A Bayesian Model for Optimizing Thermal Comfort and Indoor Air Quality

Merve Akdoğan[1], Sema Alaçam[1], Behçet Uğur Töreyin[1]

[1] Istanbul Technical University, Istanbul, Türkiye
merve.akdogan@itu.edu.tr; alacams@itu.edu.tr; toreyin@itu.edu.tr

Abstract. This study focuses on the usage of a probabilistic approach on determining the best course of action in a specific environment within the domain of architecture. More specifically, Bayesian decision theory is applied on a simplified problem of maintaining thermal comfort and air quality. An already existing comprehensive dataset is used and narrowed down for the purpose of the study. Environment measurements (indoor and outdoor temperature, indoor CO_2 level and air humidity) are taken as input variables and user preferences (open or closed window) are taken as outputs in order to address the problem as a binary classification problem. The paper can be regarded as a preliminary study on the usage of probabilistic approaches in the discipline of architecture.

Keywords: Predictive Modeling, Binary Classification, Bayesian Decision Theory, Occupant-Building Interaction, Thermal Comfort.

1 Introduction

Rapid developments in the field of artificial intelligence affect almost every discipline including the architecture and create new potentials. In every discipline, it is possible to foresee that areas that can work autonomously will gradually be left to machines and algorithms. Likewise, in the discipline of architecture, using, processing, and recalling big amounts of data emerged as a growing body of interest in the last decades. Computational approaches and methods such as data-driven design, data-informed decision making, data mining, data visualization became influential on the promoting use of data in design and optimization, also automatizing repetitive and iterate processes. These systems can be utilized in the design, production and also management processes, such as maintaining the occupants' comfort in buildings which can aim to optimize thermal comfort and air quality.

Examining occupants and building interactions is essential to optimize energy usage in buildings and life quality of humans. Moreover, energy usage and occupants’ comfort preferences directly affect building design and management processes. One of the primary parameters for energy usage and comfort is heat. Thermal features of the environment directly affect the occupant behavior and result in some actions like changing the clothing, opening or closing the windows / doors, operating HVAC systems etc. Occupants’ thermally adaptive behaviors in order to maintain their comfort can be recorded and used as a dataset and used for informed designs and management purposes. Studies on user and building interactions with statistical, computational or machine learning methods have been carried out by creating data sets on user behavior and environment interaction in different contexts. In example, Langevin et al. (2015) worked on an office building in the USA. They recorded a range of internal-external thermal effects and associated behaviors. With this data set, they created an agent-based model of the behavior of the building occupants. Mahdavi et al. (2019) collected data on user preferences and light/heat/humidity measurements in a university building and made a statistical analysis. Paige et al. (2019) focused on collecting data on residents' behavior and energy use. He et al. (2022) studied indoor quality and associated behaviors in an office building and created a model using k-means clustering algorithm, which is a machine learning method. There are also studies that create a thermal comfort model using Bayesian decision theory (Mui et al., 2020; Wong et al., 2014; Langevin et al., 2013). These studies focus on the differences between real and estimated thermal assessments and look for a better thermal comfort prediction model which is a comprehensive study topic on its own. In this study however, user and environment interaction and automating the tasks for maintaining thermal comfort and air quality is focused on with a probabilistic approach, by simplifying the problem to a basic level.

This study focuses on maintaining thermal comfort and air quality of the occupants in a specific built environment by a Bayesian model that decides on the best action according to environmental measurements. The model necessitates a dataset that is generated by recording relevant parameters and user behaviors, allowing the machine learning algorithm to detect patterns in how users respond to certain environmental conditions, such as ambient temperature, CO_2 levels, and humidity. If the environmental conditions in question are taken as input and the behaviors caused by these conditions are taken as output, this problem can be considered as a classification problem with Bayesian decision theory, which is a probabilistic approach. A probabilistic model trained by this dataset (paired environment-behavior data) works by taking certain environmental measurements as inputs, calculating the probability of possible actions/outputs to occur, and choosing the most likely action to protect the thermal comfort and air quality of the occupants under changing environmental conditions.

In order to address the problem of maintaining occupant thermal comfort and air quality with a probabilistic model, a big dataset needs to be used with the required environmental measurements and corresponding behaviors are all considered. However, in this study the problem is simplified by narrowing down the input and output variables, in order to be addressed as a binary classification problem. Also, an open source dataset is used by narrowing down to the limits of the study. Therefore, the inputs are considered as four variables, namely indoor-outdoor temperature, CO_2 level, and air humidity, and the outputs are considered as only the window open/closed condition in order to simplify the problem as much as possible and treat the study as a binary classification problem. It is recognized that the dataset with many other measurements and associated behaviors needs to be considered for a comprehensive approach, and taking only these constrained environmental measurements and constrained behavioral data will not lead to a proper model for the real environment. Moreover, opening or closing windows are not even main case of action in maintaining thermal comfort in an office space. However, this study can set an example of how Bayesian decision theory, which is the basic statistical approach in machine learning, can be applied at a fundamental level within the discipline of architecture.

2 Methodology

The problem of deciding the best course of action according to the environmental changes can be approached as a probabilistic decision problem with Bayesian decision theory. A machine learning model based on Bayesian decision theory is trained with the data set created with these input-output pairs. This probabilistic model works by taking certain environmental measurements as input, calculating the probability of possible actions as output, and choosing the most likely action.

The dataset that is selected for this study is collected by Langevin et al. (2015). They have collected the data in the Friends Center office building in Philadelphia-USA. They recorded a range of internal-external thermal measurements and their associated behaviors over a one-year period (between the summer of 2012 and 2013) in a group of twenty-four office workers. By choosing two weeks for each season during the year, they collected data through online daily surveys and datalogger readings for a total of eight weeks and obtained a large data set that is shared as open source (Langevin, 2019; URL-1). Datalogger measurements include indoor and outdoor hygrothermal (movement of moisture and heat) conditions. Occupants' positions, available control actions (heather, fan, thermostat, windows, doors, blinds), attributes and attitudes are collected via surveys. With this data set, they created an agent-based model of the behavior of the building occupants. Their dataset provides many measurements of environmental changes and the behaviors they yield in the office environment. However, only four main input variables (four environmental measurements) and two output variables (two behaviors) are selected out of this comprehensive dataset for the scope of this study in order to address the simplified binary classification problem. Since the problem is only to decide on the class of a particular environmental measurement, Bayes decision rule can be applied by comparing the posteriori probabilities of two classes and selecting the class that yields to maximum posteriori probability.

For this binary classification problem, four features defined as inputs are represented by a feature vector as follows: χ = [x1, x2, x3, x4]. The two classes are represented as: Ω = {ω1, ω2}. While applying Bayesian decision theory, when a certain χ observation (input) is made, the probabilities of it being included in the ω1 class and the ω2 class (output) are calculated separately and compared. The class with the highest posterior probability is selected. This rule is represented as follows:

If P (ω1 | χ) > P (ω2 | χ), then class ω1 is selected for that χ observation. Otherwise, the ω2 class is selected.

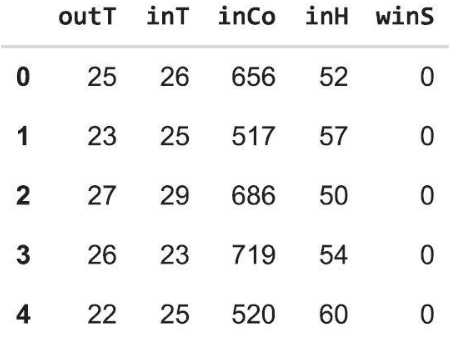
The posteriori probability of P (ω1 | χ) is the conditional probability of ω1 to χ, it expresses the probability of ω1 to occur with the condition that χ is observed.

After defining the problem, choosing the algorithm and data set to be used, the collected data is prepared before proceeding to training the algorithm. An example of this is cleaning and balancing data. Since the binary classification approach is used, the data set should contain close amounts of data from both classes. In addition, since the units of the input variables are different and their numerical values are in different ranges, all input variable data should be drawn to a standard range in order not to cause algorithmic bias. Another thing to do before proceeding to the training phase is to divide the data set into two separate parts, namely the training set and the test set, by preserving the input-output pairs. Although it is not a ratio that should be followed precisely when making this distinction, it may be correct to allocate 1/5 or 1/4 of the entire data set as the test set, and the remainder as the training set. The purpose here is to hide the test set from the algorithm during algorithm training and use it for testing after training. In this test phase, where the performance of the model is evaluated, metrics such as accuracy, precision, recall and F1-score are used.

3 Results

For this study, only indoor and outdoor thermal measurements, indoor humidity level and indoor CO_2 level measurements are used as features of the environment. Also, the study is limited to be addressed as a binary classification problem with only two classes which are window open/ closed classes. Therefore, initially, the required data-set is prepared. This preparation part is done on excel by filtering the necessary data and absent data in the set. The dataset is also narrowed down to maintain the class balance. It’s seen that out of 41804 instances, 41650 of them were mapped to “0” window state which means that window is closed. Only 154 of the rows were mapped to “1” window state. In order to balance the number of data in each class, 154 of the data points affiliated to “0” window state are chosen randomly out of 41650 of them. After equalizing the data in both classes, the prepared dataset is saved as a csv file and is taken to google colab for data analysis, visualization and training with Python programming language. The new small dataset is used for data analysis and visualization in order to gain more insight about the patterns with suitable open source libraries such as pandas, matplotlib, seaborn, numpy and scipy (URL-2). Initially the data is visualized partly with separate columns and rows (Table 1). On the table, the columns show the outdoor temperature, the indoor temperature, the indoor CO_2 level, the indoor humidity level and the window state which is either closed (0) or open (1) respectively.

Table 1: A sample of the dataset is visualized on the table from google colab.

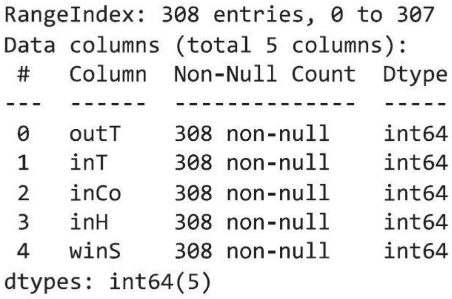


	outT	inT	inCo	inH	winS
0	25	26	656	52	0
1	23	25	517	57	0
2	27	29	686	50	0
3	26	23	719	54	0
4	22	25	520	60	0

Source: Authors, 2022.

Then the features of the dataset are checked and seen that there are 308 rows and five columns which have integer, non-null values in this dataset (Table 2). Then the last column in the dataset which shows the window state is turned into a categorical data. Therefore, the first four columns represent the inputs and the last one represents the output.

Table 2: Features of the dataset.

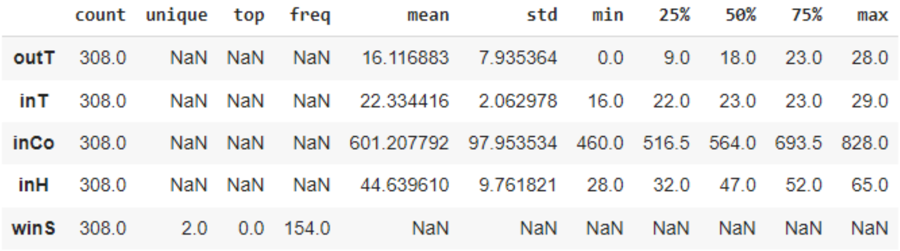


```
RangeIndex: 308 entries, 0 to 307
Data columns (total 5 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   outT    308 non-null    int64
 1   inT     308 non-null    int64
 2   inCo    308 non-null    int64
 3   inH     308 non-null    int64
 4   winS    308 non-null    int64
dtypes: int64(5)
```

Source: Authors, 2022.

Finally, some statistical analysis is done (Table 3). Other than the mean and standard deviation values of the numerical data of environmental measurements, the table also shows the number of categories and their frequencies for the categorical data representing the window state.

Table 3: Statistical information of the dataset.



	count	unique	top	freq	mean	std	min	25%	50%	75%	max
outT	308.0	NaN	NaN	NaN	16.116883	7.935364	0.0	9.0	18.0	23.0	28.0
inT	308.0	NaN	NaN	NaN	22.334416	2.062978	16.0	22.0	23.0	23.0	29.0
inCo	308.0	NaN	NaN	NaN	601.207792	97.953534	460.0	516.5	564.0	693.5	828.0
inH	308.0	NaN	NaN	NaN	44.639610	9.761821	28.0	32.0	47.0	52.0	65.0
winS	308.0	2.0	0.0	154.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Source: Authors, 2022.

After this, window state charts are drawn with respect to two different features (Figure 1, 2). Centers of the datapoints in each class are visualized as black dots on the charts.

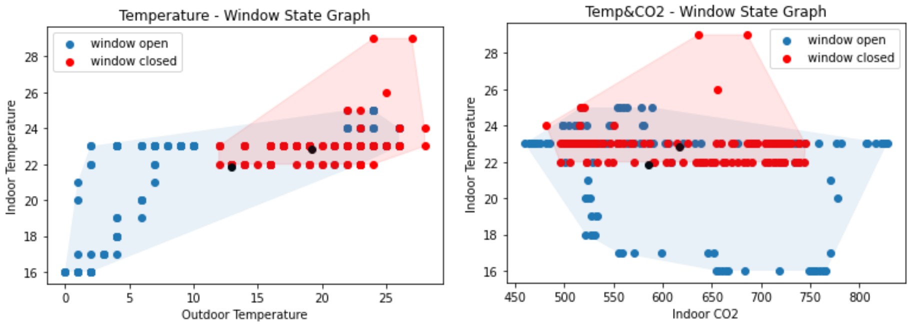


Figure 1: Window state chart with respect to outdoor and indoor temperature on the left and, with respect to indoor CO_2 level and temperature on the right (Authors, 2022).

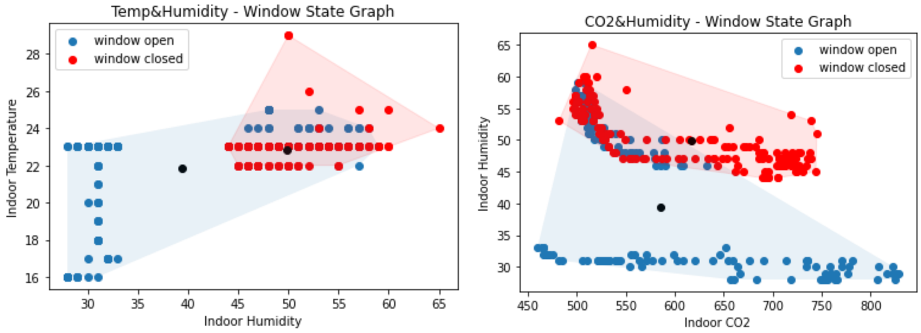


Figure 2: Window state chart with respect to indoor humidity level and temperature on the left and, with respect to indoor CO2 and humidity level on the right (Authors, 2022).

After the visualization process, a simple Bayesian algorithm which works well with small datasets is selected to train the model [URL-3]. The Gaussian Naïve Bayes algorithm is used with additional sklearn library with the assumption that the input variables have Gaussian distribution with a mean and standard deviation which is often suitable for continuous numerical data. The dataset is split into input and output values first by slicing the first 0 to 3 columns as inputs and the last column as output. Before going further, the data is split into as train and test set for both inputs and outputs. The ratio of test set size is assigned as 25 % with the code below.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test
=train_test_split(X,y,test_size= 0.25, random_state=0)
```

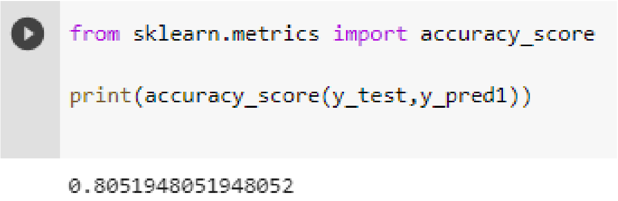
Then the dataset is standardized due to the different scales in the input variables [URL-4]:

```
from sklearn.preprocessing import StandardScaler
sc_X = StandardScaler()
X_train = sc_X.fit_transform(X_train)
X_test = sc_X.fit_transform(X_test)
```

Afterwards, Gaussian classifier is imported from sklearn libraries and used to train the model with training set, and test the model with the test set that is split in the beginning of this process, by predicting outcomes for the input test set:

```
from sklearn.naive_bayes import GaussianNB
classifer1 = GaussianNB()
classifer1.fit(X_train, y_train)
y_pred1 = classifer1.predict(X_test)
```

Then the accuracy of the predictions of the model is checked in order to evaluate how well it works. For this, the grand truth values of the test set are compared with the model’s predictions, and the accuracy rate is calculated. The model gets an accuracy of 80 % (Figure 3).



```
from sklearn.metrics import accuracy_score

print(accuracy_score(y_test,y_pred1))
```

```
0.8051948051948052
```

Figure 3: Accuracy score of the trained model (Authors, 2022).

The model is evaluated further with a confusion matrix -typically used for binary classification problems- which shows the accuracy of predictions in comparison to the grand truth values (Figure 4). For this the required modules are imported from sklearn and seaborn library which are used to plot a heatmap of the confusion matrix:

```
import seaborn as sns
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred1)
sns.heatmap(cm,annot=True)
```

On the confusion matrix horizontal “0” and “1” states depict the predicted values while the vertical ones depict the actual values. From the matrix, it can be seen that Class-0 is predicted well, while Class-1 is not predicted so well (Figure 4).

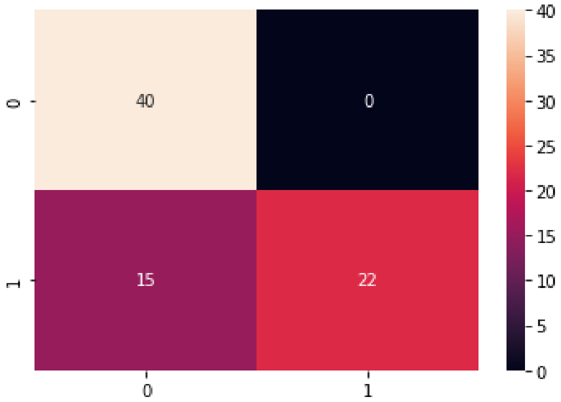
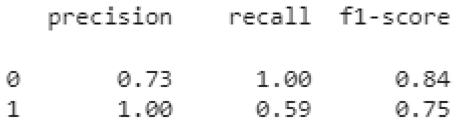


Figure 4: Confusion matrix of the model. (Authors, 2022).

Afterwards, the classification report which shows other metrics is also printed out:

```
from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred1))
```

Classification report shows the metrics that are obtained from confusion matrix (Figure 5). “Precision” shows the percentage of the predictions in each class that were correct. “Recall” shows the percentage of the data in classes that is predicted correctly. “F1-score” is a metric which considers recall and precision together.



```
       precision    recall  f1-score

0           0.73      1.00      0.84
1           1.00      0.59      0.75
```

Figure 5: Classification report of the model.

.

4 Discussion

Application of the Bayesian decision theory and binary classification in the context of occupant-building interaction within the architecture domain were the main topic of this paper. The dataset included two output variables and four input variables, totaling a modest 308 entries. Despite its small size, this dataset provided information and promising classification performance. While encouraging, the results obtained show the need for further research and development to realize the full potential of this strategy.

Evaluation of the model incorporates the evaluation of metrics focusing on the prediction performance of the model. The model obtained 0.80 “accuracy”. Considering the balanced dataset which has equal amounts of instances from both classes, a random guess would give 0.50 accuracy. Because predicting that all the instances are “Class 1” or “Class 0” would end up predicting the half of the classes true while half of them wrong. Therefore, taking this 0.50 as baseline, the accuracy of 0.80 of the trained model in this study is a reasonably good outcome which shows that the model indeed learned useful information from the data. Although an accuracy score of 0.80 is encouraging, it should be analyzed in light of the particular requirements of the task. Because 80% accuracy means also that 20% of the predictions were wrong which can mean insufficiency for different domains.

Considering other metrics, the “precision” results of the model show that 100% of the all Class-1 predictions are actually Class-1, 73% of the all Class-0 predictions are actually Class-0. It shows that the model performs perfectly well for Class-1 predictions, while having a reasonably well prediction for Class-0.

The “recall” score is 1.00 for the Class-0 and 0.59 for the Class-1. This means that, all of the actual Class-0 instances were predicted as Class-0 by the model, while 59% of actual Class-1 instances were predicted as Class-1. Therefore, the model works perfectly for recognizing Class-0 instances, while it recognizes Class-1 instances only slightly better than the rest. This difference on the score implies that while a significant number of actual window states are accurately captured by the model, some of them are also missed. This incoherence shows the trade-offs of the data simplification in this study and reminds that the model, while promising, is a just simplification of real-world complexities.

Additionally, Class-0 gets an “F1-score” of 0.84 while Class-1 gets a score of 0.75, which means a reasonable balance between recall and precision for both classes. About these metric scores, it can be said that the ones which are below 80% are open for improvement.

These findings highlight the potential usage of Bayesian decision theory, even with a small dataset and simplified inputs, in architecture domain. Nevertheless, it is accepted that this research is preliminary and, future work should go deeper into larger datasets and investigate other machine learning algorithms to handle problems to address real world scenarios. In example, going further with Bayesian decision theory can include the approaches such as minimum risk classifier (chooses the action that minimizes risk) with large datasets of multi-component features (input variables), multiple classes (output variables), and multiple actions.

5 Conclusion

In this study, a simplified occupancy comfort problem is addressed with a binary classification approach based on Bayesian decision theory. It is a study with a narrowed scope, aiming to utilize a basic statistical pattern recognition approach in an architectural context. With this narrowed scope, it is accepted that the model proposed in the study isn’t a proper model of the real world. However, the study can serve as an example of how Bayesian decision theory can be applied in the discipline of architecture and seen as a preliminary study for further problems that can be addressed with machine learning methods.

The model demonstrated the ability to distinguish between the two classes with a promising degree of accuracy, even with the limitations imposed by the small dataset. The other metric scores of the model were also reasonably good. However, the trade-off of using a simplified approach and a dataset was shown on some of the score results as discrepancies. The efficiency and resilience of the model could be improved yet more by enlarging the dataset, input and output variables and investigating other machine learning algorithms.

This study focused on bridging the realms of architecture and statistical modeling, providing an overview of the field of occupant-building interaction. Although the simplified model, based on Bayesian decision theory, provided promising results for the simplified problem in this study, it is recognized that the real world is much more complex and this model doesn't mirror it properly. This research is a preliminary one to build upon, emphasizing the potential of statistical approaches, for the future studies that encompasses required data and methods. Therefore, exploring and expanding these models with data, can lead to environments which are not only comfortable but also sustainable and responsive to the needs of their occupants.

Acknowledgements. This study is conducted as an individual project in Pattern Recognition & Analysis lecture which is the 2021-2022 Spring semester course of Istanbul Technical University, given by Prof. Dr. Behçet Uğur Töreyin. Merve Akdoğan conducts her studies under the grant of YÖK 100/2000 PhD scholarship.

References

He, M., Pen, H., Li, M., Huang, Y., Yan, D., Lou, S., & Wen, L. (2022). Investigation on typical occupant behavior in air-conditioned office buildings for South China's Pearl River Delta. *Architectural Intelligence*, 1(1), 1-19.

Langevin, J. (2019). Longitudinal dataset of human-building interactions in US offices. *Scientific data*, 6(1), 288.

Langevin, J., Gurian, P. L., & Wen, J. (2015). Tracking the human-building interaction: A longitudinal field study of occupant behavior in air-conditioned offices. *Journal of Environmental Psychology*, 42, 94-115. Author, A. A., & Author, B. B. (Date).

Langevin, J., Wen, J., & Gurian, P. L. (2015). Simulating the human-building interaction: Development and validation of an agent-based model of office occupant behaviors. *Building and Environment*, 88, 27-45.

Langevin, J., Wen, J., & Gurian, P. L. (2013). Modeling thermal comfort holistically: Bayesian estimation of thermal sensation, acceptability, and preference distributions for office building occupants. *Building and Environment*, 69, 206-226.

Mahdavi, A., Berger, C., Tahmasebi, F. & Schuss, M. (2019). Monitored data on occupants' presence and actions in an office building. *Sci Data*. https://doi.org/10.1038/s41597-019-0271-7.

Mui, K. W., Tsang, T. W., & Wong, L. T. (2020). Bayesian updates for indoor thermal comfort models. *Journal of Building Engineering*, 29, 101117.

Paige, F., Agee, P. & Jazizadeh, F. (2019). flEECe, an energy use and occupant behavior dataset for net-zero energy affordable senior residential buildings. *Sci Data*. https://doi.org/10.1038/s41597-019-0275-3.

Wong, L. T., Mui, K. W., & Cheung, C. T. (2014). Bayesian thermal comfort model. *Building and environment*, 82, 171-179.

URL-1. Kaggle. (2022, April). Human Building Office Space Interactions. https://www.kaggle.com/datasets/claytonmiller/humanbuilding-office-space-interactions

URL-2. Global AI Hub. (2022, May). Data Visualization. https://globalaihub.com/courses/data-visualization/

URL-3. Hand on Cloud. (2022, June). Implementing Naïve Bayes Classification Using Python. https://hands-on.cloud/implementing-naive-bayes-classification-using-python/

URL-4. DataCamp on Youtube. (2022, June). Spreadsheets Tutorial: Standardizing data. https://www.youtube.com/watch?v=Xg54m8f5sJI

URL-5. Towards Data Science. (2022, June) https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62