Source of Dataset: https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+

Description of Dataset:

It is an experimental dataset used to detect room occupancy; it was obtained from time stamped pictures that were taken every minute. There are total seven variables in the dataset out of which one is a target variable. List of all the variables can be found in data dictionary.

Underlying real world problem:

Conserving energy produces a higher quality of life and reduced emissions result in cleaner air quality so it is very important to save energy and it has been estimated that 30 to 42% of energy can be saved by accurate determination of occupancy detection in buildings. This occupancy data is then used as an input to HVAC control algorithms, which then distribute temperature and zone occupancy information to reduce energy consumption while in the meantime preserving the comfort condition of the occupants. Nowadays, with the affordability of sensors increasing and becoming more ubiquitous, together with affordable computing power for automation systems it makes the determination of occupancy a very promising approach to lower energy consumption by appropriate control of HVAC and lighting systems in buildings.

Other applications for occupancy detection include security and determination of building occupant behaviors. A system that could accurately detect the presence of the occupants without using a camera is very interesting due to privacy concerns.

Data Dictionary:

Column	**Type / format**	**Description**
Date and time	m/d/yyyy h:mm:ss	Date and time at which the record was taken
Temperature	Celsius	Temperature of the room
Humidity	%	Relative Humidity in the room
Light	Lux	Intensity of light in the room
CO2	ppm	Level of CO2 in the room
Humidity ratio	kgwater-vapor/kg-air	Derived quantity from temperature and relative humidity
Occupancy (target variable)	0 or 1	0 for not occupied, 1 for occupied status

Note: I will be using combination of all or the subset of the above mentioned features/variables and determine which combination classifies most accurately. Additionally, there are no missing values in the dataset.