Impacts of weather variation on air quality

Report

by

Ahmad Opeyemi ZUBAIR
Ismam BIN HASNAT

January 29, 2023

1 Introduction

Air quality in residential places is a growing concern, as it can have significant effects on the health and well-being of individuals living in these areas. Exposure to Poor air quality can lead to a variety of health complications, including respiratory problems, headaches, and fatigue. In addition, it can also contribute to the deterioration of indoor surfaces and possessions. Therefore, it is essential that steps are taken to ensure good air quality in residential spaces.

One of the primary sources of poor air quality in residential spaces is indoor air pollution. This can be caused by a variety of factors, including the presence of mold, radon, and volatile organic compounds (VOCs) emitted by certain building materials and household products. Additionally, open-air heating systems, faulty filters, and poor ventilation can also contribute to poor air quality, as it can trap pollutants such as carbon monoxide (CO) and carbon dioxide (CO_2) inside the home.

Outdoor air pollution, on the other hand, refers to the presence of pollutants in the air outside of buildings and homes. The main sources of outdoor air pollution are industrial and transportation emissions, as well as natural sources such as wildfires and dust storms. Urban areas are particularly affected by outdoor air pollution due to the concentration of industrial and transportation activity.

Both indoor and outdoor air pollution are significant environmental issues that have a negative impact on human health and the environment. While they have different sources and effects, it is important to address both types of air pollution in order to ensure clean and healthy air for all individuals. The upper limit of good air quality considered by most European countries is 1000 parts per million (ppm) [1]. Values between 1,000 - 2,000 ppm are associated with poor air quality and may cause drowsiness. The classes of air quality control and their corresponding effects are given in Figure 1 below [2].

~400 ppm	background (normal) outdoor air levels
400- 1,000 ppm	typical levels found in occupied spaces with good air exchange
1,000 – 2,000 ppm	levels associated with complaints of drowsiness and poor air
2,000 – 5,000 ppm	levels associated with headaches, sleepiness, and stagnant, stale, stuffy air Poor concentration, loss of attention, increased heart rate and slight nausea may also be present
>5,000 ppm	Exposure may lead to serious oxygen deprivation symptoms

Figure 1: CO_2 levels and their corresponding health effects

It is common knowledge that residents in a house are likely to close their windows and turn on the heating system during winter resulting in poor ventilation and buildup of CO_2 levels. The opposite is true during the summer and it is expected that the CO_2 levels will be significantly less during this period.

This project aims to investigate if such a relationship exists between weather and indoor air quality using sensor data from a smart home. Subsequently, we will use three machine learning (ML) methods: Linear Regression, Random Forest Regression, and Multi-Layer Perceptron (MLP) Regressor to determine if weather variables can be used to predict the level of air quality in a residence.

2 Data Collection

The data for this project was obtained from the smart home's database which was shared with us. Based on the website "EXPE-Smarthouse", the structure of the house is understood as well as the sensors' positions. That way, it was decided which sort of data should be considered in order to reach the main objective. In Figure 2, the sensors deployed within the house are depicted,

Figure 2: Sensors deployed in the smart house

From the figure, it can be observed that there are numerous sensors in different locations, inside or outside the house. For air quality measurement, we collect CO_2 sensor data from the main bedroom on the second floor, as well as the living room. Weather parameters are also obtained from the weather station sensor. In addition, the boiler sensor data was obtained to see if there was any correlation between the indoor heating system and the CO_2 levels in the house. Table 2 below shows the list of sensors used in this project. The data from each sensor was obtained in a CSV format.

Bedroom	Living Room	Weather Station	Boiler
CO_2	CO_2	Rain	Heater
Windows	Sound	Humidity	Hot water
Temperature/humidity		Temperature	
		Luminosity	
		Wind	

Table 1: Sensors data from bedroom, living room, weather station and boiler

The data from the five sensors at the weather station will be used to build the features matrix while the CO_2 sensor data from the bedroom and living room will be the labels. Therefore, our prediction problem will be a multivariate regression problem.

3 Data preprocessing

As is often found in machine learning projects, the dataset used in this project needed some processing before it could be used for our purposes. The first problem was the fact that most of the sensors stored the timestamps in a *string* format. Converting these values into a *DateTime* format would make analysis and visualization much easier. Therefore, we needed to implement some data transformation code.

Next, we faced the problem of dissimilar timestamps. Because each sensor recorded data at its own unique frequency, we needed to resample the data to obtain uniform timestamps across all sensors. We chose a frequency of 10 minutes when resampling the data and interpolated missing data using the mean of its two neighbors. The solutions to these two problems are implemented in a function named *import_data* which is called while importing the CSV files.

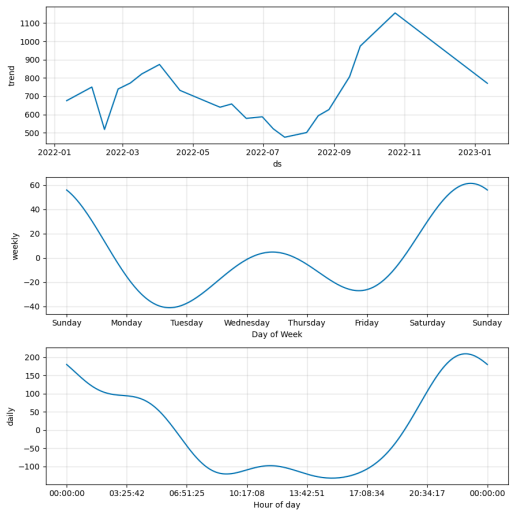
Finally, a preliminary exploration of the data showed that there was a significant difference in magnitude in our data as shown by Figure 3 below. This could influence the training process of regression models such as linear regression by assigning higher weights to the variables with larger magnitudes. Therefore, we employ a *MinMaxScaler* from *scikit-learn* to scale the values in each column between 0 and 1. This prevents any large variable from skewing the training process.

time	wind	humidity	luminosity	rain	temperature	bed_co2	living_co2
2022-01-11 15:10:00+00:00	1.000000	60.730000	1278.157000	0.000000	6.680000	788.0	900.0
2022-01-11 15:20:00+00:00	2.000000	60.270000	959.032222	0.000121	6.675000	766.0	928.0
2022-01-11 15:30:00+00:00	2.105263	59.810000	768.504000	0.000241	6.670000	753.5	931.0
2022-01-11 15:40:00+00:00	2.210526	60.513333	482.956000	0.000362	6.624167	741.0	934.0
2022-01-11 15:50:00+00:00	2.315789	61.216667	207.961000	0.000482	6.578333	743.0	956.0

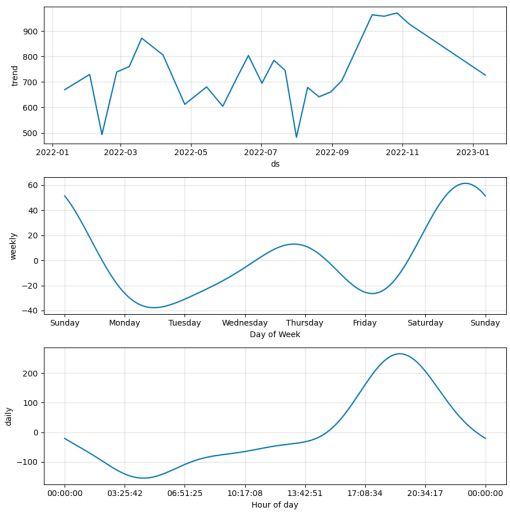
Figure 3: Matrix containing features and labels

4 Exploratory Data Analysis

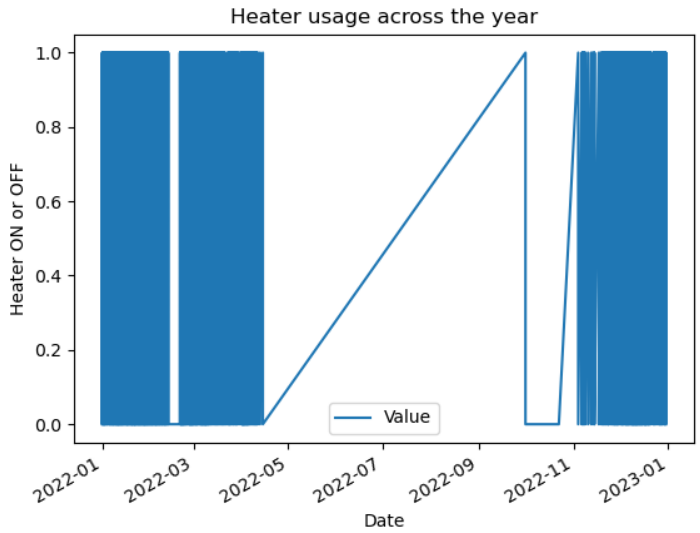
In this section, we perform some visualization and preliminary analysis of our input data. First, we analyze how the CO_2 levels in both rooms vary across the seasons to observe any visible trends. To do this, we perform a time series decomposition (TSD) on the bedroom and living room CO_2 data. The results are as shown in Figures 4a and 4b below.



(a) TSD of the bedroom data



(b) TSD of the living room data



(c) Heater usage over the year

Figure 4: Investigation of CO_2 and heating trends across the year

It can be observed that the CO_2 levels increase significantly during the winter reaching peak values around November. Conversely, the CO_2 levels decrease consistently during the summer reaching their lowest values in the month of August. Furthermore, the bimodal shape of the daily (peaks in the morning and night) and weekly seasonality (peaks during the weekend) indicate that the data was collected in a residential space which confirms prior knowledge. This lends some weight to the assumption that there is

some underlying correlation between weather variation and the air quality in the smart home.

We also analyze the use of the indoor heating system in the house as shown in Figure 4c. It is immediately obvious that the heating system is mainly used during the winter period and is turned off during the summer between the months of May and November. This seasonality also supports the assumption that the use of the heating system has some influence on the level of CO_2 present in the house.

5 Model Implementation

Based on the Exploratory Data Analysis, it was seen that there are noticeable correlations between the CO_2 level and the outdoor temperature or weather. This correlation can be used to train and test the data for machine learning and then further implement it for predictive purposes.

3 different regression models have been used to test which model is the best suited for predictive modeling. As mentioned before, the three models are Linear Regression, Random Forest Regressor, and MLP Regressor. Linear regression is one of the most common models used in machine learning while working with continuous data. For the same reason, Random Forest Regressor has been considered as a model as well. To adopt a more sophisticated approach, a neural network model - MLP (Multi-Layer Perceptron) Regressor has been included as the third choice of model. In table 2, the above-mentioned models used with their corresponding accuracy level & error level is presented.

	Linear Regression	Random Forest Regressor	MLP Regressor
Accuracy	57.36%	99.9%	67.03%
Mean Absolute Error (MAE)	0.05	0.0005	0.042

Table 2: Machine Learning model implementation

As seen from the table above, the accuracy level of the Random Fores Regressor is stunning 99.9% with a very low Mean Absolute Error (MAE) of 0.0005. The Linear Regression model has about 57.36% accuracy which is considerably low, with a comparatively significant MAE of 0.05. The MLP Regressor model also provides a mediocre accuracy of 67.03% and a considerable MAE of 0.042. Although the iteration level of the MLP regressor has been set to a high value (1000 iterations), the accuracy level did not change significantly. Even after changing the value of the split size of the test-train dataset, the models have been re-run, which also give similar results.

After training and testing of datasets with 3 different machine learning models, as the Random Forest Regressor seems to provide the near-perfect result for predictive analysis, this model has been chosen for further implementation. Furthermore, using the Random Forest Regressor, we also determine the most important weather variable which influences the indoor air quality as shown in Figure 5. It can be observed that the external temperature is the most important factor influencing the amount of CO_2 in the house. During winter, when temperatures fall, the residents are likely to turn on the heating

system and close their windows which leads to a build up of the CO_2 levels. Conversely, during summer, the heater is off and the windows are likely opened which prevents the build up of CO_2 in the house.

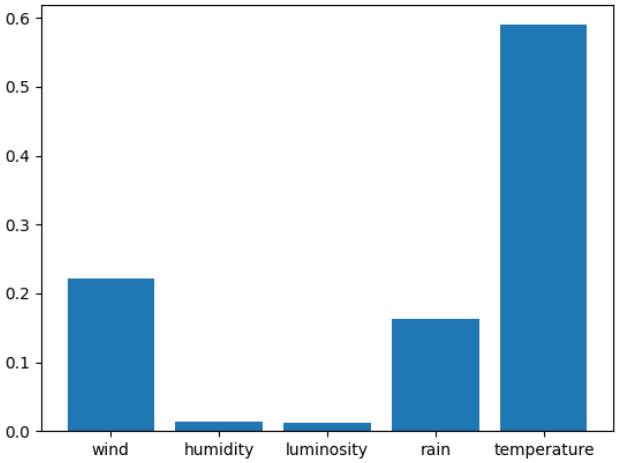


Figure 5: Feature importance

6 Conclusion

This mini-project has been a reflection of the practical world of Machine Learning Implementation. From the data collection to implementation, there were numerous real-world criteria to be considered. CO or CO_2 level in the residences is a critical factor to take into consideration in order to assure the healthiest environment possible. In the course of the project, we have observed the relationship between weather variables and air quality by developing a model capable of using the former to predict the latter. The results also show that temperature is the most important factor influencing the level of CO_2 in a house.

Based on this prediction model, some future works could be suggested, such as, predicting and scheduling the window opening and Closing or turning on/off the ventilation. This will influence the energy demand due to energy loss for window openings or energy required for ventilation. Therefore, such predictions can in turn contribute to predicting the corresponding energy demand.

References

[1] M. Griffiths and M. Eftekhari, "Control of co2 in a naturally ventilated classroom," *Energy and Buildings*, vol. 40, no. 4, pp. 556–560, 2008.

[2] CO2meter.com, "Typical co2 levels at home test — co2meter.com." https://www.co2meter.com/blogs/news/co2-levels-at-home, Oct 2022.