Aniruddha Joshi
12/12/2024
Applied Data Science
DSC680-T301 (2253-1)

Student Name: Aniruddha Joshi

Data Set: https://www.kaggle.com/datasets/claytonmiller/ashrae-global-thermal-comfort-database-ii/data

Milestone 2 – White Paper

Predicting Thermal Comfort

Business Problem

Buildings consume a significant amount of energy globally, with a large portion dedicated to Heating, Ventilation, and Air Conditioning (HVAC) systems. Optimizing HVAC usage presents a major opportunity for energy conservation, carbon reduction and cost reduction. However, this optimization should be achieved while maintaining occupant comfort level. Striking this balance is the core business problem being addressed by this project.

Background/History

Thermal comfort is a complex issue driven by various environmental factors like temperature, humidity, and air velocity. There are various standard written on thermal comfort such as ASHRAE. These standards inform building design and HVAC operation but often rely on static settings that may not adapt well to individual preferences and dynamic environmental conditions.

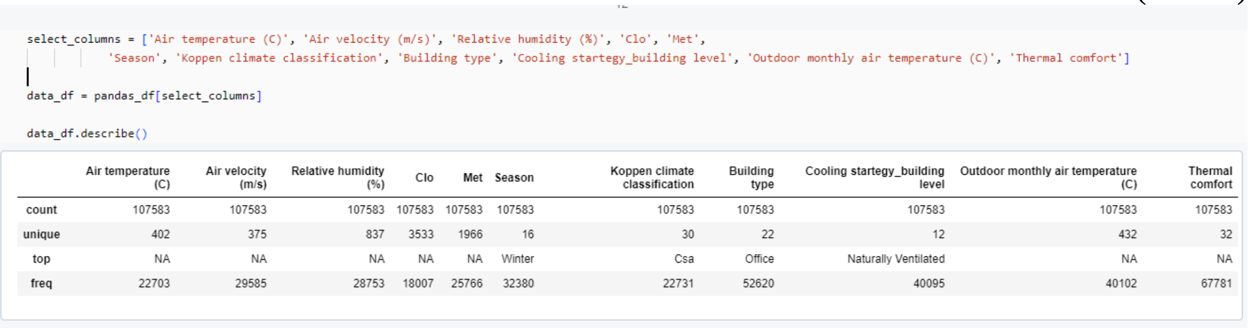
By leveraging large datasets of real-world comfort evaluations and corresponding environmental data, predictive models can be developed to inform more intelligent and responsive HVAC control strategies.

Data Explanation

This project utilizes the ASHRAE Global Thermal Comfort Database II, a publicly available dataset hosted on Kaggle.

Information about the data

1. The dataset contains over 80,000 records of indoor climate observations, each accompanied by individual comfort evaluations.
2. Python libraries like Pandas and NumPy are being used to clean the data. The process involved:
    a. Removing invalid or incomplete rows. E.g. NA records
    b. Converting categorical data into a suitable format for analysis.
    c. Keeping only relevant columns
3. Factors that influence the correlation
    a. **Environmental Factors:** Temperature, humidity, air velocity
    b. **Personal Factors:** Clothing level, metabolic rate (activity level)
    c. **Location Information:** building type
    d. **Individual Feedback:** Thermal sensation vote (e.g., too hot, too cold, comfortable)



```
select_columns = ['Air temperature (C)', 'Air velocity (m/s)', 'Relative humidity (%)', 'Clo', 'Met',
                 'Season', 'Koppen climate classification', 'Building type', 'Cooling startegy_building level', 'Outdoor monthly air temperature (C)', 'Thermal comfort']

data_df = pandas_df[select_columns]

data_df.describe()
```

	Air temperature (C)	Air velocity (m/s)	Relative humidity (%)	Clo	Met	Season	Koppen climate classification	Building type	Cooling startegy_building level	Outdoor monthly air temperature (C)	Thermal comfort
count	107583	107583	107583	107583	107583	107583	107583	107583	107583	107583	107583
unique	402	375	837	3533	1966	16	30	22	12	432	32
top	NA	NA	NA	NA	NA	Winter	Csa	Office	Naturally Ventilated	NA	NA
freq	22703	29585	28753	18007	25766	32380	22731	52620	40095	40102	67781

Figure 1 Selected columns

Feature Engineering

Given the high dimensionality of the data (70 variables), feature engineering will be employed to select and transform the most relevant variables.

Figure below shows the co-relational analysis between the key parameters. The trained variable is ‘Thermal Comfort’.

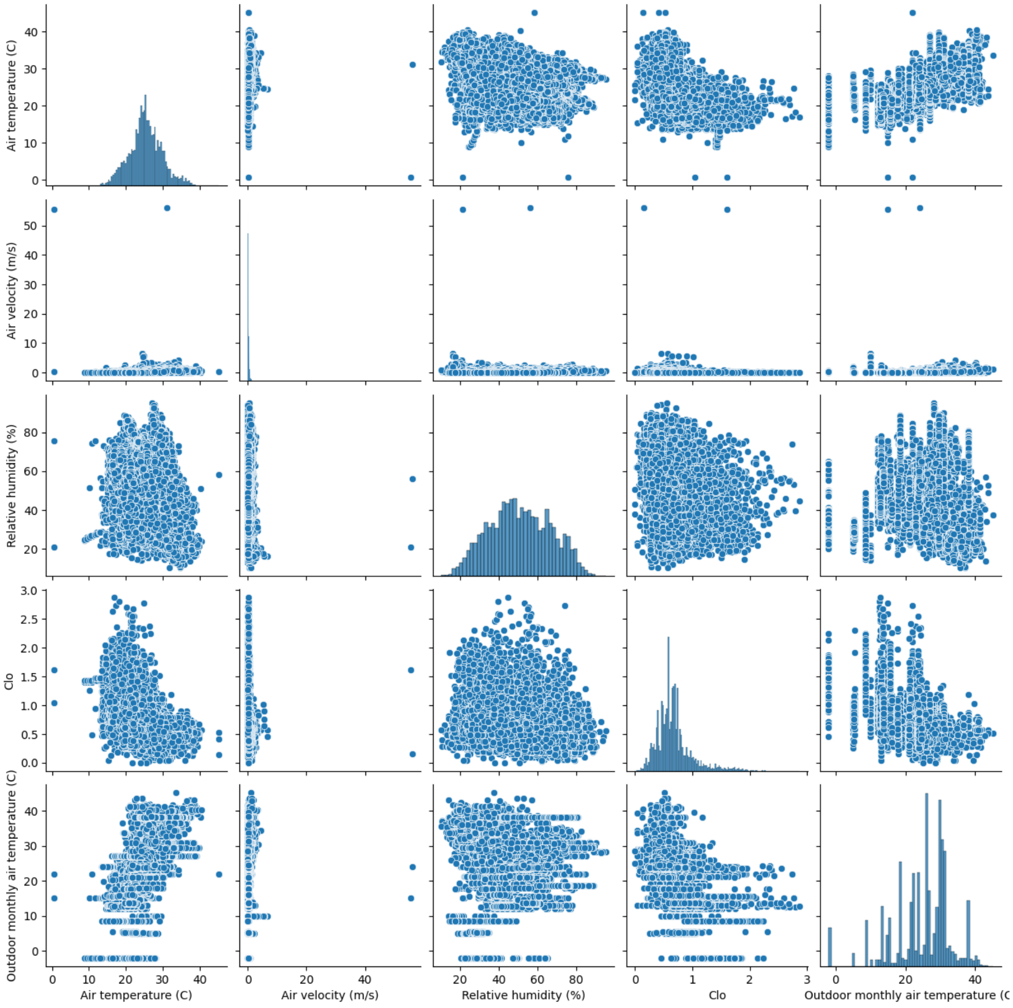


Figure 2: Co-relation Analysis

Methods

The project employs a machine learning and model evaluation methodology.

Data Exploration:

Data visualization using Matplotlib and Seaborn libraries are conducted to understand variable distributions, identify trends, and uncover potential relationships between variables.

Identifying the co-relation

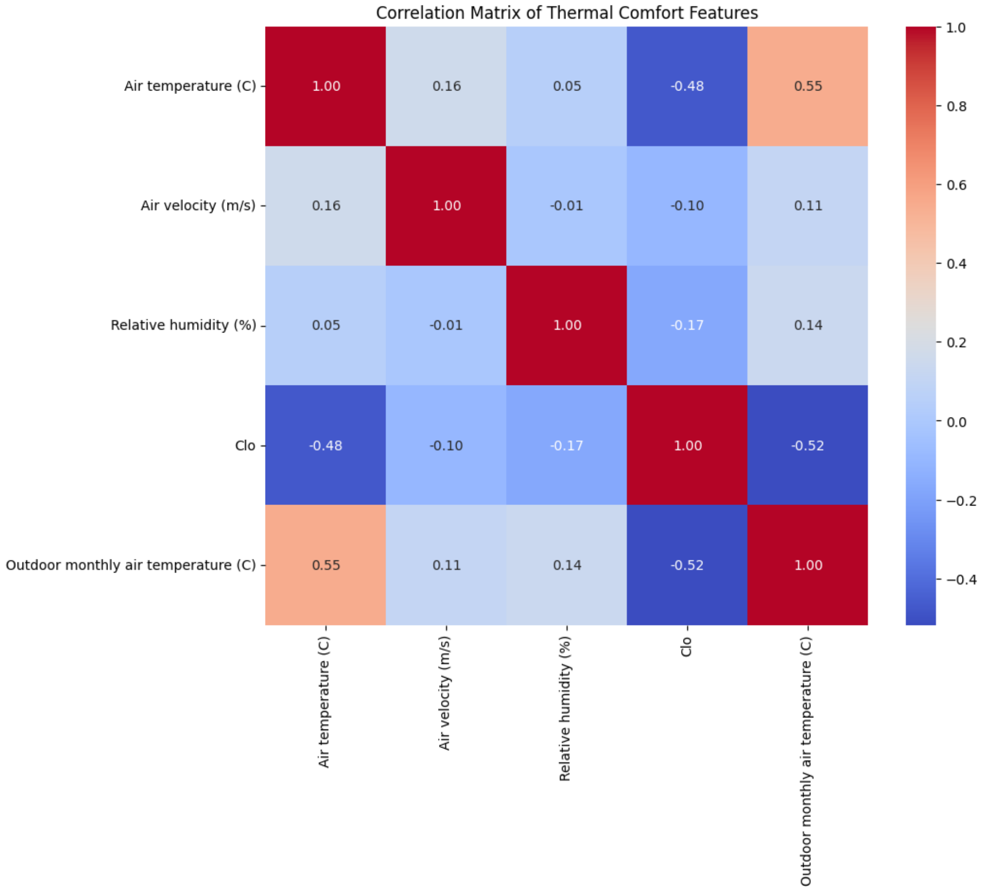


Figure 3: Co-relation Analysis

From the diagram above, we can see a negative correlation between Clo (clothing insulation) and outdoor temperature and air temperature.

To identify the density of the data distribution.

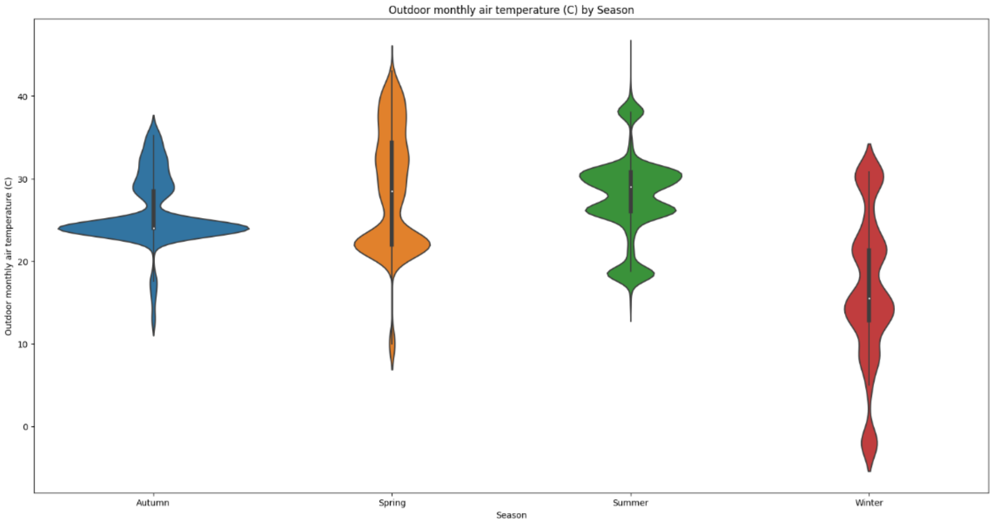


Figure 4 Data Density and Distribution

Machine Learning Modelling:
'Thermal comfort' is the categorical data; the RandomForestClassifier machine learning algorithm are used to train and predict thermal comfort.

Model Evaluation and Selection:
Trained models is evaluated on unseen data. The following metrics is used to assess predictive accuracy:

a. `Accuracy: 0.5397902529302899`
b. `Precision: 0.26713634558451604`
c. `Recall: 0.26910204787599834`
d. `F1-score: 0.2639454386959716`

Analysis

a. **Accuracy (0.54):** This parameter shows the overall correctness of the model. In this case, the model predicted the correct class about 54% of the time. This seems low.
b. **Precision (0.27):** This measures how often the model is correct when it predicts the positive class. A precision of 0.27 means that when the model predicts the positive class, it's only correct about 27% of the time. This suggests a high number of false positives (incorrectly predicting the positive class).
c. **Recall (0.27):** This measures how often the model correctly identifies the positive class out of all the actual positive instances. A recall of 0.27 indicates that the model only

captures about 27% of the actual positive cases. This suggests a high number of false negatives (failing to identify actual positive instances).

d. **F1-score (0.26):** This is the harmonic mean of precision and recall. It provides a balanced measure of the model's accuracy, taking both false positives and false negatives into account. A low F1-score like 0.26 suggests that the model is struggling to find a good balance between precision and recall.

Conclusion

This project aims to demonstrate the feasibility of using machine learning to predict individual thermal comfort based on readily measurable environmental data. The model helps develop intelligent HVAC control systems that optimize energy consumption while prioritizing occupant comfort.

Assumptions

It is assumed that the selected machine learning algorithms are capable of capturing the complex relationships between environmental factors and thermal comfort.
It is also assumed that the people provided true feelings about the temperature conditions during data collection.

Challenges

a. **Data Cleaning and Feature Engineering:** Handling missing data and reducing the dimensionality of the dataset effectively pose significant challenges.

b. **Model Selection:** Selecting the most appropriate machine learning model for this specific prediction task requires careful consideration and experimentation.

Future Uses/Additional Applications

a. **Personalized Comfort Control:** The developed model can be integrated into building management systems to provide personalized comfort control based on individual preferences and real-time environmental monitoring. For example, the people inside the building are analyzed based on who checks in.

b. **Predictive HVAC Control:** Predictions of future thermal comfort levels can inform anticipatory HVAC adjustments, further optimizing energy usage.

Recommendations

Explore advanced machine learning techniques: Investigate the potential of more sophisticated algorithms, such as deep learning models, to further enhance predictive accuracy.

Implementation Plan

First, acquire good quality data on Thermal Comfort, then handle missing values and prepare it for analysis. Then, exploratory data analysis to be conducted to uncover patterns and relationships. Experiment with various machine learning algorithms, tuning hyperparameters to optimize performance. Evaluate model performance using appropriate metrics to select the best model. Finally, develop a prototype for real-world deployment, potentially integrating it with existing building management systems and gather feedback from the stakeholders, and retune as needed.

Ethical Assessment

a. **Privacy:** Ensure that the dataset and any subsequent data collection processes adhere to privacy regulations and protect individuals' anonymity. Avoid collecting or storing any personally identifiable information (PII).

b. **Fairness:** Carefully evaluate the model for potential biases that could lead to unfair or discriminatory outcomes. Paying particular attention to sensitive variables like gender to ensure the model does not build any inequalities.

References:

- (Miller, n.d.) ASHRAE Standard 55: This standard provides guidelines for thermal comfort conditions in occupied spaces.
- (Venturewell, n.d.), this website provide information on Controls for Thermal Comfort

10 Questions an audience would ask :

1. Based on current score of accuracy, precision and F1-score, how confident are you in the model's ability to make accurate predictions?
2. What can you do to improve the model's accuracy?
3. What other algorithm can you test/experiment with?
4. Currently, removing NA significantly reduces the data size; what else can you do to keep the data size?
5. How do you plan to address the ethical considerations related to privacy and potential bias in the model's predictions, especially concerning gender?
6. What are the challenges in deploying these models?
7. How will the model's predictions be used in practice to optimise HVAC systems and improve energy efficiency?
8. What are the expected benefits of this project in terms of energy savings and improved occupant comfort?
9. How will the model be maintained and updated over time to ensure its continued effectiveness?
10. Is it possible to build individual comfort level HVAC systems?