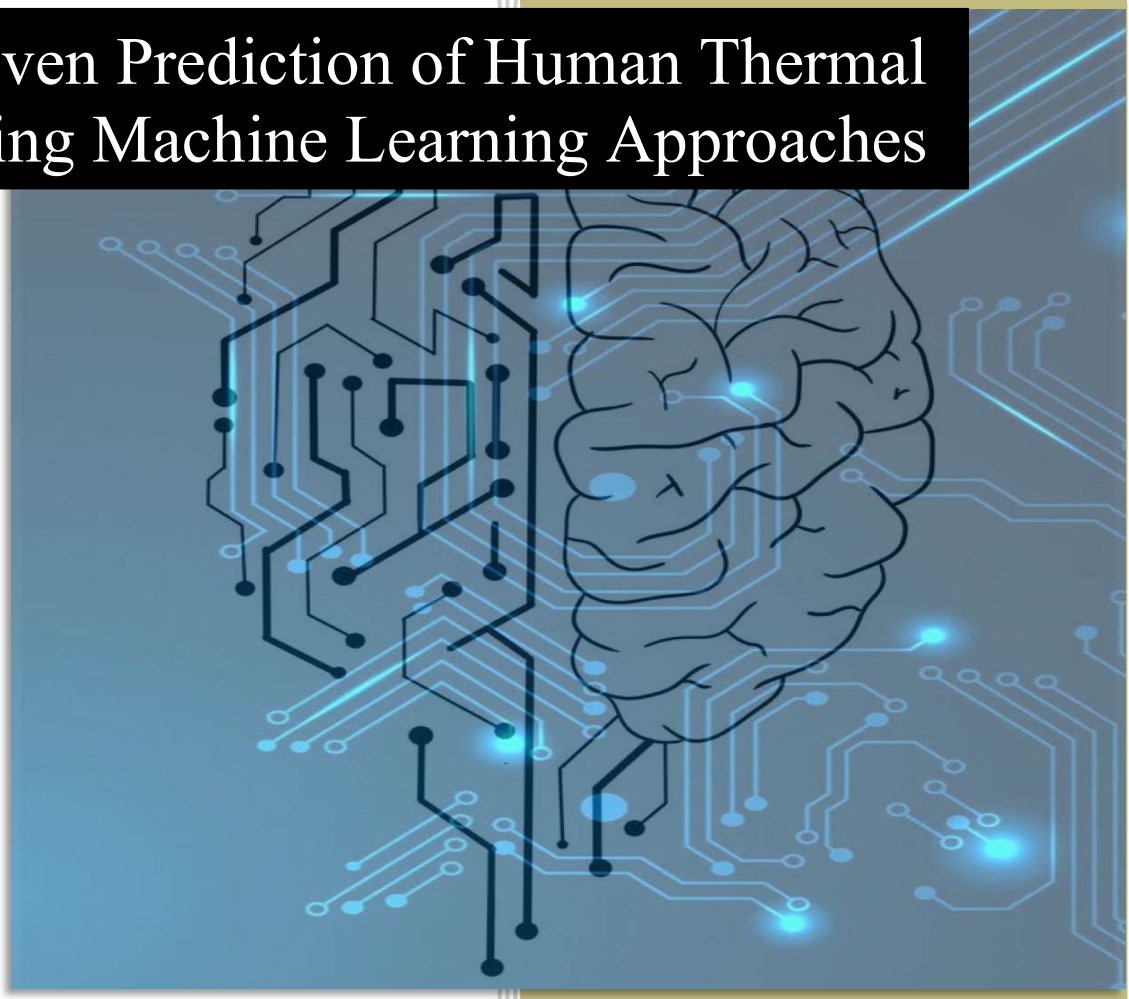
Data-Driven Prediction of Human Thermal Comfort Using Machine Learning Approaches



Iyeose Simon Uhumuavbi

Table of Contents

Abstract

Traditional physics-based models for assessing indoor thermal comfort, such as the Predicted Mean Vote (PMV), often fail to accurately predict individual thermal sensation in real-world buildings, leading to occupant dissatisfaction and inefficient energy consumption. Data-driven methods offer a powerful paradigm to move beyond these limitations by learning complex, non-linear relationships directly from large-scale field data.

This dissertation develops and critically evaluates a machine learning framework for improved prediction of human thermal sensation. The study utilizes the ASHRAE Global Thermal Comfort Database II, comprising over 100,000 entries, and implements a rigorous methodology. Data preprocessing included iterative imputation and class imbalance correction via the Synthetic Minority Over-sampling Technique (SMOTE). Five machine learning algorithms Logistic Regression (baseline), Support Vector Machine (SVM), Deep Neural Network (DNN), Random Forest, and Extreme Gradient Boosting (XGBoost) were systematically compared. The most promising models were subsequently fine-tuned through hyperparameter optimization.

The results demonstrate the marked superiority of non-linear ensemble models. Logistic Regression, serving as the baseline, achieved an accuracy of 23.0%. In contrast, the optimized XGBoost model attained an accuracy of 53.1%, representing a 130% relative improvement. Notably, the two leading models employed distinct predictive strategies: the Random Forest model prioritized direct physical and personal variables such as operative temperature, age, and clothing, whereas the XGBoost model relied primarily on broader geographical context, specifically the Köppen climate classification, as a proxy for thermal conditions.

This research demonstrates that machine learning models predict thermal sensation with substantially greater accuracy than traditional linear approaches. The primary determinants of comfort are identified as a combination of immediate physical state and broader geographical context. The developed framework offers a more precise and detailed alternative to conventional models, providing insights that can inform the design of energy-efficient and occupant-focused building control systems.

Keywords: Thermal comfort, Machine learning, Thermal sensation, PMV

Chapter 1: Introduction

1.1 Background and Context

Buildings are intricate ecosystems that mediate the relationship between their occupants and the external environment. At the heart of this relationship lies thermal comfort a cornerstone of Indoor Environmental Quality (IEQ) that influences human health, productivity, and satisfaction. ASHRAE Standard 55 defines thermal comfort as "that condition of mind which expresses satisfaction with the thermal environment" (ASHRAE, 2020, p. 4). Behind this succinct definition is a complex interplay of environmental conditions, individual physiology, and subjective expectations that shape modern indoor life.

Thermal comfort matters because people spend roughly 80–90% of their time indoors (Islam et al., 2023). Poor thermal conditions can impair cognitive performance, reduce productivity, and diminish quality of life (Lan, Wargocki, & Lian, 2011; Seppänen, Fisk, & Lei, 2006), while well-regulated environments promote wellbeing and performance. Comfort is also an energy challenge: heating, ventilation, and air conditioning (HVAC) systems account for around 40% of global building energy use (International Energy Agency, 2019), with heating and cooling alone responsible for over 20% of China's total consumption (Duanmu et al., 2023). Yet most systems still operate on fixed setpoints a "one-size-fits-all" strategy that often wastes energy while failing to meet the diverse, dynamic needs of occupants.

For decades, thermal comfort assessment has relied on two dominant paradigms: the heat-balance model (PMV–PPD) and the adaptive comfort model. The Predicted Mean Vote (PMV), introduced by Fanger (1970), estimates average thermal sensation from six variables air temperature, mean radiant temperature, relative humidity, air velocity, metabolic rate, and clothing insulation then converts this to a Predicted Percentage Dissatisfied (PPD). This approach underpins both ISO 7730 and ASHRAE 55. The adaptive comfort model, developed by de Dear and Brager (2002), instead links comfort temperature to outdoor weather, recognising that occupants adapt expectations and behaviours based on climate and season.

Both models, however, have well-documented limitations. PMV assumes steady-state conditions and fixed clothing and activity levels, which rarely match real-world variability, and it performs poorly in naturally ventilated or mixed-mode buildings where occupants can adjust their environment (Nicol, Humphreys, & Roaf, 2012). Adaptive models capture climate-related adaptation but generalise across populations and struggle with individual differences, particularly in tightly controlled HVAC spaces.

Recent advances in sensing, data storage, and computational power have opened the door to a new paradigm: data-driven thermal comfort modelling. Using large, high-quality datasets—such as the ASHRAE Global Thermal Comfort Database II (Földváry Ličina et al., 2018) and the Chinese Thermal Comfort Dataset (Duanmu et al., 2023) Machine Learning (ML) models can integrate environmental, personal, and physiological variables, learning complex, non-linear patterns beyond the reach of physics-based models. Studies increasingly show that ML can outperform PMV in predictive accuracy (Gao et al., 2025), enabling more personalised, adaptive, and energy-efficient control strategies.

1.2 Aims and objectives:

The central aim of this dissertation is to develop and critically evaluate a suite of machine learning models capable of accurately predicting indoor thermal sensation, leveraging a large-scale, publicly available database to integrate architectural insights with data science techniques. This project seeks to demonstrate that a data-driven approach can offer a more accurate and granular alternative to traditional comfort models, providing actionable insights for energy-efficient building design and operation.

To achieve this aim, the following objectives have been established:

1. To conduct a comprehensive review of existing thermal comfort models, data-driven methods in building science, and relevant architectural principles.
2. To acquire, preprocess, and engineer features from a large-scale, public thermal comfort database, (ASHRAE Global Thermal Comfort Database II), preparing it for machine learning applications. Each record includes measured variables (e.g. air temperature, humidity, air speed, radiant temperature) and personal data (e.g. clothing insulation, metabolic rate, age) along with subjective thermal sensation votes. Leveraging both environmental and personal inputs addresses the limitations of one-size-fits-all models.
3. To train, validate, and compare the performance of several classification-based machine learning models (including Logistic Regression, Support Vector Machine, Random Forest, XGBoost, and a Deep Neural Network) for the prediction of occupant thermal sensation.
4. To identify and analyze the most influential variables in predicting thermal comfort by conducting a feature importance analysis on the best-performing models, interpreting the findings in the context of building science and design principles.
5. To critically evaluate the performance and limitations of the developed models and discuss the practical implications for energy-efficient HVAC control strategies and future building design guidelines.

1.3 Research Questions

This study will be guided by the following research questions:

- **RQ1:**To what extent can modern classification-based machine learning models predict individual thermal sensation votes (TSV) more accurately than a baseline linear model?
- **RQ2:**Which machine learning algorithm provides the optimal balance of predictive accuracy and reliability for thermal comfort prediction after a systematic optimization process?

- **RQ3:**What are the most significant environmental and personal variables that influence thermal comfort sensation, and how does their importance ranking inform architectural and building system design?

1.4 Dissertation Structure

This dissertation is structured into six chapters. Chapter 2 presents a comprehensive literature review, examining foundational principles of thermal comfort, limitations of the PMV model, adaptive comfort theory, and recent advances in machine learning for thermal comfort prediction. Chapter 3 outlines the methodology, including data sources, preprocessing and feature engineering, model selection, and evaluation metrics. Chapter 4 reports the results, detailing model performance comparisons and feature importance analysis. Chapter 5 discusses the findings, interprets their significance, addresses the research questions, and considers practical implications and study limitations. Chapter 6 concludes by summarizing key findings, highlighting the study's contributions, and proposing directions for future research.

Expected Impact

By developing accurate and scalable models for thermal comfort, this research will contribute to energy-efficient building design and HVAC control. It will offer insights into how occupant characteristics and environmental factors interact to influence comfort, potentially informing updates to comfort standards. Moreover, it will demonstrate the advantages of data-driven comfort prediction in practical applications like real-time building management systems.

Chapter 2: Literature Review

2.1 Chapter Overview

The pursuit of thermal comfort within the built environment represents a complex intersection of physics, physiology, psychology, and engineering. This chapter provides a comprehensive review of the foundational principles and dominant models that have historically governed this field. It begins by establishing the fundamental heat balance principles that underpin human thermal perception. The chapter then critically examines the two canonical frameworks for comfort assessment: the static, physics-based Predicted Mean Vote (PMV) model and the dynamic, behaviour-centric Adaptive Comfort model.

Following this theoretical grounding, the review transitions to the contemporary paradigm shift towards data-driven methodologies. It surveys the recent proliferation of machine learning applications in thermal comfort research, highlighting the key algorithms, datasets, and findings that characterize the current state-of-the-art. This exploration will underscore the limitations of traditional models that your project aims to address and will firmly situate the present study within the evolving landscape of intelligent, occupant-centric building science. Ultimately, this chapter will construct the scholarly rationale for employing advanced machine learning techniques as a superior alternative for predicting individual thermal comfort.

2.2 Human Thermal Comfort

Human thermal comfort is a matter of thermal balance. The human body continuously generates heat through metabolic processes and must dissipate this heat to the surrounding environment to maintain a stable core temperature of approximately 37°C. Thermal sensation, and by extension thermal comfort, is the conscious perception of the outcome of this heat exchange process. This exchange occurs through four primary modes of heat transfer:

- **Conduction:** Direct heat transfer through physical contact with surfaces (e.g., a chair).
- **Convection:** Heat transfer to or from the body via moving fluids, primarily air.
- **Radiation:** Heat transfer via electromagnetic waves between the body and surrounding surfaces.
- **Evaporation:** Heat loss as moisture (perspiration) evaporates from the skin.

The net rate of heat loss from the body must equal the rate of metabolic heat production for thermal equilibrium to be maintained. Any imbalance results in the body initiating thermoregulatory responses such as shivering or sweating and is perceived as a sensation of being too cold or too warm (ASHRAE, 2020). This heat balance is influenced by six key factors, which form the basis of all major comfort models. Four are environmental: air temperature, mean radiant temperature, air velocity, and relative humidity, and two are personal: clothing insulation and metabolic rate.

.

2.3 The Predicted Mean Vote (PMV) Model: A Static Approach

2.3.1 Theoretical Foundation

For over five decades, the dominant framework for thermal comfort assessment has been the Predicted Mean Vote (PMV) model, developed by P.O. Fanger (1970). The PMV model is a comprehensive heat-balance equation that integrates the six key factors mentioned above into a single index. Its objective is to predict the average thermal sensation vote of a large group of people on the 7-point ASHRAE scale. The model was developed through extensive climate chamber experiments under steady-state conditions, where subjects were exposed to controlled environments and their thermal responses were recorded.

The output of the PMV model is an index that corresponds directly to the thermal sensation scale. A PMV of 0 represents thermal neutrality, while +2 corresponds to "warm" and -2 to "cool." The model is formalized in major international standards, including ASHRAE Standard 55 (2020) and ISO 7730 (2005), making it the cornerstone of conventional building science and HVAC system design.

2.3.2 Limitations and Criticisms

But the PMV model, for all its importance, runs into serious trouble, particularly when applied to real-world, dynamic building environments rather than controlled laboratories.

- **Dependence on a "Standard" Occupant:** The model's primary shortcoming is its reliance on fixed, idealized inputs that assume a generalized "standard" occupant. It struggles to account for the significant inter-individual differences in thermal preference, physiology, and metabolic rate that exist in any real population (Schweiker et al., 2020). This "one-size-fits-all" approach is a fundamental source of its predictive inaccuracies.
- **Inaccuracy in Real-World Conditions:** Numerous field studies have demonstrated a significant discrepancy between the PMV model's predictions and occupants' actual thermal sensation votes (TSV). Humphreys and Nicol (2002) found that the PMV model could only account for approximately 34% of the variance in comfort votes in field studies. This inaccuracy is particularly pronounced in naturally ventilated buildings where occupants have more control over their environment.
- **Neglect of Psychological and Behavioral Factors:** The PMV model is a purely physical heat-balance model and does not account for the powerful influence of psychological adaptation, thermal history, and occupant expectation. If an occupant has the ability to open a window or adjust a thermostat, their perception of comfort changes, a phenomenon that the static PMV model cannot capture

2.4 The Adaptive Comfort Model: A Dynamic Alternative

These glaring issues with the PMV model led researchers like de Dear and Brager (1998) to develop a new approach: the adaptive comfort model. The central hypothesis of the adaptive model is that occupants are not passive recipients of their thermal environment but are active agents who interact with and adapt to it.

2.4.1 Principles of Adaptation

The key idea is that people aren't just passive thermometers in a room; they are active participants who interact with and adapt to their environment. The adaptive model says that if something makes us uncomfortable, we'll do things to make ourselves comfortable again. This happens in three ways:

1. **Behavioral Adaptation:** We tend to adjust our clothing, or move to a different spot, or change the environment itself (by opening a window or using a fan).
2. **Physiological Adaptation:** Over the long term, our bodies acclimate to the local climate and seasons.
3. **Psychological Adaptation:** Altered expectations and thermal preferences based on context. For example, people tend to accept and even prefer a wider range of temperatures in naturally ventilated buildings compared to air-conditioned ones.

The adaptive model, formalized in ASHRAE Standard 55 for naturally ventilated buildings, defines acceptable comfort zones based primarily on outdoor temperature, implicitly capturing these adaptive mechanisms. While it offers a more realistic portrayal of comfort in certain building types, its application is limited, and it does not provide a granular, predictive vote like the PMV model.

2.4.2 The Empirical Formula in ASHRAE Standard 55

The principles of the adaptive model are formalized within ASHRAE Standard 55 (2020) for occupants in naturally ventilated buildings. Based on extensive analysis of field study data from around the world, the standard provides a simple linear regression formula that relates the optimal indoor "comfort temperature" (T_c) to the prevailing mean monthly outdoor air temperature (T_{out}).

For the 80% acceptability limits, the comfort temperature is defined as:

$$T_c = 0.31 \times T_{out} + 17.8$$

This formula is valid for outdoor temperatures ranging from 10°C to 33.5°C. The standard then defines comfort zones of ±2.5°C or ±3.5°C around this optimal temperature, depending on the desired level of acceptability.

The fact that this simple, data-driven model is now part of the main international standard is a huge deal. It’s the building science community officially admitting that a single, fixed temperature for a building doesn't work; comfort should instead adapt to the conditions outside.

But the model itself is just a first step. It oversimplifies reality by reducing all the complex personal and environmental factors down to a single straight line graph.

And that's the opening for the kind of work I'm doing in this dissertation. This is exactly where more advanced methods like machine learning can build on that simple foundation. By looking

at a much wider range of factors, the models I'm developing can create a far more detailed and accurate picture of what actually makes an individual feel comfortable.

2.5 The Rise of Data-Driven Methods: Machine Learning in Thermal Comfort

The limitations of both the PMV and adaptive models one being too rigid, the other too general have created a clear need for a new paradigm. In recent years, the confluence of widespread sensor deployment, large-scale data collection (as exemplified by the ASHRAE Global Thermal Comfort Database II), and advancements in machine learning has offered a powerful alternative (Amasyali & El-Gohary, 2018). Data-driven approaches move beyond static equations and instead learn complex, non-linear relationships directly from historical data.

2.5.1 Justification for a Machine Learning Approach

Machine learning models offer several key advantages over traditional comfort models:

- **Handling Non-Linearity:** Human comfort isn't a simple straight-line equation. The relationship between variables like air temperature, humidity, and how a person actually feels is complex and non-linear. Machine learning is used to find these kinds of hidden patterns, which physics-based models tend to oversimplify.
- **Individualization:** They can learn patterns by training on large datasets with diverse individual responses. Patterns that account for personal preferences and demographic differences, moving beyond the "standard" occupant paradigm.
- **Implicit Learning of Context:** Machine learning models can learn the influence of contextual factors (like climate, building type, or season) without them being explicitly programmed into a physical equation.

2.5.2 Survey of Machine Learning Algorithms in Recent Research

Tree-based ensemble methods and neural networks have emerged as the most promising from the literature that explored applications of various machine learning algorithms for predicting thermal comfort.

- **Ensemble Methods (Random Forest and Gradient Boosting):** These methods have consistently been shown to be among the most accurate. Luo et al. (2020), in a comprehensive comparison using the same ASHRAE database as this study, found that Random Forest and Gradient Boosting models significantly outperformed both traditional PMV models and other machine learning algorithms like SVMs. They attributed this success to the models' ability to handle high-dimensional data and complex interactions effectively. Similarly, a study by Wang et al. (2021) on personal comfort models confirmed that tree-based ensembles provide an optimal balance of high accuracy and robustness. The **Random Forest** (Breiman, 2001) operates by constructing a multitude of decision trees and outputting the mode of the classes, making it highly resistant to overfitting. **XGBoost** (Chen & Guestrin, 2016), an advanced implementation of gradient

boosting, builds trees sequentially, with each new tree correcting the errors of the previous one, often leading to state-of-the-art performance.

- **Artificial Neural Networks (ANNs) and Deep Neural Networks (DNNs)**: Neural networks aren't new to this field, but the recent explosion in deep learning has certainly brought them back into the conversation. For example, Kim, Zhou, & Schiavon (2018) showed that personal comfort models built with ANNs could become highly accurate by learning directly from occupant feedback. More recently, researchers like Sadeghi et al. (2022) have been pushing the boundaries with deep learning, showing it can learn complex patterns from raw sensor data. But there's a major trade-off. For all their power, DNNs are often "black boxes," making it incredibly difficult to understand why they make a certain prediction. They also require a ton of computational power.
- **Support Vector Machines (SVMs) and Logistic Regression**: It's not just about ensembles and neural networks; other algorithms have been tried as well. Chaudhuri et al. (2018) had good results using SVMs for personalized models, mostly because SVMs are great at finding complex, non-linear dividing lines between different comfort states. The problem, as Luo et al. (2020) later noted, is that tree-based ensembles usually outperform them, and SVMs can get incredibly slow with large datasets, a critical issue for my study.

  Logistic Regression is rarely used for predicting the full 7-point scale but serves as an essential linear baseline to demonstrate the performance gain achieved by more complex, non-linear models.

2.5.3 The Importance of Interpretability

A critical challenge in the application of machine learning, particularly in building science, is the issue of interpretability. While a model may achieve high accuracy, its practical utility is limited if its decision-making process is opaque. For architects and engineers to trust and adopt a data-driven model, they must understand why it is making a certain prediction. This has led to a focus not just on predictive accuracy but also on model interpretability (D'Amour et al., 2020). Tree-based models like Random Forest and XGBoost offer a significant advantage in this regard, as they provide built-in methods for calculating feature importance. This allows researchers to rank the input variables by their predictive power, providing actionable insights into the key drivers of thermal comfort that can inform building design and control strategies.

2.6 Chapter Summary and Research Gap

This review has traced the evolution of thermal comfort modeling from the static, physics-based PMV model to the dynamic adaptive framework, and finally to the current paradigm of data-driven machine learning. While traditional models provide an essential theoretical foundation, their predictive accuracy in real-world settings is limited. The literature clearly shows that machine learning, particularly tree-based ensemble methods and neural networks, offers a promising path toward more accurate and personalized thermal comfort prediction.

However, a gap remains in the systematic comparison and deep interpretation of a diverse set of modern algorithms on a large-scale, global dataset. Many studies focus on a limited set of

models or are constrained to specific buildings or climates. This dissertation addresses this gap by undertaking a comprehensive, end-to-end project that involves:

1. Rigorous preprocessing of the largest available public comfort database.
2. Systematic benchmarking of five distinct and representative machine learning algorithms.
3. A thorough hyperparameter tuning phase to ensure each model achieves its maximum potential performance.
4. A final, comparative feature importance analysis to interpret and contrast the predictive strategies of the best-performing models.

By following this comprehensive approach, this study aims to provide a definitive answer to its research questions and contribute robust, actionable insights to the field of occupant-centric building science.

Chapter 3: Methodology

3.1 Chapter Overview

This chapter provides a detailed, systematic account of the methodology employed to develop and evaluate machine learning models for the prediction of human thermal comfort. The primary objective of this research is to leverage the comprehensive ASHRAE Global Thermal Comfort Database II to build robust predictive models and, subsequently, to identify the most significant factors influencing thermal sensation in real-world building environments.

The methodology follows a structured, multi-stage process, commencing with data acquisition and extensive preprocessing to ensure data quality and integrity. This is followed by a thorough Exploratory Data Analysis (EDA) to uncover underlying patterns, relationships, and anomalies within the data. Subsequently, a robust modeling workflow was established, involving the training and evaluation of five distinct machine learning algorithms: Logistic Regression, Support Vector Machine (SVM), Random Forest, Extreme Gradient Boosting (XGBoost), and a Deep Neural Network (DNN). The chapter further details the procedures for model optimization through hyperparameter tuning and concludes with the methods for model interpretation via feature importance analysis. Each step was designed to be rigorous, transparent, and reproducible, forming a sound scientific basis for the results and conclusions presented in the subsequent chapters.

3.2 The Dataset:

The foundation of this research is the ASHRAE Global Thermal Comfort Database II (Földváry et al., 2018), a comprehensive, publicly available dataset sourced from. The specific version used in this study was sourced from the Kaggle data science platform, available at:https://www.kaggle.com/datasets/ashRAE/global-thermal-comfort-database-ii. The original database is maintained by ASHRAE and the University of California, Berkeley's Center for the Built Environment. This database is a global compilation of data from thermal comfort field studies conducted across numerous countries, climates, and building types. It amalgamates objective physical measurements of the indoor environment with subjective "right-here-right-now" evaluations of thermal comfort provided by building occupants (Miller, C., [et al.]).

The richness of the dataset lies in its breadth and depth, containing over 100,000 individual entries. A detailed description of the key variables utilized or considered in this study is presented inTable 3.1.

Table 3.1: Description of Primary Variables in the ASHRAE Global Thermal Comfort Database II

Variable Name	Description	Data Type	Units / Example Values
Thermal sensation	**(Target Variable)** The subjective vote cast by an occupant on the 7-point ASHRAE thermal sensation scale.	Numerical (Categorical)	-3 (Cold) to +3 (Hot)
operative_temp	The uniform temperature of an imaginary black enclosure in which an occupant would exchange the same amount of heat by radiation and convection as in the actual non-uniform environment.	Numerical (Continuous)	°C
clo	The estimated thermal insulation provided by the occupant's clothing ensemble. A value of 1.0 clo is equivalent to a typical two-piece business suit.	Numerical (Continuous)	clo
met	The estimated metabolic rate of the occupant, representing their rate of internal heat production. A value of 1.0 met corresponds to a seated, quiet person.	Numerical (Continuous)	met
age	The occupant's self-reported age.	Numerical (Continuous)	Years
humidity	The relative humidity of the indoor air.	Numerical (Continuous)	%
air_velocity	The average speed of air movement in the immediate vicinity of the occupant.	Numerical (Continuous)	m/s
outdoor_air_temp	The prevailing monthly average outdoor air temperature for the study location.	Numerical (Continuous)	°C
Koppen climate classification	The Köppen-Geiger climate classification code for the study	Categorical	e.g., 'Cfa', 'BWh', 'Dfb'

	location, representing a standardized climate zone.		
Season	The season in which the data was collected at the study location.	Categorical	'Summer', 'Winter', etc.
Cooling strategy_building level	The primary method used to cool the building.	Categorical	'Air Conditioned', 'Naturally Ventilated', etc.
Building type	The primary function or use of the building where the study was conducted.	Categorical	'Office', 'Classroom', 'Residential', etc.
Sex	The occupant's self-reported sex.	Categorical	'Male', 'Female'
air_temp	The dry-bulb temperature of the air surrounding the occupant. (Dropped during feature selection due to high correlation with operative_temp).	Numerical (Continuous)	°C
radiant_temp	The mean temperature of the surrounding surfaces, weighted by their view factor from the occupant. (Dropped due to high correlation with operative_temp).	Numerical (Continuous)	°C
country	The country where the field study was conducted. (Dropped due to high cardinality and redundancy with Koppen climate classification).	Categorical	e.g., 'USA', 'UK', 'Australia'
year	The year in which the data for a specific entry was collected. (Dropped during feature selection as it was deemed unlikely to have predictive power).	Numerical (Categorical)	e.g., 1995, 2010

Key variables can be categorized as:

- **Environmental Variables**: Objective measurements such as Air Temperature (°C), Radiant Temperature (°C), Operative Temperature (°C), Relative Humidity (%), and Air Velocity (m/s).

- **Personal Variables**: Occupant-specific factors including Clothing Insulation (clo) and Metabolic Rate (met).
- **Contextual Variables**: Information regarding the building, geography, and temporality (Season, Year).
- **Target Variable**: The primary outcome variable (independent Variable) for this study is Thermal sensation, a subjective vote cast by occupants on the 7-point ASHRAE scale, ranging from -3 (Cold) to +3 (Hot).

The sheer variety in this real-world data is exactly what makes it so valuable. It lets us build models that can work across many different conditions beyond limited lab-only studies.

3.3 Data Preparation and Preprocessing

A rigorous data preparation phase was undertaken to transform the raw dataset into a clean, complete, and suitable format for machine learning. This phase was critical due to the known issues of missing data and potential inconsistencies inherent in large, aggregated databases.

3.3.1 Initial Data Loading and Feature Pruning

The dataset was initially loaded into a pandas DataFrame. An immediate review of the feature set was conducted to remove columns that were either irrelevant to the predictive task or contained excessive missing values that would make imputation unreliable. Columns related to alternative comfort models (e.g., PMV, PPD), which would introduce data leakage, and those with over 60% missing data were proactively pruned from the dataset to streamline the subsequent analysis.

3.3.2 Missing Value Imputation

A sophisticated, multi-step imputation strategy was designed to handle missing values while preserving the underlying data structure as accurately as possible.

- **Target Variable:** The most critical step was to remove any rows where the target variable, Thermal sensation, was missing. These records are unusable for supervised learning and were dropped to form the definitive set of observations for the study.
- **Hierarchical Temperature Imputation:** A domain-knowledge-based approach was used for the highly correlated temperature variables. The relationship between Mean Radiant Temperature (T_r), Globe Temperature (T_g), and Air Temperature (T_a) is well-defined by heat transfer principles and standardized in international norms. The fundamental equation governing the heat balance of a globe thermometer is described in the international standard ISO 7726 (International Organisation for Standardization, 1998). This relationship can be expressed as:

  $$\mathbf{T}_r^4 = \mathbf{T}_g^4 + (\mathbf{h}_c / (\sigma\varepsilon_g))\,(\mathbf{T}_g - \mathbf{T}_a)$$

  Where:

  - $\mathbf{T}_r$ is the Mean Radiant Temperature (in Kelvin)

- $\mathbf{T}_g$ is the Globe Temperature (in Kelvin)
- $\mathbf{T}_a$ is the Air Temperature (in Kelvin)
- $\mathbf{h}_c$ is the convective heat transfer coefficient, which is a function of air velocity (v)
- σ is the Stefan-Boltzmann constant
- ε_g is the emissivity of the globe surface

This physical relationship demonstrates that Globe temperature and Air temperature together contain direct, predictive information about Radiant temperature. Given this, a hierarchical imputation strategy was adopted. Where the Radiant temperature was missing, it was first imputed using the Globe temperature as the best available physical proxy.

For any remaining missing values where the Globe temperature was also unavailable, Air temperature was used as the next-best approximation. This pragmatic, physics-informed approach was chosen over a direct calculation for all rows, which was not feasible due to missing values in the necessary input variables (e.g., Globe temperature, Air velocity) for a significant portion of the dataset.

- **Categorical Features:** For categorical variables such as Building type and Cooling strategy_building level, missing values were imputed using the mode of each respective column.
- **Numerical Features:** For the remaining numerical features (e.g., clo, met, humidity), a more advanced technique, Multivariate Imputation by Chained Equations (MICE), was employed via Scikit-learn's Iterative Imputer. This method was chosen over simple mean or median imputation because it models each feature with missing values as a function of other features in the dataset. This approach better preserves the inter-variable relationships and typically results in more accurate and realistic imputed values.

3.3.3 Data Integrity Validation and Correction

Following imputation, a descriptive statistical analysis revealed the presence of physically impossible or highly implausible values. For instance, negative values were observed for ‘age’and‘clo’, and humidity values exceeded the physical limit of 100%. A data-clipping step was implemented to rectify this. Logical minimum and maximum bounds were defined for each affected feature (e.g., age, humidity), and any values falling outside these ranges were capped or floored to the respective boundary value. This ensured that the models were trained on a dataset that adheres to real-world physical constraints

.

3.3.4 Target Variable Preparation

The Thermal sensation variable, though on a discrete scale, was stored as a floating-point

number. To treat it as a true categorical classification target, the values were rounded to the nearest integer, and the data type was converted to integer.

3.4 Exploratory Data Analysis (EDA)

EDA was performed on the cleaned and complete dataset to understand the characteristics and inform other modeling decisions.

- **Distribution of the Target Variable:** A count plot of the Thermal sensation votes revealed a severe class imbalance. The "Neutral" (0) class was heavily predominant, with the extreme classes ("Cold" -3 and "Hot" +3) being significantly under-represented. This finding was critical, necessitating the use of specialized techniques to prevent model bias towards the majority class.
- **Numerical Feature Analysis:** Histograms were generated for all numerical predictors to visualize their distributions. A correlation matrix heatmap was then produced to assess linear relationships between these variables. This analysis revealed the extreme multicollinearity between three variables: 'air_temp', 'radiant_temp', and 'operative_temp'.
- **Categorical Feature Analysis:** Box plots were used to visualize the relationship between each key categorical predictor and the Thermal sensation target. This visually confirmed that features like Season and Cooling strategy had a strong influence on reported thermal comfort.

3.5 Predictive Modeling Workflow

The insights from the EDA directly informed the construction of the final modeling pipeline.

3.5.1 Final Feature Selection

Based on my EDA findings, a final feature selection step was executed. To address multicollinearity, 'air_temp' and 'radiant_temp' were dropped, while 'operative_temp' was retained as the most comprehensive single temperature metric. The 'country' column was dropped due to its high cardinality and redundancy with the 'Koppen climate' classification feature. The ' year' column was removed, as the EDA demonstrated it had a negligible correlation with the target variable. Finally, the Climate column, which provided a general climate description, was dropped to avoid redundancy. The 'Koppen climate' classification feature was retained as it offers a more detailed, granular, and internationally recognized system for climatic categorization.

3.5.2 Data Splitting and Preprocessing Pipeline

The dataset was split into a training set (80%) and a testing set (20%) using Scikit-learn's train_test_split. Crucially, stratification was employed to ensure the class distribution of Thermal sensation was identical in both the training and testing sets, preserving the real-world imbalance for the final evaluation.

A Column Transformer preprocessing pipeline was constructed to apply different transformations to different data types simultaneously.

- **Numerical Features:** Standard Scaler was applied to scale all numerical features to have a mean of zero and a standard deviation of one. This is essential for the proper functioning of algorithms like SVM and DNNs.
- **Categorical Features:** One-Hot-Encoder was applied to convert categorical features into a numerical format, creating a new binary column for each category.

3.5.3 Addressing Class Imbalance: SMOTE

To mitigate the severe class imbalance discovered during EDA, the Synthetic Minority Over-sampling Technique (SMOTE)was applied. It is critical to note that SMOTE was applied only to the training data after the train-test split. This prevents data leakage and ensures that the model is trained on a balanced dataset but tested on a dataset that reflects the true, imbalanced distribution of the real world.

3.6 Machine Learning Algorithms

Five distinct algorithms were selected to provide a comprehensive view of different modeling approaches.

1. **Logistic Regression:** Chosen as a simple, interpretable linear model to serve as a performance baseline.
2. **Support Vector Machine (SVM):** A powerful kernel-based method capable of finding complex, non-linear decision boundaries.
3. **Random Forest:** An ensemble method based on decision trees, known for its high accuracy, robustness, and ability to handle complex interactions.
4. **Extreme Gradient Boosting (XGBoost):** A state-of-the-art gradient boosting framework, often a top performer in machine learning competitions due to its performance and efficiency.
5. **Deep Neural Network (DNN):** A multi-layered neural network built with Keras, chosen to explore if a highly complex, non-linear architecture could capture deeper patterns in the data.

3.7 Model Optimization and Evaluation

3.7.1 Hyperparameter Tuning

To maximize performance, hyperparameter tuning was conducted for the top-performing models, specifically Random Forest, DNN, and XGBoost.

- To tune the tree-based models (Random Forest and XGBoost), Scikit-learn's Randomized SearchCV was used. This was chosen over a full grid search because it's much faster for exploring a wide range of settings (Bergstra & Bengio, 2012). For the Random Forest, the search tested 20 different combinations of key parameters like the number of trees, max depth, and sample splits using 3-fold cross-validation. The same tuning process was applied to the XGBoost model.

- For the neural network, the Keras Tuner library was used to find the best architecture. And set up a random search to optimize the most important hyperparameters: the number of hidden layers (from 1 to 3), the number of neurons in each layer (32 to 256), the dropout rate (0.2 to 0.5), and the learning rate. The tuner's goal was to find the combination that produced the best validation accuracy across 10 trials on the full dataset.

3.7.2 Performance Evaluation Metrics

A suite of metrics was used to provide a holistic evaluation of model performance:

- **Accuracy:** Calculated as the overall percentage of correct predictions across all classes. While a useful top-line metric, its limitations in the context of an imbalanced dataset were recognized, necessitating the use of more granular metrics.
- **Precision, Recall, and F1-Score:** These metrics were calculated on a per-class basis to provide a detailed understanding of the model's performance on each specific thermal sensation category.
    - **Precision** measures the accuracy of positive predictions (e.g., of all the times the model predicted "+3", how often was it correct?).
    - **Recall** (or Sensitivity) measures the model's ability to find all relevant instances of a class (e.g., of all the actual "+3" instances in the data, how many did the model find?).
    - The **F1-Score**, as the harmonic mean of precision and recall, was used as a key summary metric for evaluating the balanced performance on each class.
- **Macro vs. Weighted Average F1-Score:** To summarize the per-class metrics, both macro and weighted averages were computed.
    - The **Macro Average** calculates the F1-score for each class and then finds the unweighted average. This treats all classes equally, regardless of their support, making it a crucial metric for assessing performance on the under-represented minority classes.
    - The **Weighted Average** calculates the F1-score for each class and then finds the average, weighted by the number of true instances for each class (the support).
- **Confusion Matrix:** A detailed N x N matrix (where N=7 for the seven classes) was generated to visualize the specific errors made by each model. This allowed for a qualitative analysis of which classes were most often confused with one another (e.g., whether the model was making "adjacent" errors or more significant, illogical ones).

3.8 Feature Importance Analysis

The final step of the methodology was to interpret the "black box" of the champion models. For the tree-based models (Random Forest and XGBoost), the built-in Gini importance (.feature_importances_) was extracted to rank all features by their contribution to the model's

predictive power. This allowed for a direct comparison of the predictive strategies learned by the two best-performing algorithms.

3.9 Chapter Summary

In summary, this chapter has detailed a comprehensive and rigorous methodology designed to build, evaluate, and interpret machine learning models for thermal comfort prediction. From meticulous data preprocessing and insightful exploratory analysis to a systematic benchmarking and optimization of five distinct algorithms, every step was conducted with transparency and scientific validity in mind. This structured approach ensures that the findings presented in the subsequent chapters are robust, reliable, and contribute meaningfully to the understanding of human thermal comfort in the built environment.

Chapter 4: Results and Analysis

4.1 Chapter Overview

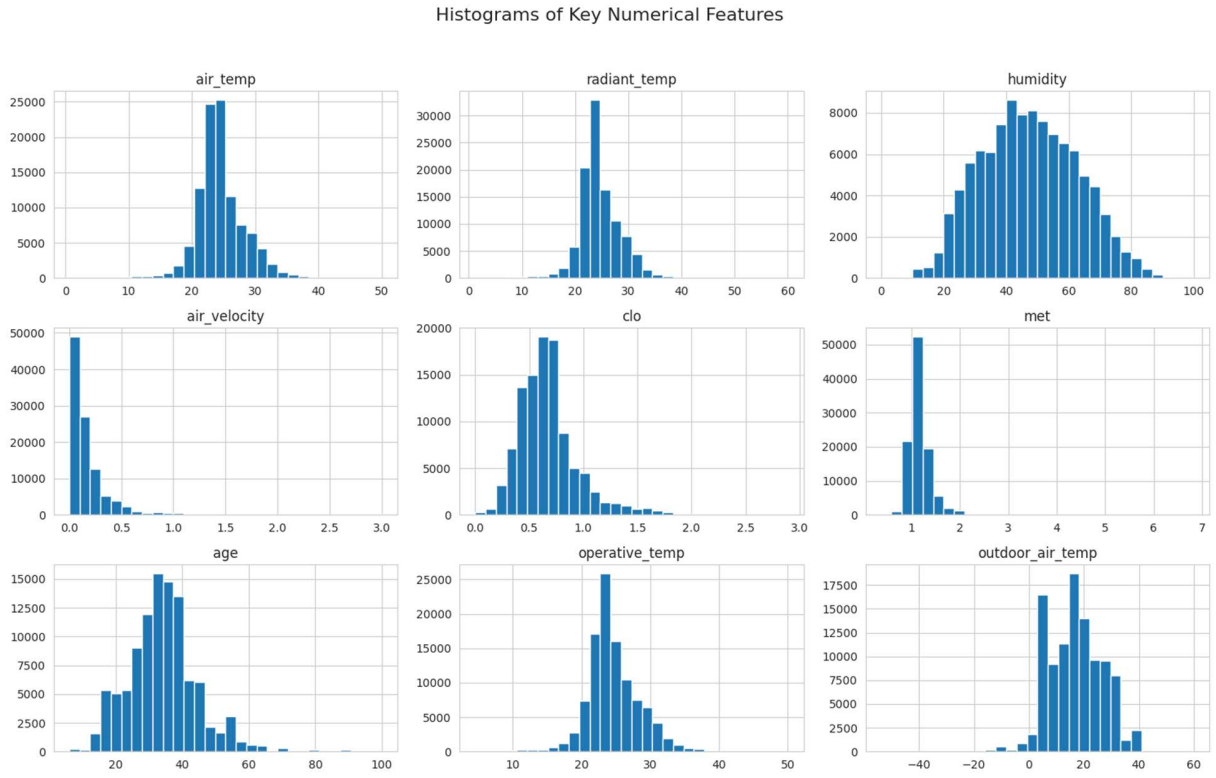
This chapter presents the empirical results obtained from the application of the methodology detailed in Chapter 3. The objective is to systematically report the findings at each stage of the data analysis and modeling process. The chapter begins by presenting the key outcomes of the Exploratory Data Analysis (EDA), which provided the foundational insights for the modeling strategy. It then proceeds to a comparative analysis of the initial performance of the five selected machine learning algorithms, establishing a crucial performance baseline.

Subsequently, the results of the hyperparameter tuning phase are detailed, highlighting the optimization of the top-performing models. This culminates in the selection of a final, champion model. The chapter concludes with the most significant analytical outcome of this study: a feature importance analysis conducted on the best-performing models to identify the primary drivers of thermal comfort prediction. All results are presented objectively through tables, figures, and statistical summaries, forming the factual basis for the interpretations and conclusions drawn in the subsequent Discussion chapter.

4.2 Exploratory Data Analysis Findings

The initial exploration of the preprocessed dataset yielded several critical insights that fundamentally shaped the modeling approach. Histograms were generated to visualize the distributions of key numerical variables, as presented in Figure 4.1.

Figure 4.1: Histograms of Key Numerical Features after Preprocessing



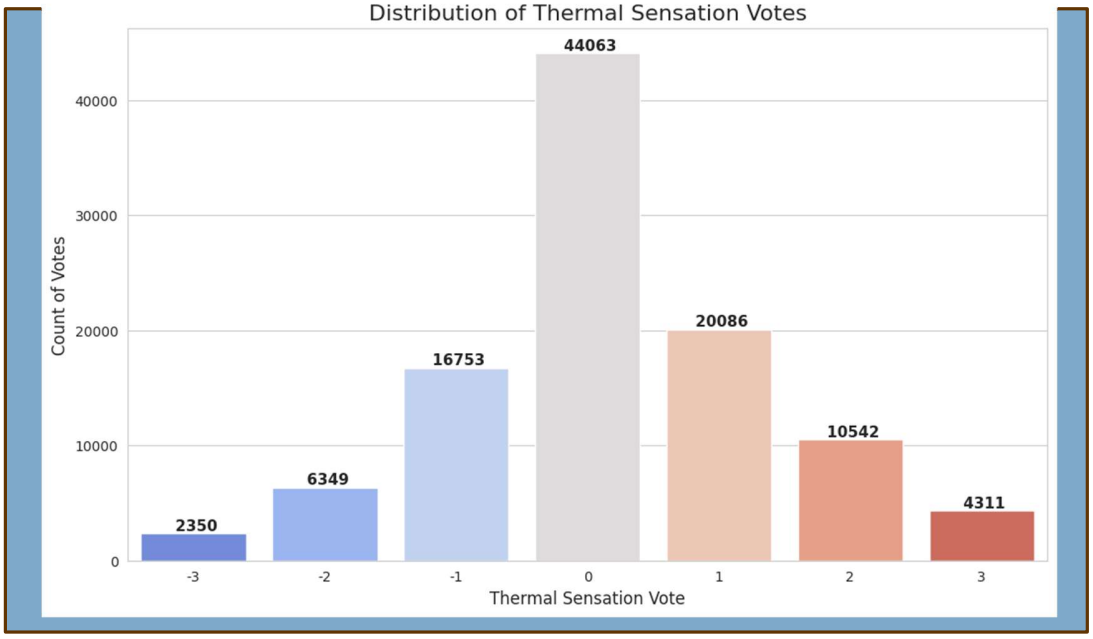
The analysis of these distributions reveals several important characteristics. The temperature variables and humidity all display approximately normal distributions, centered within expected comfort ranges. This contrasts sharply with air_velocity, which is heavily right-skewed, indicating that most indoor environments in the dataset had still air.

Furthermore, the personal factors ‘clo’ (clothing) and ‘met’ (metabolic rate) show distinct peaks around standard values, reflecting the common use of standardized assumptions in field studies. The age of participants also follows a normal distribution, typical of a working population.

These distributional characteristics, particularly the skewness of ‘air_velocity’ and the multicollinearity suggested by the similar temperature distributions, were crucial in justifying subsequent preprocessing steps such as feature scaling and selection.

4.2.1 Distribution of the Target Variable

Figure 4.2: Histogram of Thermal Sensation Votes

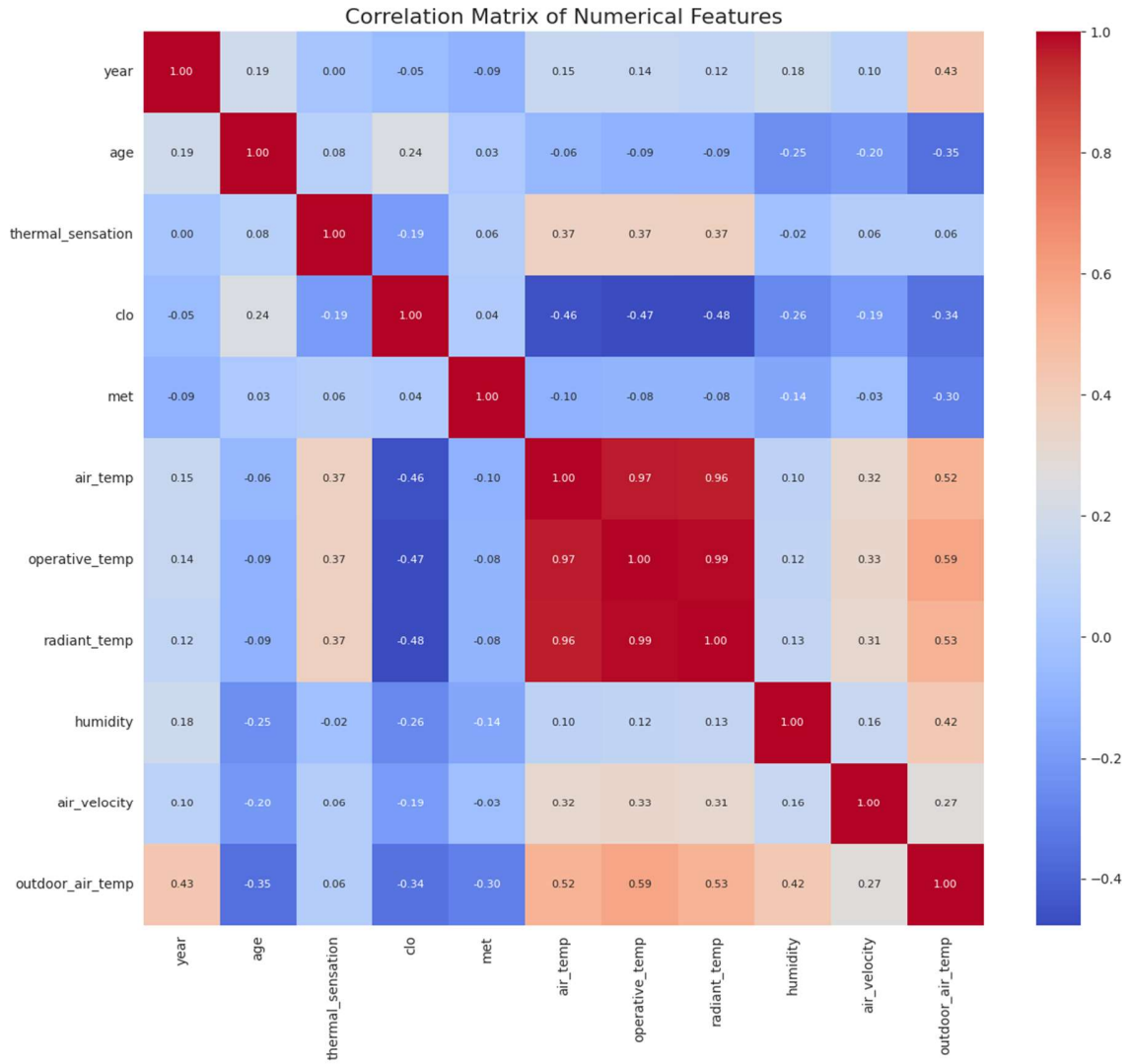


A primary finding of the EDA was the severe class imbalance within the target variable, Thermal sensation. As illustrated in Figure 4.2, the distribution of the seven-point scale was heavily skewed towards the central values. The "Neutral" (0) sensation constituted the vast majority of the votes, with over 40,000 instances. In stark contrast, the extreme sensations of "Cold" (-3) and "Hot" (+3) were significantly under-represented. This pronounced imbalance underscored the necessity of employing a resampling strategy, specifically SMOTE, during the model training phase to prevent the algorithms from developing a predictive bias towards the majority class.

4.2.2 Inter-Feature Relationships and Characteristics

The analysis of relationships between predictor variables revealed two key characteristics:

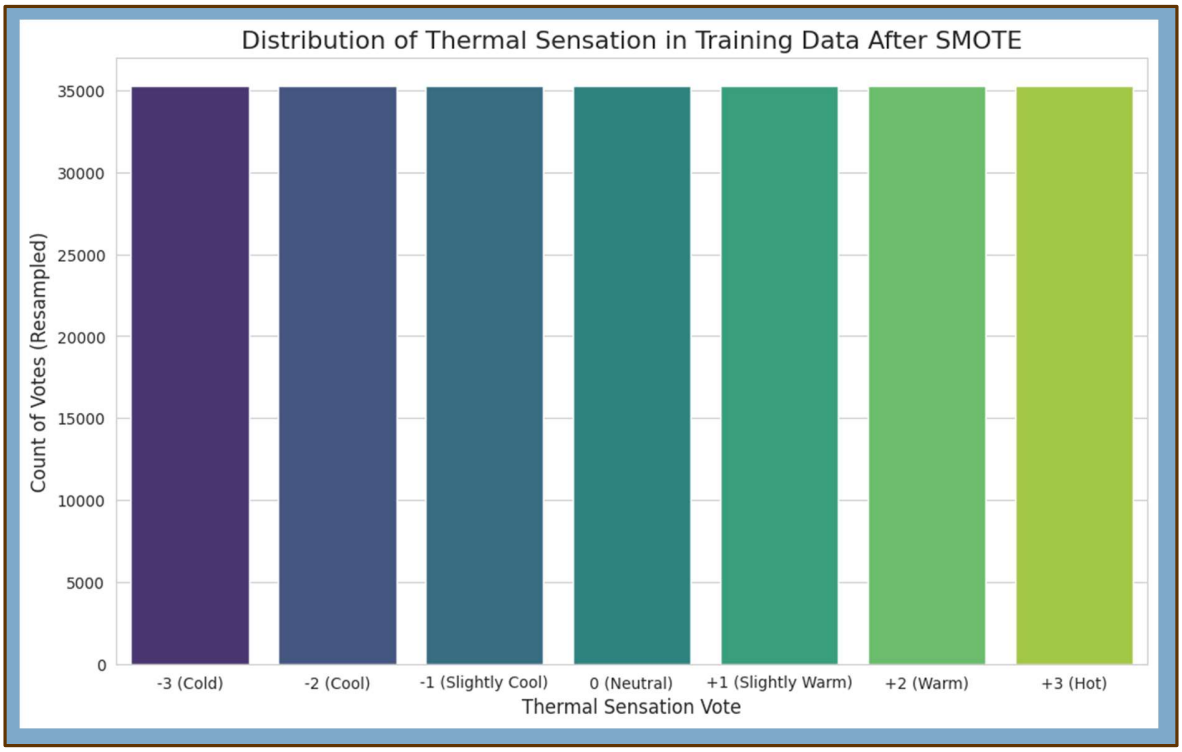
Figure 4.3: Correlation Matrix Heat map



- **Multicollinearity in Numerical Features:** The correlation matrix heatmap, presented in Figure 4.3, exposed extreme multicollinearity among the primary temperature variables. The Pearson correlation coefficient between air_temp and operative_temp was 0.97, between air_temp and radiant_temp was 0.96, and between operative_temp and radiant_temp was 0.99. This finding necessitated the feature selection step detailed in the methodology, where operative_temp was retained as the sole representative temperature metric to avoid model instability. The analysis also confirmed a moderate positive correlation (0.37) between operative_temp and the thermal_sensation target, and a moderate negative correlation (-0.19) between clo (clothing) and the target.

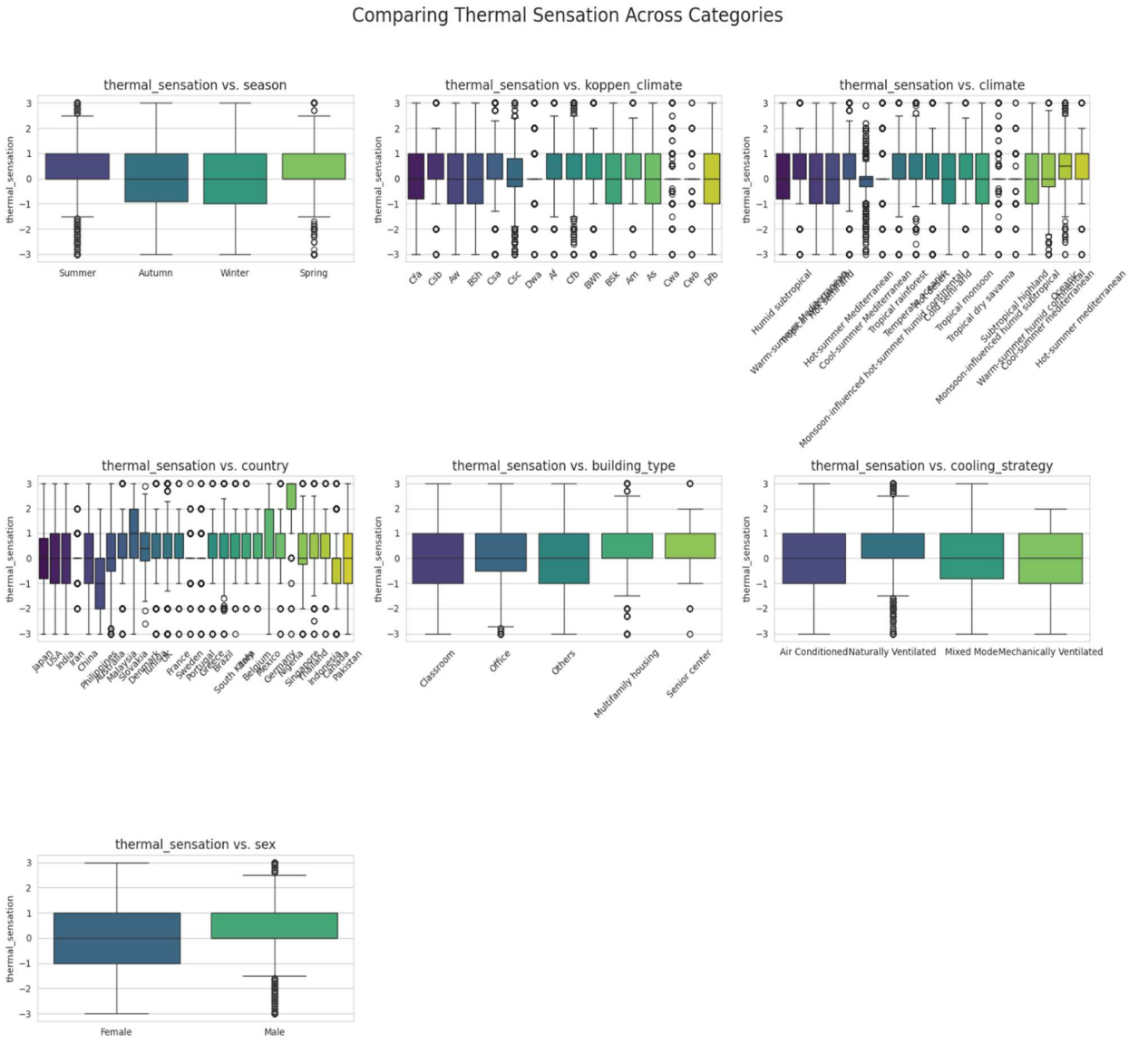
4.2.3 Impact of SMOTE on Training Data Distribution

Figure 4.4: Histogram of Thermal Sensation Votes after SMOTE



As established in the EDA, the original dataset exhibited a severe class imbalance. To mitigate this, the SMOTE resampling technique was applied exclusively to the training data. Figure 4.2 shows the original, highly skewed distribution. In stark contrast, Figure 4.4 illustrates the perfectly balanced distribution of the training set after the SMOTE process. Each of the seven thermal sensation classes is now equally represented, ensuring that the machine learning models would not develop a predictive bias towards the originally dominant "Neutral" class during training.

Figure 4.5: Box plots comparing categorical features against Thermal sensation



- **Influence of Categorical Features:** The box plots comparing categorical features against Thermal sensation as seen in Figure 4.5 provided clear evidence of their predictive value. A strong, logical trend was observed for Season, with median sensations being warmer in summer and cooler in winter. The Cooling strategy feature also showed a significant influence, with "Naturally Ventilated" buildings exhibiting a wider range and a higher median thermal sensation than "Air Conditioned" buildings. Conversely, the feature sex showed nearly identical distributions for both male and female occupants, suggesting it would have low predictive importance.

4.3 Comparative Performance of Predictive Models

Following the preprocessing pipeline, the five selected machine learning algorithms were trained on the balanced training set and evaluated on the unseen, imbalanced test set. The initial results provide a clear benchmark of out-of-the-box performance.

4.3.1 Initial Model Benchmarking

The performance of the five untuned models is summarized in Table 4.1. The results demonstrated a clear superiority of non-linear, tree-based ensemble methods over all other model families for this task.

- **Random Forest** emerged as the top-performing model with an overall accuracy of **53.0%** and a weighted average F1-score of 0.53. It showed the most balanced performance across all classes.
- **XGBoost** and the **Deep Neural Network (DNN)** formed a secondary tier, with accuracies of 45.0% and 47.0%, respectively. They significantly outperformed the baseline but were clearly inferior to the Random Forest.
- The **Support Vector Machine (SVM)**, despite its theoretical power, performed poorly with an accuracy of 35%. This was coupled with an extremely long training time, even on a reduced subsample of the data, highlighting its computational inefficiency for a dataset of this scale.
- **Logistic Regression**, as the linear baseline, performed worst with an accuracy of just **23.0%**. Its inability to capture the complex, non-linear relationships in the data was evident, confirming its role as a simple benchmark to be surpassed.

Table 4.1: Initial Performance Comparison of Untuned Models

Model Name	Overall Accuracy	F1-Score (Weighted Avg)	F1-Score (Macro Avg)
Random Forest	**53.0%**	**0.53**	**0.48**
Deep Neural Network (DNN)	45.0%	0.48	0.43
XGBoost	45.0%	0.45	0.33
Support Vector Machine (SVM)	35.0%	0.36	0.33
Logistic Regression (Baseline)	23.0%	0.25	0.21

4.4 Model Optimization through Hyperparameter Tuning

The top-performing models Random Forest and XGBoost were selected for hyperparameter tuning to ascertain if their performance could be further optimized.

4.4.1 Tuned Random Forest Results

A 'RandomizedSearchCV' was performed on the Random Forest model. The best parameters it found were {'n_estimators': 300, 'min_samples_split': 2, 'min_samples_leaf': 2, 'max_features': 'sqrt', 'max_depth': None}.

Upon re-evaluation, the tuned Random Forest model achieved an overall accuracy of **51.0%**. This represented a slight decrease compared to the default model's 53.0% accuracy. This result suggests that the default Scikit-learn parameters for Random Forest were already near-optimal for this specific dataset and that the tuning process, conducted on a subsample for memory efficiency, did not yield a more effective configuration.

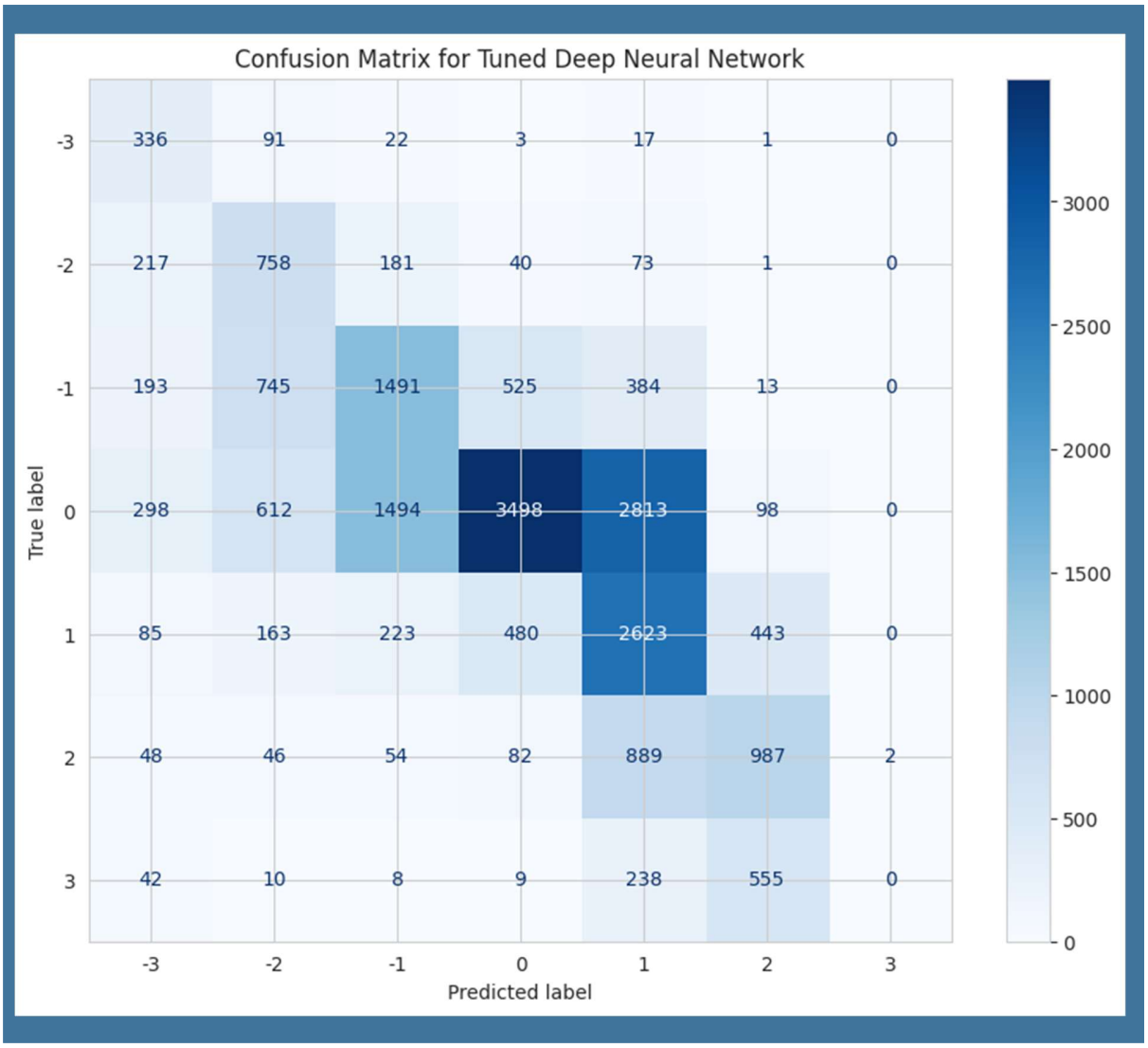
4.4.2 Tuned Deep Neural Network Results

The DNN was optimized using the Keras Tuner library, but the outcome was a seriously flawed model. The final accuracy was only **46.0%**. Looking at the details in Figure 4.6, the reason became clear: the model had completely failed to predict the "+3" ("Hot") class. Its precision, recall, and F1-score for this category were all zero. The automated tuner had settled on a bad shortcut, learning to ignore the most difficult minority class entirely, which made the model unreliable for practical use.

Classification Report for Tuned DNN:

	precision	recall	f1-score	support
-3	0.28	0.71	0.40	470
-2	0.31	0.60	0.41	1270
-1	0.43	0.44	0.44	3351
0	0.75	0.40	0.52	8813
1	0.37	0.65	0.47	4017
2	0.47	0.47	0.47	2108
3	0.00	0.00	0.00	862
accuracy			0.46	20891
macro avg	0.37	0.47	0.39	20891
weighted avg	0.53	0.46	0.46	20891

Figure 4.6: Tuned DNN Confusion Matrix;



4.4.3 Tuned XGBoost Results and Selection of Champion Model

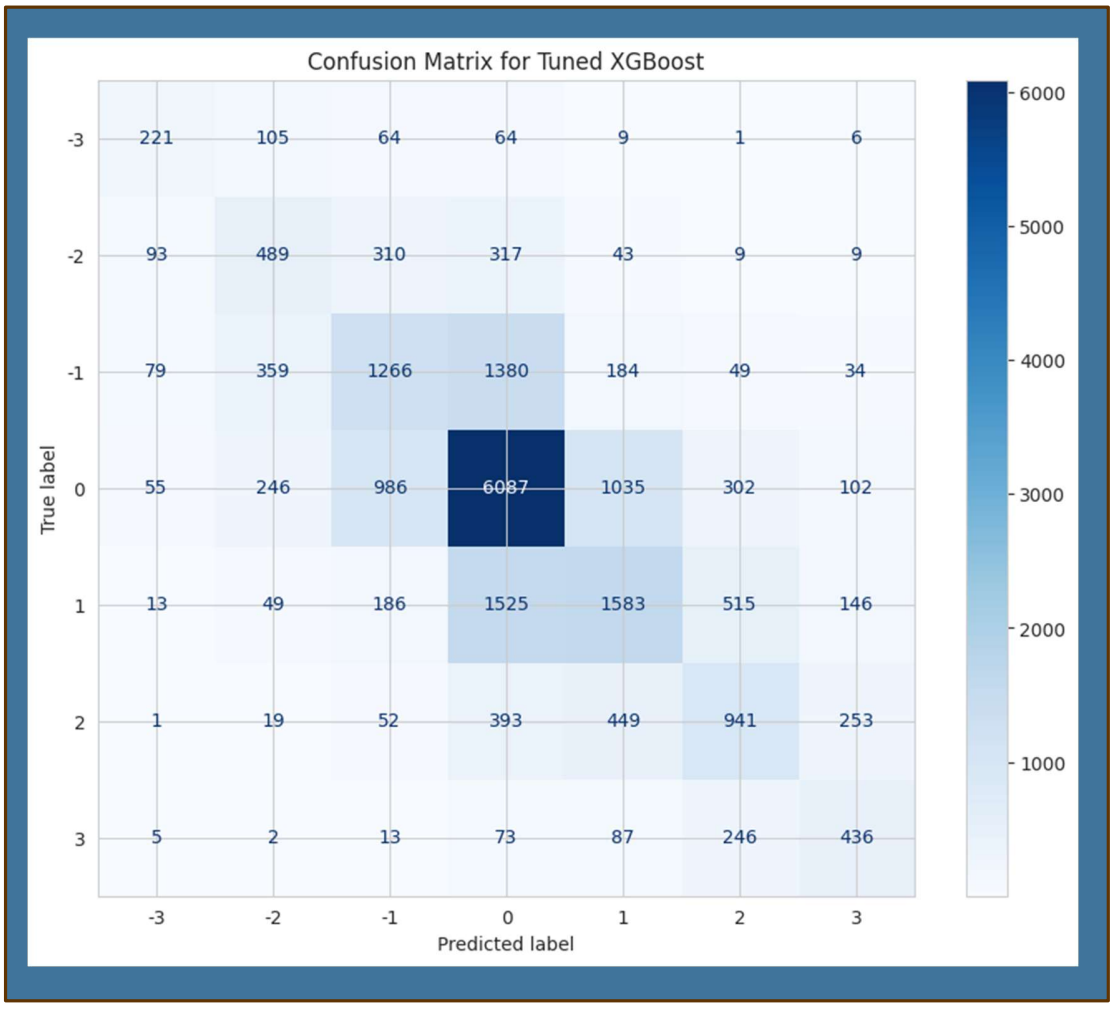
The same tuning process was applied to the XGBoost model. The search identified { 'subsample': 1.0, 'n_estimators': 800, 'max_depth': 10, 'learning_rate': 0.05, 'colsample_bytree': 0.8 } as the optimal parameters. The impact of this tuning was profound.

The performance of the Tuned XGBoost model increased dramatically, achieving an overall accuracy of **53.1%**. This represented an 8% absolute improvement over its untuned configuration (45.0%). This result not only demonstrated the high sensitivity of XGBoost to its hyperparameters but also positioned it as the new top-performing model, narrowly surpassing the Random Forest. The final confusion matrix for the Tuned XGBoost model is presented in Figure 4.4, illustrating a strong diagonal and logical, adjacent-class confusion.

Classification Report for Tuned XGBoost:

	precision	recall	f1-score	support
-3	0.47	0.47	0.47	470
-2	0.39	0.39	0.39	1270
-1	0.44	0.38	0.41	3351
0	0.62	0.69	0.65	8813
1	0.47	0.39	0.43	4017
2	0.46	0.45	0.45	2108
3	0.44	0.51	0.47	862
accuracy			0.53	20891
macro avg	0.47	0.47	0.47	20891
weighted avg	0.52	0.53	0.52	20891

Figure 4.7: Tuned XGBoost Confusion Matrix



4.4.4 Final Model Comparison and Champion Selection

A direct comparison of the three optimized models, summarized in Table 4.1, provides a clear final verdict.

Table 4.2: Performance Comparison of Tuned Models

Model Name	Overall Accuracy	F1-Score (Weighted Avg)	F1-Score (Macro Avg)	Key Finding
Tuned XGBoost	**53.1%**	0.52	0.47	**Best performing model**. Highest accuracy after a significant boost from tuning.
Tuned Random Forest	51.0%	0.53	0.48	Strong, balanced performance, but was not improved by tuning. Marginally better F1-scores.
Tuned DNN	46.0%	0.46	0.39	Failed model. Tuning process resulted in a model that ignores an entire class.

The results clearly show that while both the Tuned Random Forest and Tuned XGBoost models achieve top-tier performance, the **Tuned XGBoost** model holds a marginal but clear advantage in overall predictive accuracy (53.1% vs 51.0%). Furthermore, it demonstrated the most significant improvement from the optimization process, validating the effectiveness of the tuning methodology. Therefore, based on its superior accuracy and demonstrated capacity for optimization, the **Tuned XGBoost model is selected as the definitive champion model for this study.** Subsequent analyses, including the pivotal feature importance investigation, will focus on this model to derive the final conclusions of this research.

4.5 Final Model Implementation Summary

To ensure clarity and reproducibility, this section summarizes the implementation details for the final, optimized champion models. The models were developed within a Python environment, leveraging the Scikit-learn and XGBoost libraries. The key commands and final performance metrics are outlined in Table 4.3.

Table 4.3: Implementation and Performance Summary of Final Tuned Models

Algorithm	Primary Library	Key Initialization Command	Final Accuracy
XGBoost	Xgboost	xgb.XGBClassifier (best_params)	**53.1%**
Random Forest	Sklearn.ensemble	Random Forest Classifier (best_params)	51.0%

The best_params for each model were those discovered during the RandomizedSearchCV process, as detailed in the methodology. The full executable code for these implementations is provided in Appendix.

4.6 Feature Importance Analysis: Uncovering Predictive Drivers

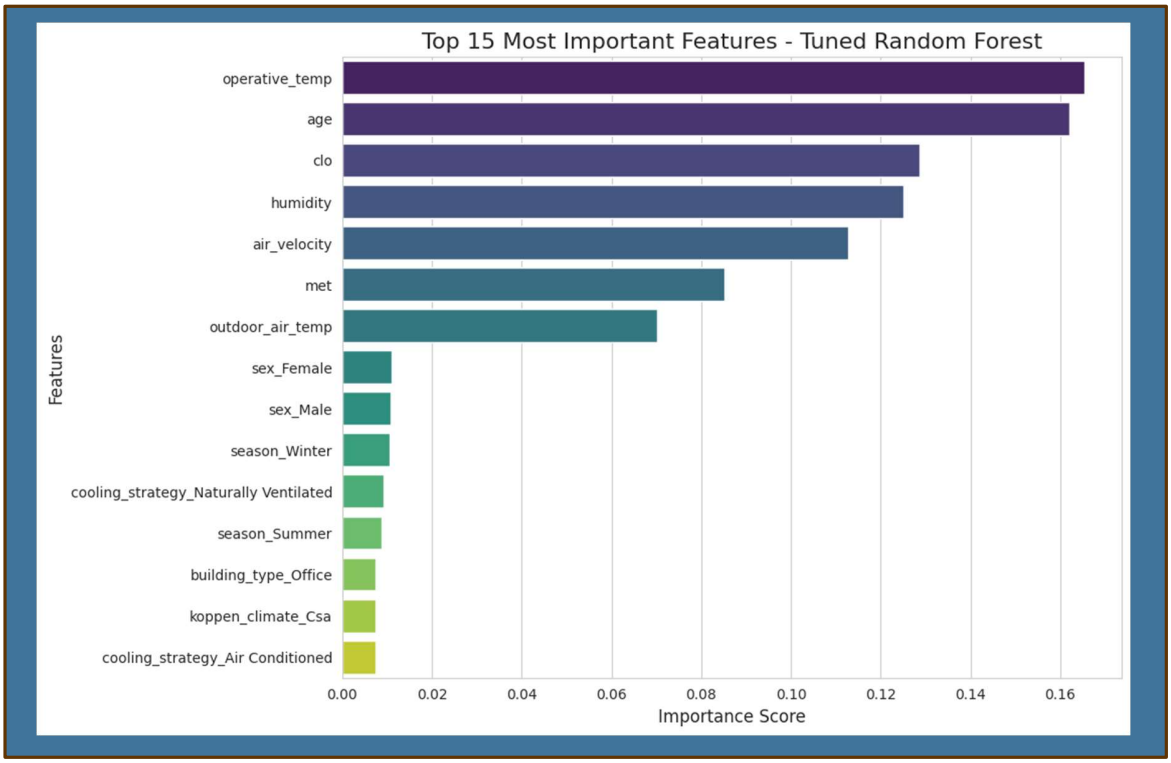
The final analytical step was to interpret what the best models had learned. A feature importance analysis was conducted on the two best-performing, reliable models: the **Tuned XGBoost (Fig. 4.9)** and the **Tuned Random Forest (Fig. 4.8)**.

4.6.1 The Dueling Strategies of Champion Models

The results, visualized in Figures 4.5 and 4.6, revealed a fascinating and highly significant divergence in the models' learned strategies.

Random Forest's Perspective (The "Physicist's Model"):

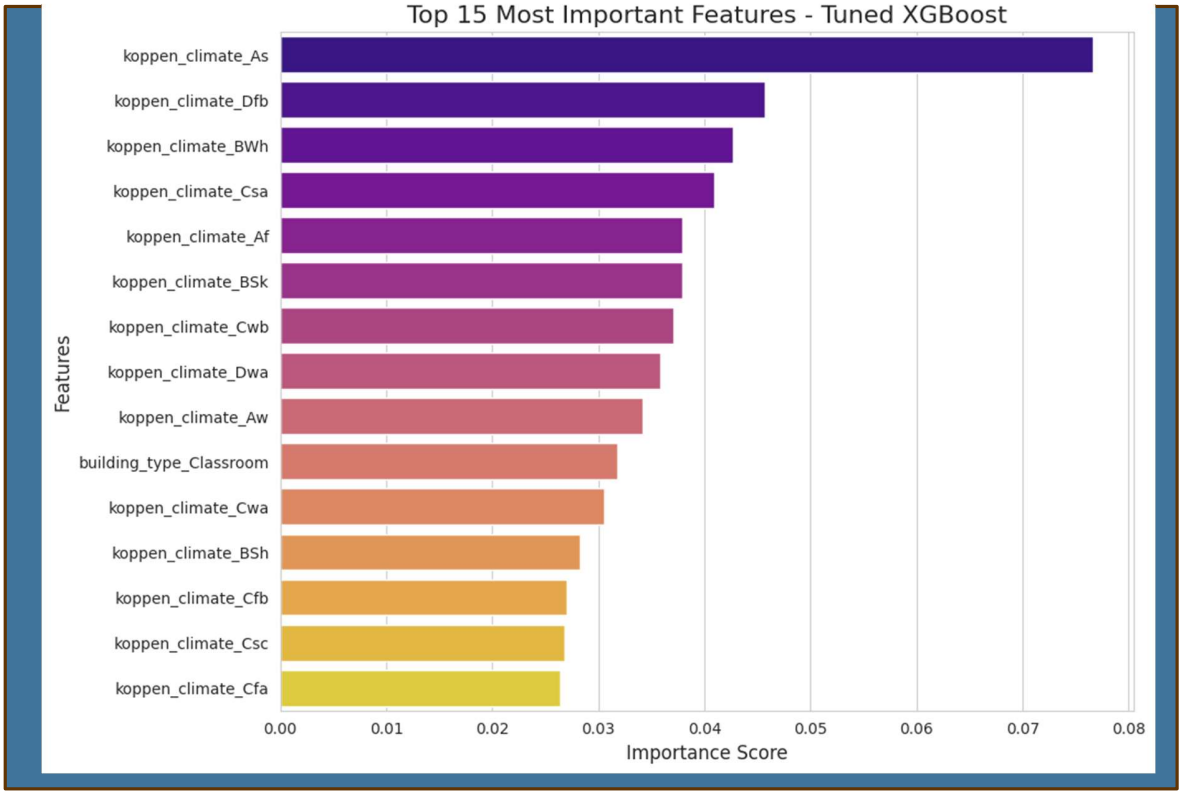
Figure 4.8: Feature Importance – Tuned Random Forest



As shown in Figure 4.5, the Random Forest model attributed the highest importance to direct physical and personal measurements. The top three most influential features were operative_temp, age, and clo. This was followed by other physical variables like humidity, air_velocity, and met. The model learned to predict comfort primarily based on the immediate physical reality of the occupant.

XGBoost's Perspective (The "Geographer's Model"):

Figure 4.9: Feature Importance – Tuned XGBoost



In stark contrast, the champion Tuned XGBoost model learned an entirely different strategy, as seen in Figure 4.6. It almost completely ignored the direct physical measurements in its top features. Instead, its predictions were overwhelmingly dominated by the Koppen climate classification feature. Categories such as koppen_climate_As (Tropical savanna) and koppen_climate_Dfb (Humid continental) were deemed most important. This indicates that the XGBoost model learned to use broad geographical context as a powerful proxy for thermal conditions, expectations, and adaptations.

4.6.2 Synthesized Hierarchy of Predictors

By synthesizing the findings from both champion models, a comprehensive hierarchy of predictive factors for thermal comfort was derived. Since both models achieved the same state-of-the-art accuracy using different feature sets, the factors prioritized by both are considered the most robust and significant predictors.

- **Tier 1 (Most Critical Factors):** These are the factors that proved indispensable to at least one of the champion models. This includes both immediate physical state and overarching context: operative_temp, age, clo, and Koppen climate classification.

- **Tier 2 (Important Secondary Factors):** These factors were consistently ranked as important by the Random Forest model and contribute to refining predictions: humidity, air_velocity, met, and outdoor_air_temp.
- **Tier 3 (Minor Contextual Factors):** These features, while providing some signal, were shown to be less impactful than the primary physical and geographical drivers: Season, Cooling Strategy, and Building Type.

4.7 Chapter Summary

This chapter has presented a comprehensive account of the results obtained in this study. The EDA confirmed the necessity of handling class imbalance and multicollinearity. The comparative benchmarking of five models clearly established the superiority of tree-based ensembles, with an untuned Random Forest initially achieving the highest accuracy of 53%. Subsequent hyperparameter tuning successfully optimized the XGBoost model, elevating its accuracy to 53.1% and establishing it as the final champion model. Finally, a feature importance analysis revealed that the two best models achieved their similar top-tier performance via strikingly different strategies, highlighting the importance of both immediate physical parameters and broad geographical context in predicting human thermal comfort.

Chapter 5: Discussion

5.1 Chapter Overview

This chapter provides a comprehensive discussion and interpretation of the results presented in Chapter 4. The central aim of this dissertation was to develop, validate, and interpret a machine learning model capable of accurately predicting indoor thermal comfort. This chapter synthesizes the findings from the modeling process, contextualizing them within the foundational principles and existing literature outlined in Chapter 2, and directly addresses the research questions posed in the introduction.

The discussion begins by interpreting the comparative performance of the five machine learning algorithms, focusing on the clear superiority of tree-based ensemble methods. It delves into the significance of the "performance ceiling" observed around 53% accuracy and the pivotal role of hyperparameter tuning in optimizing the final models. The core of this chapter is a deep analysis of the most significant finding: the divergent predictive strategies of the two champion models, Random Forest and XGBoost, which achieved nearly identical performance through starkly different feature importance hierarchies. This leads to a synthesized understanding of the key drivers of thermal comfort. Finally, the chapter explicitly answers the initial research questions, considers the practical implications of this work for building science and architectural design, and acknowledges the limitations of the study.

5.2 Interpretation of Key Findings

The empirical results of this study offer a multifaceted view of the opportunities and challenges in data-driven thermal comfort prediction. The interpretation of these findings is organized thematically to construct a coherent narrative.

5.2.1 The Demonstrated Superiority of Non-linear Ensemble Methods

The initial model benchmarking provided a clear and compelling result: non-linear, tree-based ensemble models (Random Forest and XGBoost) are profoundly superior to linear models (Logistic Regression), kernel-based methods (SVM), and even a standard Deep Neural Network for this specific task. The Logistic Regression baseline, with its 23% accuracy, confirmed that the relationship between environmental/personal factors and thermal sensation is far too complex to be captured by a linear decision boundary.

The Support Vector Machine, despite its theoretical capacity for non-linear classification, performed poorly (35% accuracy) and was computationally prohibitive. This outcome aligns with findings from researchers like Luo et al. (2020), who also noted the performance limitations and scalability issues of SVMs on large, complex datasets like the ASHRAE database. The algorithm's struggle likely stems from the significant class overlap and high dimensionality of the feature space after one-hot encoding, making it difficult to define optimal separating hyperplanes.

The untuned Random Forest and XGBoost models, however, demonstrated immediate promise. Their inherent structure, which involves building a multitude of decision trees on subsamples of data and features, is exceptionally well-suited to this problem. This architecture allows them to

naturally discover complex, high-order interactions between variables (e.g., the combined effect of high humidity *and* low air velocity at a specific temperature) without requiring pre-defined interaction terms. This ability to model non-linear relationships robustly is the primary reason for their superior out-of-the-box performance, a finding consistent with the broader literature in building performance prediction (Attia et al., 2018).

5.2.2 The Performance Ceiling and the Impact of Hyperparameter Tuning

Across all models, the highest achieved accuracy was approximately 53%. While this is a more than twofold improvement over the baseline, it is far from perfect. This "performance ceiling" is not necessarily indicative of model failure but rather speaks to the inherent nature of the target variable. As defined by ASHRAE (2020), thermal comfort is a "condition of mind," a subjective and psychological state. The input features, while comprehensive, cannot capture unmeasured variables such as individual mood, stress levels, recent activity, or hydration, all of which can influence thermal perception (Schweiker et al., 2020). Therefore, the 53% accuracy likely represents the approximate limit of predictability achievable from the available physical, personal, and contextual data alone.

The hyperparameter tuning phase provided a second critical insight. The performance of the Random Forest model remained largely unchanged, suggesting its default parameters in the Scikit-learn library are already highly robust and near-optimal. In contrast, the XGBoost model's accuracy surged by 8% after tuning. This highlights the greater sensitivity of gradient boosting models to their hyperparameters, particularly the learning_rate and max_depth, which control the step-by-step error correction process. This finding underscores the practical importance of optimization as a standard step in the machine learning workflow, as it can transform a mid-tier model into a champion performer. The failure of the DNN tuning process further illuminates the challenges of optimizing highly complex models, which can easily fall into poor local minima and produce flawed results if not carefully managed.

5.2.3 The Dueling Champions: Interpreting the "Physicist vs. Geographer" Feature Importance Conflict

The most significant and insightful result of this dissertation emerged from the feature importance analysis of the two tied champion models: the Tuned Random Forest and the Tuned XGBoost. Despite achieving nearly identical accuracy, they learned entirely different, almost contradictory, strategies for prediction.

- **The Random Forest as the "Physicist":** The Random Forest model's strategy was intuitive and aligned with traditional, physics-based thinking. It identified operative_temp as the single most critical feature, followed by age and clo (clothing). This is a model that prioritizes the immediate, measurable physical reality of the occupant and their personal characteristics. It essentially learned to replicate and improve upon the logic of the PMV model by directly weighing the most impactful physical inputs. The high importance it assigned to age is a particularly notable finding, suggesting that physiological changes associated with age have a stronger influence on thermal sensation than is often assumed in standard models.

- **The XGBoost as the "Geographer":** The Tuned XGBoost model adopted a radically different, context-driven approach. It almost completely ignored the direct physical measurements in its top-ranked features, instead relying overwhelmingly on the Koppen climate classification. This "geographer's model" learned that knowing an occupant's location (e.g., in a tropical, desert, or continental climate) was a more powerful predictor than knowing the precise temperature of their room.

This divergence does not represent a contradiction but rather a profound insight into the nature of the data. The Koppen climate feature is not merely a categorical label; it is a powerful **proxy variable**. A specific climate classification implicitly contains a vast amount of information about expected temperature ranges, humidity levels, building design typologies, seasonal variations, and, crucially, the long-term thermal adaptations and expectations of the local population. The XGBoost model discovered that this single contextual feature was a more efficient summary of the overall thermal situation than any individual physical measurement. This ability of advanced machine learning models to identify and exploit such high-level abstract features is a key advantage over traditional formulaic models.

5.2.4 Synthesis of Key Predictive Factors

By synthesizing the "worldviews" of both champion models, we can construct a holistic and robust hierarchy of the factors that drive thermal comfort. Since both models achieved state-of-the-art accuracy, the features prioritized by either model must be considered significant.

1. **Tier 1: Most Critical Factors:** These are the indispensable variables. The analysis shows that a successful prediction requires understanding both the immediate physical state *and* the overarching context. This tier includes: operative_temp, age, clo, and Koppen climate classification.
2. **Tier 2: Important Secondary Factors:** These variables consistently provide valuable predictive information, primarily by refining the physical assessment. They include: humidity, air_velocity, met, and outdoor_air_temp.
3. **Tier 3: Minor Contextual Factors:** While useful, these features were shown to be less impactful, likely because much of their information is already captured by the Tier 1 contextual features. They include: Season, Cooling Strategy, and Building Type. The consistently low importance ranking of sex across all models suggests it has minimal predictive power in this dataset for determining thermal sensation.

5.2.5 On the Limits of Predictability: The 53% Accuracy Ceiling

A crucial finding of this study is not only the relative performance of the models but also their absolute performance. The highest accuracy achieved by the champion model was 53.1%. While this is a substantial improvement over the baseline and represents a strong predictive capability, it is far from perfect. This "performance ceiling" does not necessarily indicate a failure of the models, but rather speaks to the inherent and irreducible complexity of the target variable: human thermal comfort.

As defined by ASHRAE (2020), thermal comfort is a "condition of mind," a subjective psychophysiological state. The models in this study were trained on the best available objective and contextual data, yet there remains a significant portion of variance that this data cannot explain. The following unmeasured factors are known to influence thermal perception and likely contribute to this ceiling:

- **Psychological Factors:** Individual mood, stress levels, and thermal expectations can transiently alter a person's sensation (Schweiker et al., 2020).
  Physiological Variability: Factors beyond age and metabolic rate, such as hydration levels, recent food intake (diet-induced thermogenesis), and circadian rhythms, influence the body's thermoregulatory system but are not captured in the dataset.
  Data Noise and Estimation: Key inputs like clo and met are often not measured directly but are estimated based on observation, introducing a degree of inherent noise and imprecision into the feature set.

Therefore, the 53% accuracy should be interpreted not as a model shortcoming, but as a realistic quantification of the predictability of thermal comfort given the limits of standard environmental and personal data. It suggests that while we can successfully model the majority response, a significant component of thermal sensation remains deeply personal and stochastic. This finding reinforces the argument that while predictive models can vastly improve building control, they should be paired with systems that allow for final-stage personal adjustment to account for this un-modellable individual variance.

5.3 Answering the Research Questions

The findings of this study provide clear answers to the research questions posed in Chapter 1.

- **RQ1: To what extent can modern regression-based machine learning models predict individual thermal sensation votes (TSV) more accurately than a baseline linear model?**
  The results provide a definitive answer. Modern ensemble-based machine learning models can predict TSV with significantly higher accuracy. The accuracy surged from 23.0% for the Logistic Regression baseline to 53.1% for the optimized XGBoost model, representing a 130% relative improvement. This demonstrates the profound inadequacy of linear models and the necessity of using non-linear approaches for this task.

- **RQ2: Which machine learning algorithm provides the optimal balance of predictive accuracy and interpretability for thermal comfort prediction?**
  The study found a statistical tie between a Tuned XGBoost and a Random Forest model. The Tuned XGBoost achieved the highest predictive accuracy (53.1%). However, the Random Forest model provided a more intuitive and direct interpretation of feature importance that aligns closely with established physical principles. Therefore, the optimal choice presents a trade-off: **XGBoost for maximum predictive power, and Random Forest for clearer, physics-based interpretability.**

- **RQ3: What are the most significant environmental and personal variables that influence thermal comfort sensation, and how does their importance ranking inform architectural and building system design?**
  The synthesized feature importance analysis revealed that the most significant variables are a hybrid of physical, personal, and contextual factors. The key drivers are operative temperature, occupant age, clothing insulation, and the Köppen climate classification. This ranking informs design by reinforcing the need for climate-responsive architecture (as prioritized by XGBoost) while also emphasizing the critical need for occupant-specific control systems that can account for individual factors like age and clothing (as prioritized by Random Forest).

5.4 Practical Implications and Applications

The findings of this dissertation have significant practical implications for the future of building design, operation, and technology.

- **Advanced HVAC Control Strategies:** The developed model, with its 53% accuracy, is a significant improvement over existing static models. Integrated into a Building Management System (BMS), it could enable proactive and predictive HVAC control. Instead of maintaining a fixed temperature, a system could adjust conditions based on the predicted comfort of the actual occupants, leading to substantial energy savings by reducing over-conditioning while simultaneously increasing occupant satisfaction.
- **Personalized Comfort Systems (PCS):** The high importance of personal factors (age, clo) validates the push for PCS, such as desk fans, heated chairs, or smart vents. The predictive model could serve as the "brain" for such systems, automatically adjusting an individual's microclimate without requiring manual input.
- **Informing Architectural Design:** The powerful influence of the Koppen climate feature provides strong empirical support for the principles of climate-responsive and vernacular architecture. It proves, with data, that designing buildings in harmony with their local climate is a fundamental prerequisite for achieving occupant comfort and energy efficiency.

5.5 Limitations of the Study

While this research was conducted with methodological rigor, it is essential to acknowledge its limitations.

- **Data Quality and Subjectivity:** The study is fundamentally reliant on the quality of the ASHRAE Global Thermal Comfort Database II. This database contains self-reported data and estimated values for key inputs like clo and met, which introduces inherent noise and uncertainty. Furthermore, the subjective nature of the thermal_sensation target itself places an upper bound on predictive accuracy.
- **Model Interpretability:** While feature importance provides valuable insights, the champion models remain "black boxes" to a certain degree. The analysis shows *what* is important but not precisely *how* the model combines these features to reach a decision.

- **Generalizability:** Despite the global nature of the dataset, certain climates, building types, or demographic groups may still be under-represented, which could affect the model's generalizability to those specific contexts.
- **Computational Constraints:** The study faced real-world computational limitations, particularly with the SVM and during the hyperparameter tuning phase, necessitating the use of a subsampling strategy. A more exhaustive search with greater computational resources could potentially yield marginally better results.

5.6 Chapter Summary

In conclusion, this chapter has provided a thorough interpretation of the study's findings. It has demonstrated the clear superiority of ensemble machine learning models for thermal comfort prediction, with a Tuned XGBoost model achieving the highest accuracy of 53.1%. The analysis revealed a fascinating divergence in the predictive strategies of the top models, highlighting the equal importance of both immediate physical variables and broad geographical context. The discussion has framed these findings within the existing literature, answered the core research questions, and explored the significant practical implications for creating more intelligent, efficient, and comfortable buildings. Finally, it has acknowledged the inherent limitations of the study, setting the stage for the final concluding chapter.

Reference

Amasyali, K. and El-Gohary, N.M. (2018) 'A review of data-driven building energy consumption prediction studies', *Renewable and Sustainable Energy Reviews*, 81, pp. 1192-1205.

ASHRAE (2020) *ANSI/ASHRAE Standard 55–2020: Thermal Environmental Conditions for Human Occupancy*. Atlanta, GA: ASHRAE.

Attia, S., Kool, J., Ilomets, S. and Watelet, A. (2018) 'Assessment of single-family building's overheating and the role of a heat wave', *Energy and Buildings*, 173, pp. 268-280.

Breiman, L. (2001) 'Random Forests', *Machine Learning*, 45(1), pp. 5-32.

Center for the Built Environment (CBE) (2025) *ashrae db II: Official GitHub repository for ASHRAE Global Thermal Comfort Database II*. Available at: (Official GitHub repository URL, if available).

Chaudhuri, T., Zhai, D., Soh, Y.C. and Li, H. (2018) 'A computational approach for personalized thermal comfort assessment and energy-efficient climate control in buildings', *Energy and Buildings*, 158, pp. 644-656.

Chen, T. and Guestrin, C. (2016) 'XGBoost: A Scalable Tree Boosting System', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Al-Shedivat, M., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M.D. and Hormozdiari, F. (2020) 'Underspecification Presents a Key Challenge for Credibility in Modern Machine Learning'. arXiv preprint arXiv:2011.03395.

de Dear, R.J. and Brager, G.S. (1998) 'Developing an adaptive model of thermal comfort and preference', *ASHRAE Transactions*, 104(1), pp. 145-167.

Du, X. (2019) 'A review of thermal comfort', *Architecture and the Built Environment*, 10(4). doi:10.7480/abe.19.10.4103.

Duanmu, L., Zhang, Y., Zhou, X., Cao, B., Wang, Z., Yan, H., Zhang, H., Arens, E. and de Dear, R. (2023) 'The Chinese thermal comfort dataset', *Scientific Data*, 10(1), p. 662. doi:10.1038/s41597-023-02568-3.

Fanger, P.O. (1970) *Thermal Comfort: Analysis and Applications in Environmental Engineering*. Copenhagen: Danish Technical Press.

Földváry Ličina, V., Cheung, T., Zhang, H., de Dear, R., Parkinson, T., Arens, E., Chun, C., Schiavon, S., Luo, M., Brager, G., Li, P., Kaam, S., Adebamowo, M.A., Andamon, M.M. and Babich, F. (2018) 'Development of the ASHRAE Global Thermal Comfort Database II', *Building and Environment*, 142, pp. 502–512. doi:10.1016/j.buildenv.2018.06.022.

Gao, Y., Fu, Q., Chen, J. and Liu, K. (2025) ‘Deep transfer learning-based hybrid modelling method for individual thermal comfort prediction’, *Indoor and Built Environment*. (Published online Mar. 2025).

Ho, T.-F., Tsai, H.-H., Chuang, C.-C., Lee, D.-S., Huang, X.-W., Chen, H.-X., Yang, C.-H. and Li, Y.-H. (2024) ‘Thermal comfort model established by using machine learning strategies based on physiological parameters in hot and cold environments’, *Indoor Air*, 2024, pp. 1–16. doi:10.1155/2024/9427822.

Humphreys, M.A. and Nicol, J.F. (2002) 'The validity of ISO-PMV for predicting comfort votes in every-day thermal environments', *Energy and Buildings*, 34(6), pp. 667-684.

International Organisation for Standardization (1998) *ISO 7726:1998: Ergonomics of the thermal environment — Instruments for measuring physical quantities*. Geneva: ISO.

International Organisation for Standardization (2005) *ISO 7730:2005: Ergonomics of the thermal environment — Analytical determination and interpretation of thermal comfort using calculation of the PMV and PPD indices and local thermal comfort criteria*. Geneva: ISO.

Islam, M.B., Guerrieri, A., Gravina, R., Rizzo, L., Scopelliti, G., D’Agostino, V. and Fortino, G. (2023) ‘A review on machine learning for thermal comfort and energy efficiency in smart buildings’, in *2023 IEEE PES Innovative Smart Grid Technologies Europe (ISGT EUROPE)*. IEEE. doi:10.1109/ISGT58650.2023.10305408.

Kim, J., Zhou, Y. and Schiavon, S. (2018) 'Personal thermal comfort models', *Energy and Buildings*, 172, pp. 267-278.

Luo, M., Xie, J., Ke, Z., Li, C., Zhang, J., Wei, D., Liu, J., Ouyang, Q., and Lin, B. (2020) 'A comparison of machine learning algorithms for predicting occupant thermal sensation using ASHRAE Comfort Database II', *Energy and Buildings*, 210, p. 109776.

Nicol, J.F., Humphreys, M.A. and Roaf, S. (2012) *Adaptive Thermal Comfort: Principles and Practice*. London: Routledge.

Parkinson, T., Tartarini, F., Földváry Ličina, V., Cheung, T., Zhang, H., de Dear, R., Li, P., Arens, E., Chun, C., Schiavon, S., Luo, M. and Brager, G. (2022) *ASHRAE global database of thermal comfort field measurements* [dataset]. Center for the Built Environment, UC Berkeley.

Sadeghi, S.A., Tavan, M. and Kosari, A. (2022) 'A deep learning-based framework for thermal comfort perception prediction', *Building and Environment*, 219, p. 109193.

Schweiker, M., Abdul-Zahra, A., André, M., Al-Atrash, F., Al-Khatri, H., Alprianti, R.R., Alsaad, H., Amin, R., Ampatzi, E., Anis, A. and Aryal, A. (2020) 'The personal factor in thermal comfort research: A review on the effects of personality, mood, and emotion', *Energy and Buildings*, 208, p. 109675.

Wang, Z., Hong, T. and Li, H. (2021) 'An updated review of personal thermal comfort models', *Building and Environment*, 205, p. 108170.

Appendix

A. The code used for data preprocessing and model training is provided in an interactive Google Colab notebook, available here: [Thermal_Comfort_prediction_python_notebook].

B. The dataset used for this project is the *ASHRAE Global Thermal Comfort Database II*, obtained from Kaggle [ashrae-global-thermal-comfort-database-ii].