESM y un hogar residencial. Inicialmente se planeó que las dos recolecciones se fueran en el gimnasio, ya que se quería que la segunda recolección fuera una mejora en la observación del ground truth, sin embargo, tras declararse la pandemia de COVID 19 a finales de marzo, todos los espacios del ITESM fueron cerrados por lo que fue necesario realizar la segunda medición en un espacio diferente. Dadas las circunstancias sanitarias globales y la cuarentena obligatoria en México, se optó por realizar la recolección en la sala del hogar de uno de los asistentes encargados de la recolección.

*Gimnasio*

El espacio monitoreado corresponde al gimnasio de empleados del ITESM campus Monterrey en Monterrey, NL, el cual cuenta con climatización y ventilación mecánica. La recolección se llevo a cabo durante seis días entre 18 de septiembre al 02 de octubre del 2019. Los datos se midieron en sesiones de aproximadamente de 20 minutos en tres horarios distintos: mañana, tarde, noche. Dichos horarios corresponden a los períodos de mayor transito de personas en el gimnasio por lo que representa un escenario adecuado para la investigación. La resolución de las mediciones fue de 1 segundo obteniendo al final aproximadamente 10k registros correspondientes a aproximadamente 180 minutos. Debido a la distribución del gimnasio y los recursos disponibles no fue posible utilizar cámaras o sensores auxiliares para el registro de la ocupación, por lo que esta tuvo que ser recabada manualmente por el encargado de la recolección. Una debilidad adicional es que no se registró el número de personas observado si no el nivel de ocupación directamente, según fue percibido por el encargado de la recolección. Esto puede considerarse una fuente de variación adicional en los datos.

*Living room*

El espacio monitoreado corresponde al living room de un edificio residencial. El área monitoreada es aproximadamente de 8m x 4m y cuenta con ventilación mecánica: un minisplit, un ventilador de techo y en algunos momentos un segundo ventilador de piso. El espacio carece de ventanas y por lo que no cuenta con ventilación natural. El sensor se colocó entre la sala y el comedor, aproximadamente a unos 5 metros del minisplit, esto con la finalidad de evitar que el sensor tenga un impacto directo con el minisplit en caso de ser encendido. La recolección se llevó acabo durante 11 días entre el 14 de mayo y el 4 de junio del 2020, en sesiones con duración promedio de 12 horas, aunque algunas sesiones duraron 24 horas, por lo que cuenta con periodos de inactividad correspondientes a la madrugada. La resolución de la medición fue de 1 segundo, por lo que al final se obtuvieron aproximadamente 300k registros correspondientes a 5000 minutos. Para la obtención del ground truth se utilizó una cámara para tomar fotos cada 10 min y posteriormente obtener la ocupación. Esto es una ventaja respecto al dataset del gimnasio, al ser un valor objetivo de la ocupación. La ocupación máxima registrada fue de 7 personas y la mínima 0.

**Preprocesamiento**

Los datasets fueron exportados de su almacenamiento en la nube como archivos CSV. Para el caso del living room, cada día de medición se encontraba en un archivo distinto, y adicionalmente el ground truth de toda la recolección se encontraba también en un archivo separado. Al realizar el cross de los archivos, se encontraron pequeños periodos donde se contaba con mediciones de datos pero no el ground truth, esto debido a detalles técnicos durante la recolección (i. e. cambio de memoria en la cámara). Dichos períodos fueron una minoría que representaban menos del 1% de los datos. Al ser pocos registros se decidió descartarlos. Igualmente, al revisar ambos datasets en busca de valores perdidos, se encontró únicamente 6 mediciones con valores faltantes en el dataset del living room. Dichos valores se completaron usando los valores promedio del valor anterior y el valor posterior. Esto se consideró adecuado debido a la alta resolución del dataset comparado con trabajos anteriores que utilizaron resoluciones menores (1 minuto o más) [adeogun, zemoury, Chitu,Kumar ]. Finalmente se descartó de los datasets finales la información de la altitud. Esto debido a que la altitud no tiene relación con el problema de la ocupación por lo que carece de sentido dejar dicha información. Adicionalmente, de acuerdo con la hoja técnica del sensor BM280, la información de la altitud se calcula usando las mediciones de la presión, por lo que la capacidad discriminante que esta variable pudiera tener, se encontraría contenida en la variable de presión. Esto se comprobó mediante un análisis de correlación Pearson en el cual ambas variables mostraron una correlación positiva perfecta (1.0).

*Establecimiento de los niveles de ocupación*

Como mencionamos anteriormente, en el caso del gimnasio, los niveles de ocupación fueron registrados directamente por el encargado de la recolección de los datos. Dichos niveles son alto, medio y bajo. No se cuenta con el nivel correspondiente a vacío, sin embargo, dado que la recolección se realizó en los horarios de mayor transito de personas, es poco probable que existiera un período vacío. Al no contar con el rango de personas correspondiente a cada nivel, el valor de la ocupación es considerado subjetivo.

Para el caso del living room, como mencionamos anteriormente la ocupación se obtuvo de fotografías en intervalos de 10 minutos. Los niveles de ocupación se establecieron siguiendo el procedimiento descrito en [4], y quedaron de la siguiente manera: vacío (0), bajo (1-2), medio (3, 5), y alto (5-7).

*Generacion de datasets adicionales*

Adicionamente, para cada dataset se seconsideraron las siguientes resoluciones alternativas: 10 segundos, 30 segundos, 1 minuto, y 5 minutos. Por cada resolución se generaron dos versiones: una versión tomando únicamente una muestra, y las otras promediando las muestras dentro de cada período. En total se obtuvieron 9 datasets correspondientes al gymansio, y 9 datasets correspondientes al livingroom. En los datasets promediados, por cada feature (humedad, presión y temperatura) se agregaron dos features extras, correspondientes a la kurtosis y la desviación estándar, quedando estos finalmente con 9 features.

*Partición y Balanceo*

Cada dataset fue particionado usando un 20% para pruebas y el resto para entrenamiento. Adicionalmente, se encontró que los datasets se encontraban desbalanceados por lo que se realizó un balanceo utilizando el algoritmo ADASYN para realizar un oversampling. En la tabla [tabla 1] se compara el tamaño de los datasets de entrenamiento antes y después de ser balanceados.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Before** | | | | **After** | | | |
| **Dataset** | Empty | Low | Medium | High | Empty | Low | Medium | High |
| Living room 10 sec | 4098 | 16325 | 2830 | 547 | 16353 | 16325 | 16330 | 16329 |
| Living room 30 sec | 1345 | 5450 | 956 | 188 | 5445 | 5450 | 5436 | 5452 |
| Living room 1 min sec | 673 | 2730 | 480 | 92 | 2720 | 2730 | 2720 | 2732 |
| Living room 5 min sec | 136 | 553 | 96 | 19 | 556 | 553 | 555 | 554 |
| Gym 10 sec | NA | 202 | 430 | 189 | NA | 430 | 430 | 430 |
| Gym 30 sec | NA | 149 | 149 | 59 | NA | 149 | 149 | 150 |
| Gym 1 min sec | NA | 33 | 77 | 34 | NA | 77 | 77 | 77 |
| Gym 5 min sec | NA | 21 | 9 | 5 | NA | 21 | 21 | 21 |

**Modelos**

Decision Trees

K-NN

SVM

**Experimentos**

Se diseñaron dos experimentos con 3 objetivos en mente: i) hallar el subconjunto de features óptimo para la predicción de la ocupación, ii)

Features Selectoin

<>

Living Room

<>

Resultados y Discusión

Gimnasio

Living Room

Tablas

[1] Resultados del balanceo

Imágenes

[1] Diagrama del circuito

[2] Diagrama del Gimnasio

[3] Diagrama del Livingroom

Referencias

[1] https://www.bosch-sensortec.com/products/environmental-sensors/humidity-sensors-bme280/

[2] <https://www.espressif.com/en/products/socs/esp32/overview>

[3] Referencia a trabajos anteriores con resoluciones mayores

[4] Yuan 2020