MINISTERUL EDUCAȚIEI

UNIVERSITATEA TEHNICĂ
DIN CLUJ-NAPOCA
FACULTATEA DE AUTOMATICĂ ȘI CALCULATOARE

MODELAREA CONFORTULUI TERMIC PENTRU OPTIMIZAREA SISTEMELOR HVAC

PROIECT DE SEMESTRU

Student: **Domide Maria**

Disciplina: **Sisteme bazate pe Cunoaștere**

2025

Cuprins

1 Introducere

1.1 Context general

Sistemele de încălzire, ventilaţie şi aer condiţionat (HVAC) joacă un rol esenţial în asigurarea confortului termic şi a sănătăţii mediului interior, fiind critice atât pentru locuinţe, cât şi pentru spaţiile comerciale sau industriale.

Progresul tehnologic din ultimele decenii a facilitat integrarea unor tehnologii avansate bazate pe date şi algoritmi de învăţare automată, care pot îmbunătăţi semnificativ performanţa acestor sisteme.

În acest context, proiectul nostru se încadrează într-un domeniu multidisciplinar, la intersecţia dintre ingineria mediului interior, analiza datelor şi inteligenţa artificială.

Introducerea acestor tehnologii în sistemele HVAC aduce beneficii considerabile, inclusiv optimizarea consumului de energie, creşterea eficienţei operaţionale şi îmbunătăţirea calităţii vieţii utilizatorilor prin personalizarea condiţiilor de mediu.

Într-o eră în care schimbările climatice şi sustenabilitatea devin priorităţi globale, optimizarea sistemelor HVAC este esenţială pentru reducerea emisiilor de carbon şi atingerea obiectivelor de eficienţă energetică.

Datele climatice şi subiective, cum ar fi temperatura, umiditatea, viteza aerului sau percepţiile utilizatorilor, au un impact direct asupra confortului termic. Astfel, dezvoltarea unor modele predictive care să coreleze aceste variabile este vitală pentru proiectarea unor sisteme HVAC inteligente şi adaptative.

Pentru acest proiect, s-a utilizat baza de date **ASHRAE Global Thermal Comfort Database II**, o resursă cu peste 100.000 de înregistrări care oferă o combinaţie valoroasă de măsurători obiective şi percepţii subiective.

Prin analiza acestor date, proiectul urmăreşte să dezvolte un model avansat care să optimizeze reglajele sistemelor HVAC în timp real, contribuind la o gestionare mai eficientă a energiei şi la crearea unui mediu interior plăcut.

Structura lucrării este organizată astfel:

Capitolul 1 introduce analiza exploratorie a datelor, incluzând analiza univaria(n)tă, bivariată şi multivariată. Se prezintă distribuţia variabilelor, corelaţiile dintre acestea şi vizualizările relevante pentru înţelegerea setului de date.

Capitolul 2 detaliază metodele de preprocesare a datelor utilizate, incluzând curăţarea datelor, interpolarea valorilor lipsă şi eliminarea anomaliilor pentru a asigura o bază de date curată şi coerentă.

Capitolul 3 se concentrează pe analiza datelor, incluzând analiza statistică şi vizualizările realizate pentru a înţelege relaţiile dintre variabile şi structura datelor.

Capitolul 4 detaliază metodologia utilizată în modelare, explicând tehnicile aplicate, evaluarea performanței modelelor predictive și selecția celui mai performant model.

Capitolul 5 prezintă concluziile generale ale proiectului, oferă o sinteză a rezultatelor obținute și propune direcții viitoare de dezvoltare pentru îmbunătățirea soluției.

Această lucrare nu doar că prezintă o soluție inovatoare pentru optimizarea sistemelor HVAC, dar oferă și o bază solidă pentru cercetări ulterioare.

1.2 Obiective

Obiectivele proiectului sunt următoarele:

Dezvoltarea unui model predictiv avansat pentru estimarea confortului termic, folosind variabile climatice obiective (ex. temperatura, umiditatea) și date subiective din baza de date ASHRAE.

Analiza factorilor determinanți ai confortului termic, precum temperatura, umiditatea, viteza aerului și alte condiții de mediu, pentru a înțelege modul în care aceștia influențează percepția utilizatorilor.

Optimizarea performanței modelelor predictive prin aplicarea mai multor algoritmi de învățare automată (regresie liniară, regresii Ridge și Lasso, Random Forest, XGBoost).

Evaluarea performanței modelelor utilizând metrici precum MSE (Mean Squared Error), R^2 Score și, pentru clasificare, acuratețea, precizia și F1-score.

Reducerea consumului energetic al sistemelor HVAC, utilizând modele predictive pentru reglarea automată a parametrilor sistemului, astfel încât să fie adaptate nevoilor utilizatorilor și condițiilor de mediu.

Integrarea unei metodologii reproductibile, care să poată fi implementată în alte contexte sau sisteme similare.

Aceste obiective subliniază abordarea sistematică a proiectului, de la analiza datelor la implementarea practică a unui sistem HVAC inteligent.

1.3 Specificaţii

Proiectul urmăreşte să îndeplinească următoarele specificaţii tehnice şi funcţionale:

1.3.1 Funcţionalităţi principale:

- Dezvoltarea unui model predictiv pentru estimarea PMV (Predicted Mean Vote).
- Analizarea datelor climatice şi subiective pentru identificarea relaţiilor semnificative.
- Optimizarea reglajelor sistemului HVAC pe baza predicţiilor modelului.

1.3.2 Mediul de implementare:

- Limbaj de programare: Python 3.9.
- Biblioteci utilizate: pandas, numpy, matplotlib, scikit-learn, xgboost.
- Mediu de dezvoltare: Jupyter Notebook.
- Date: Baza de date ASHRAE Global Thermal Comfort Database II, disponibilă în format CSV.

1.3.3 Performanţă şi calitate:

- Modelele predictive vor fi evaluate utilizând metrici statistice pentru a asigura o precizie ridicată.
- Sistemul trebuie să ofere predicţii în timp util pentru a putea fi aplicat în medii practice.

1.3.4 Limitări:

- Performanţa modelelor depinde de calitatea şi volumul datelor.
- Datele utilizate sunt limitate la înregistrările disponibile în baza ASHRAE, ceea ce poate afecta generalizarea în alte contexte.

1.3.5 Fiabilitate şi securitate:

- Codul este structurat şi documentat pentru a fi reproductibil.
- Proiectul respectă bunele practici în manipularea şi stocarea datelor pentru a preveni pierderea sau compromiterea acestora.

Prin respectarea acestor specificaţii, proiectul oferă o soluţie robustă, reproductibilă şi eficientă pentru optimizarea confortului termic, cu potenţial de aplicare în diverse medii comerciale şi rezidenţiale.

2 Cunoașterea și analiza setului de date

2.1 Definiție

Confortul termic reprezintă starea de satisfacţie a ocupanţilor unui spaţiu în raport cu mediul termic. Este influenţat de factori subiectivi, precum percepţiile şi preferinţele individuale, dar şi de factori obiectivi, cum ar fi temperatura aerului, umiditatea relativă şi viteza aerului.

2.2 Introducere

În cadrul acestui proiect, am utilizat datele disponibile în **arhiva 1** a bazei de date **ASHRAE Global Thermal Comfort Database II**, care conţine informaţii despre confortul termic colectate în diverse locaţii şi condiţii între 1995 şi 2015. Setul de date include informaţii esenţiale pentru înţelegerea relaţiei dintre mediul interior şi percepţiile subiective ale ocupanţilor. Cele mai relevante caracteristici selectate pentru analiza şi modelarea ulterioară includ:

Caracteristici obiective ale mediului interior:

- *Air temperature (C)*: Temperatura aerului.
- *Relative humidity (%)*: Umiditatea relativă.
- *Air velocity (m/s)*: Viteza aerului.
- *Radiant temperature (C)*: Temperatura radiantă.
- *Operative temperature (C)*: Temperatura operativă.

Indici termici calculaţi:

- *PMV (Predictive Mean Vote)*: Indicele de confort termic prezis.
- *PPD (Predicted Percentage of Dissatisfied)*: Procentajul prezis de persoane nemulţumite.
- *SET (Standard Effective Temperature)*: Temperatura standard echivalentă.
- Preferinţe şi percepţii ale ocupanţilor:
- *Thermal sensation*: Senzaţia termică percepută.
- *Thermal preference*: Preferinţele ocupanţilor în ceea ce priveşte confortul termic.
- *Air movement preference*: Preferinţa pentru mişcarea aerului.

Scopul analizei este de a înţelege structura datelor, de a verifica calitatea acestora şi de a identifica factorii principali care influenţează confortul termic. Această etapă este esenţială pentru a selecta variabilele relevante ce vor fi utilizate în modelele predictive ulterioare.

2.3 Mediu de implementare

Pentru implementarea şi analiza setului de date utilizat în acest proiect, am folosit un mediu tehnic bine configurat, ce include atât biblioteci Python esenţiale pentru prelucrarea datelor şi modelare, cât şi un mediu de dezvoltare interactiv.

Am utilizat biblioteca **pandas** pentru manipularea şi prelucrarea eficientă a datelor. Aceasta mi-a permis să citesc, să filtrez şi să restructurez setul de date în funcţie de cerinţele analizei. **NumPy** a fost folosit pentru efectuarea de calcule numerice şi operaţii matriciale, fiind indispensabil în etapele de pregătire a datelor şi în implementarea algoritmilor de învăţare automată.

Vizualizările grafice au fost realizate cu ajutorul bibliotecilor **matplotlib** şi **seaborn**, care au permis crearea de grafice clare şi intuitive pentru interpretarea relaţiilor dintre variabile. Aceste vizualizări au fost fundamentale în etapa de analiză exploratorie a datelor, oferind o perspectivă vizuală asupra distribuţiilor şi corelaţiilor dintre caracteristici.

Pentru partea de modelare a datelor, am utilizat biblioteca **scikit-learn**, care a furnizat funcţionalităţile necesare pentru implementarea unor modele predictive precum regresia liniară, regresia logistică, Random Forest şi Gradient Boosting.

Codul a fost scris şi executat folosind **Python 3.13**, într-un mediu de dezvoltare **Jupyter Notebook**, care oferă un spaţiu interactiv pentru rularea secvenţială a codului şi vizualizarea imediată a rezultatelor. Sistemul de operare utilizat a fost **Windows 10**, configurat cu toate pachetele şi dependinţele necesare pentru buna desfăşurare a procesului.

Acest mediu de lucru a permis o implementare eficientă şi o analiză detaliată, asigurând condiţiile necesare pentru obţinerea rezultatelor dorite.

2.4 Citirea şi explorarea iniţială a datelor

Importul setului de date

Pentru încărcarea şi explorarea iniţială a setului de date, am utilizat biblioteca `pandas`, care permite manipularea eficientă a datelor. Codul utilizat pentru citirea fişierului CSV şi afişarea primelor rânduri este:

```
import pandas as pd
# Citirea datelor din arhiva 1
ashrae_data = pd.read_csv("C:\\Users\\asus\\Desktop\\SBCproiect\\archive1\\ashrae_db2.01.csv")
# Afişarea primelor rânduri pentru a verifica structura
print(ashrae_data.head())
```

Dimensiunea și structura datelor

Pentru a înțelege mai bine structura setului de date, am analizat numărul total de rânduri şi coloane, precum şi tipurile de date din fiecare coloană.

```
print(f"Număr de rânduri şi coloane ASHRAE: {ashrae_data.shape}")
print(ashrae_data.info())
```

Rezultatele obţinute au fost : număr total de rânduri şi coloane: 107,583 rânduri şi 70 coloane ; variabile numerice (float64, int64) cum ar fi PMV, PPD, SET, Air temperature (C), Relative humidity (%), Air velocity (m/s), şi variabile categorice (object) Season, Climate, Building type, Cooling strategy, Sex,

Verificarea calităţii datelor

Înainte de analiza detaliată a datelor, este important să verificăm existența valorilor lipsă şi a valorilor duplicate**.**

```
# Verificarea valorilor lipsă
print("Valori lipsă în setul de date:")
print(ashrae_data.isnull().sum())

# Verificarea valorilor duplicate
print(f"Numărul de duplicate în setul de date: {ashrae_data.duplicated().sum()}")
```

Rezultate:

Valori lipsă: Setul de date conţine valori lipsă în mai multe coloane, printre care:Publication (Citation): 1,655 valori lipsa ,Air velocity (m/s): 9,143 valori lipsa ,PMV: 13,780 valori lipsa ,SET: 15,000 valori lipsa ,Outdoor monthly air temperature (C): 28,245 valori lipsa , Door, Heater: peste 90,000 valori lipsa;

Valori duplicate: 5,429 rânduri duplicate au fost identificate în setul de date.

2.5 Analiza Univariată

2.5.1 Scopul analizei univariate

Analiza univariată reprezintă primul pas în explorarea setului de date și are ca scop înțelegerea distribuției fiecărei variabile luată individual. Aceasta ajută la detectarea unor posibile valori aberante, dezechilibre în distribuție și tipare care pot influența rezultatele analizei ulterioare.

Pentru această etapă, am analizat **variabilele numerice** utilizând histograme și grafice KDE, iar variabilele categorice au fost examinate prin countplot-uri pentru a observa frecvențele fiecărei categorii.

2.5.2 Distribuția variabilelor numerice

Scopul analizei distribuției variabilelor numerice este de a identifica tipare generale, cum ar fi media, variația sau posibilele valori aberante, care ar putea influența modelele predictive.

Pentru a vizualiza distribuția variabilelor numerice, s-au generat **histograme** și **grafice KDE** (Kernel Density Estimation). Aceste grafice permit observarea formei distribuției și identificarea eventualelor anomalii.

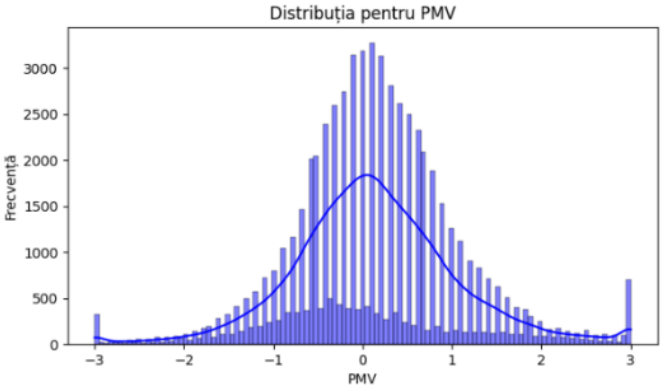

Figura 2.5.2 a) Distribuția pentru PMV

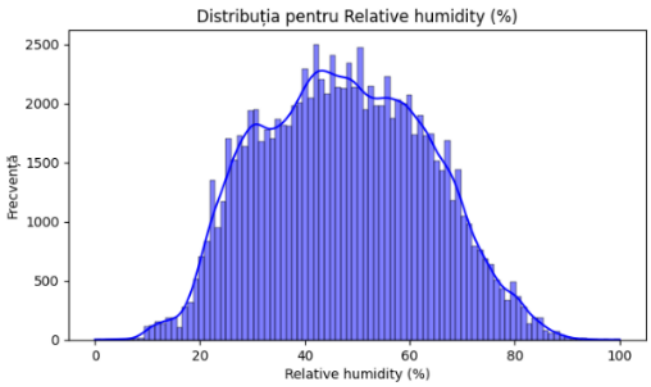

Figura 2.5.2 b) Distribuția pentru Umiditate

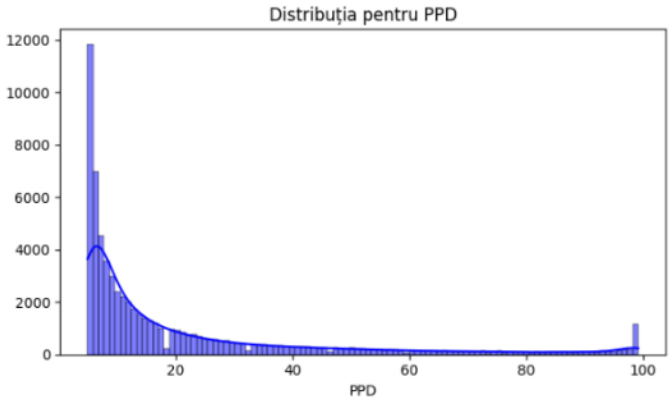

Figura 2.5.2 c). Distribuția pentru PPD (Procentul prezis de persoane nemulțumite)

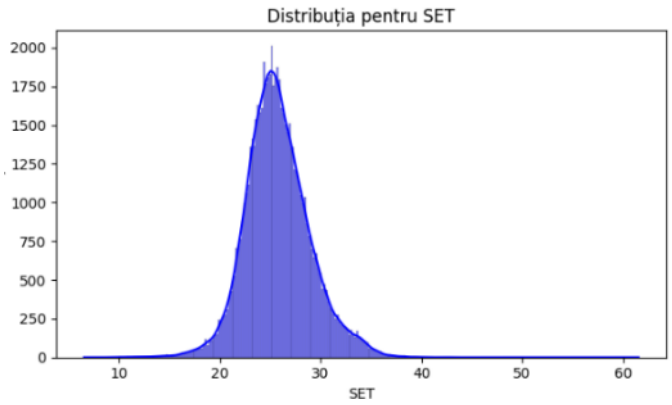

Figura 2.5.2 d). Distribuția pentru SET (Temperatura efectivă standard)

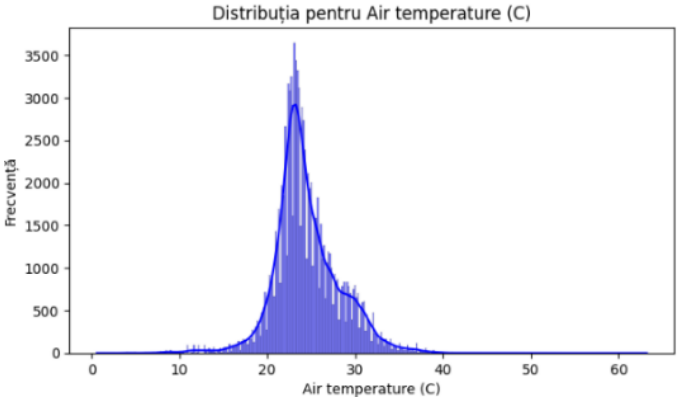

Figura 2.5.2 e). Distribuția pentru Temperatura

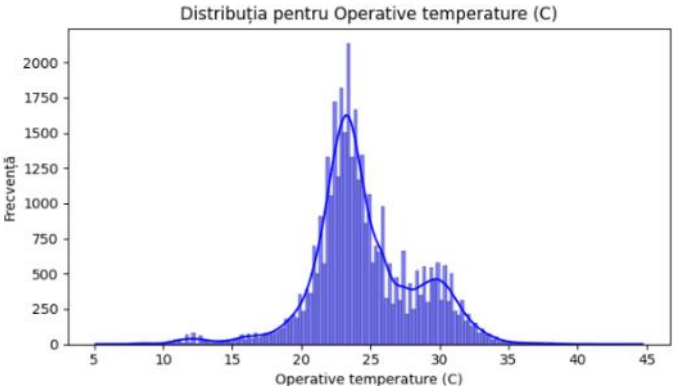

Figura 2.5.2 f). Distribuția pentru Temperatura

Se observă că distribuția variabilei PMV este aproape normală, cu cele mai multe valori în jurul intervalului de confort (-0.5 și 0.5).

Variabilele precum 'SET' și 'Air temperature (C)' prezintă o distribuție concentrată în jurul valorii medii, indicând o consistență în măsurătorile efectuate.

2.5.3 Distribuția variabilelor categorice

Analiza distribuției variabilelor categorice ajută la identificarea tiparelor și a frecvenței categoriilor distincte, oferind informații despre structura și echilibrul setului de date. Aceasta este esențială pentru a înțelege cum anumite caracteristici categorice, cum ar fi sezonul sau tipul clădirii, influențează variabilele numerice sau ținta modelului.

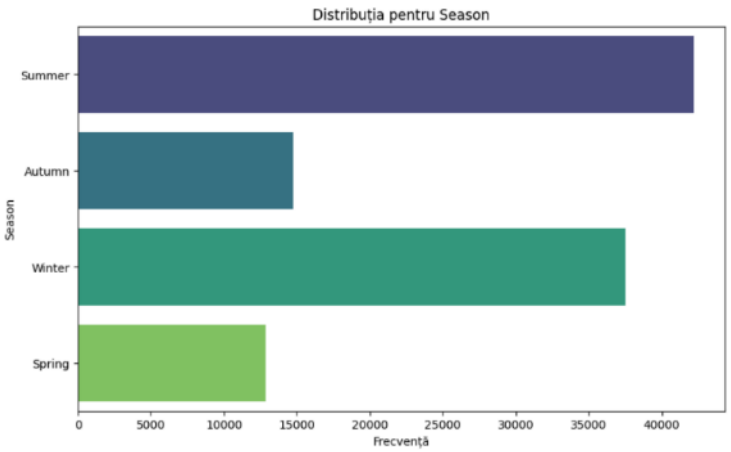

Figura 2.5.3 a). Distribuția pentru Sezon

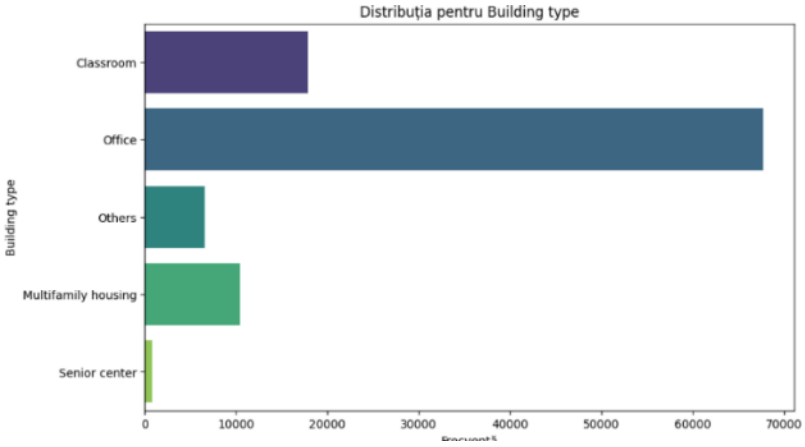

Figura 2.5.3 b). Distribuția pentru Tipul Clădirii

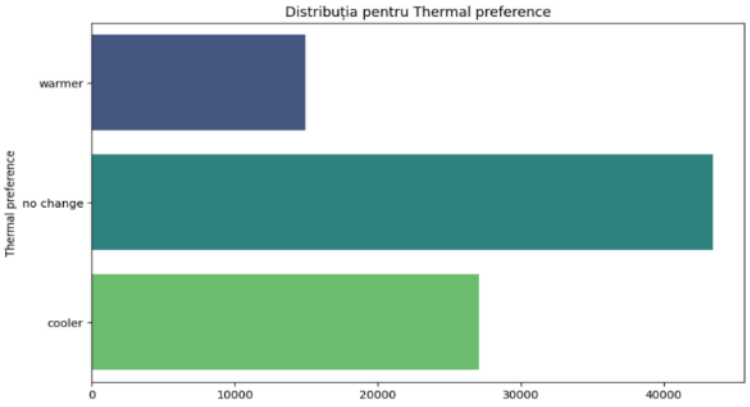

Figura 2.5.3 c). Distribuția pentru Preferința Termică

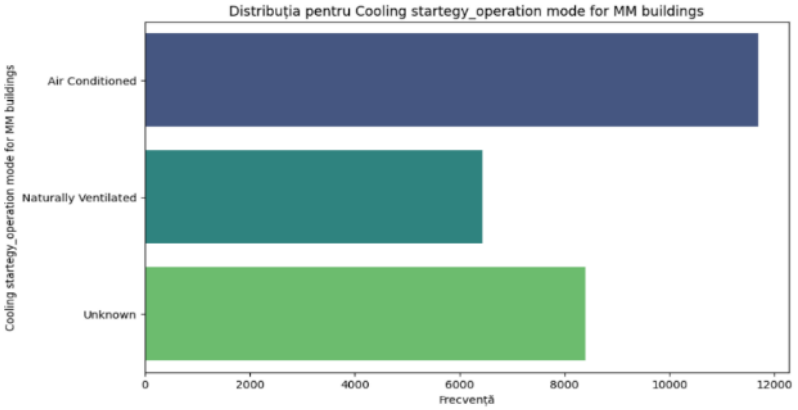

Figura 2.5.3 d). Distribuția pentru Modul de Operare al Strategiei de Răcire

Distribuția variabilelor categorice evidențiază o predominanță a datelor din sezoanele extreme (vară și iarnă) și clădirile de birouri, majoritatea participanților percepând mediul ca fiind confortabil (fără preferințe de schimbare termică). De asemenea, sistemele de aer condiționat sunt cele mai frecvent utilizate, reflectând tendința generală de optimizare a confortului termic în spații controlate.

2.6 Analiza bivariate

Scopul analizei bivariate este de a explora relațiile dintre două variabile, în special pentru a identifica corelații sau tipare care ar putea influența rezultatele modelelor predictive. Pentru aceasta, s-au generat grafice de dispersie (scatter plots) pentru a vizualiza legătura dintre variabilele de interes.

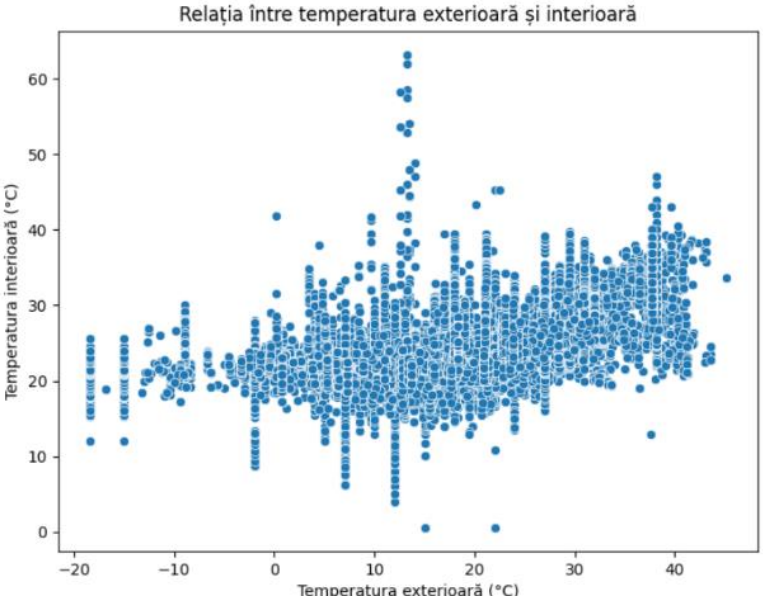

Figura 2.6 a). Relația între temperatura exterioară și temperatura interioară

Graficul arată o corelație pozitivă, ceea ce indică faptul că pe măsură ce temperatura exterioară creşte,temperatura interioară tinde să crească, dar într-o manieră controlată.

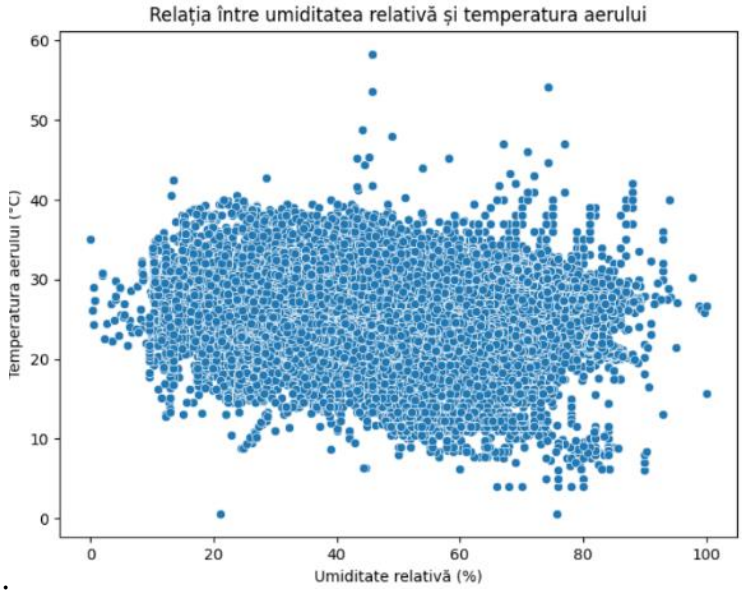

Figura 2.6 b). Relația între umiditatea relativă și temperatura aerului

Graficul nu evidențiază o corelație puternică între umiditatea relativă şi temperatura aerului. Distribuția datelor sugerează o variabilitate mare a umidității la diferite temperaturi.

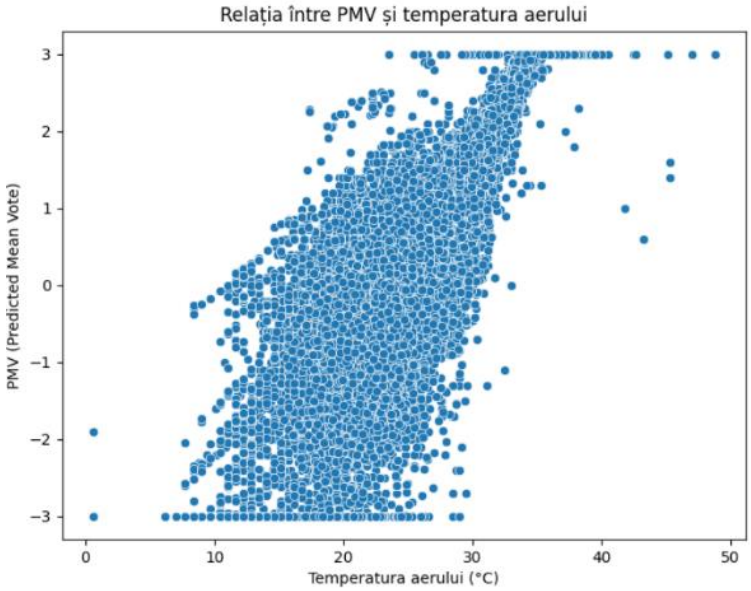

Figura 2.6 c). Relația între PMV și temperatura aerului

Graficul arată o relație clară, aproape liniară, între PMV și temperatura aerului. Pe măsură ce temperatura aerului crește, PMV devine mai mare, indicând o percepție de confort termic redus la temperaturi mai mari.

Aceste observații sunt utile pentru a înțelege cum variabilele cheie interacționează între ele, oferind o bază pentru selectarea variabilelor relevante pentru modelele predictive.

2.7 Analiza multivariată

Analiza multivariată este utilizată pentru a explora relațiile complexe dintre mai multe variabile simultan, identificând modele și structuri ascunse în date. În cadrul acestui proiect, am aplicat două tehnici principale de reducere a dimensionalității (PCA și t-SNE) și am analizat matricea de corelație extinsă pentru a înțelege mai bine interacțiunile dintre variabilele numerice.

Pentru a înțelege structura datelor și pentru a vizualiza în două dimensiuni distribuția sezonală a variabilelor, s-au aplicat următoarele metode:

2.7.1 PCA (Principal Component Analysis):

PCA este o tehnică liniară care reduce dimensionalitatea datelor, păstrând cât mai multă variație posibilă. Prima componentă principală explică cel mai mult din variația datelor, iar a doua componentă principală explică variația rămasă.

În acest proiect, PCA a fost aplicat pe un set de variabile numeric standardizate pentru a vizualiza distribuția în funcție de sezon.

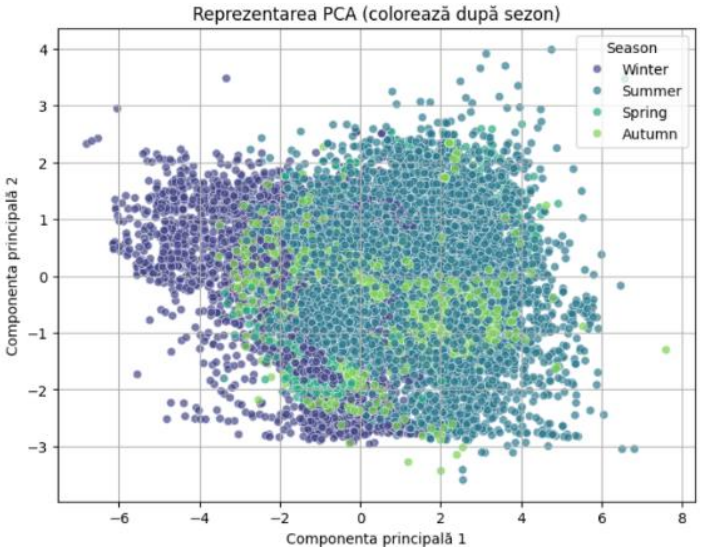

Figura 2.7.1 Reprezentarea PCA

Reprezentarea PCA arată o suprapunere a datelor pentru diferite sezoane, ceea ce sugerează că variaţia principală a datelor nu este strict influenţată de anotimpuri. Totuşi, această tehnică a permis o reducere semnificativă a dimensionalităţii datelor, păstrând structura general

2.7.2 t-SNE (t-distributed Stochastic Neighbor Embedding):

t-SNE este o tehnică de reducere a dimensionalităţii utilizată pentru vizualizarea relaţiilor complexe într-un spaţiu de dimensiuni reduse. Aceasta este mai potrivită pentru analiza clusterelor în date complexe.

Am utilizat un subset al datelor pentru a reduce timpul de calcul şi am colorat rezultatele în funcţie de anotările sezoniere.

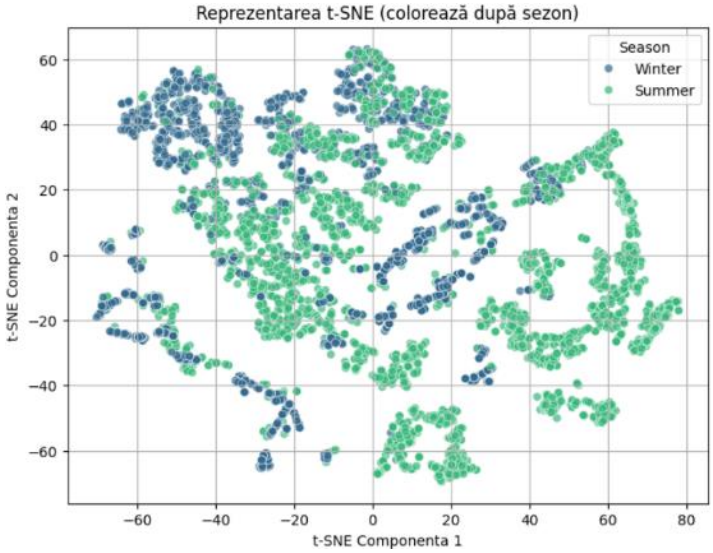

Figura 2.7.2 Reprezentarea t-SNE

Vizualizarea t-SNE evidențiază grupuri mai distincte între anumite sezoane, cum ar fi vara și iarna. Aceasta indică faptul că există variații sezoniere în datele colectate, dar și o anumită suprapunere între aceste grupuri.

2.7.3 Matricea de corelație

Pentru a evalua relațiile directe dintre variabilele numerice, am generat o matrice de corelație utilizând coeficientul Pearson. Această matrice oferă informații esențiale despre cât de puternică este legătura dintre două variabile.

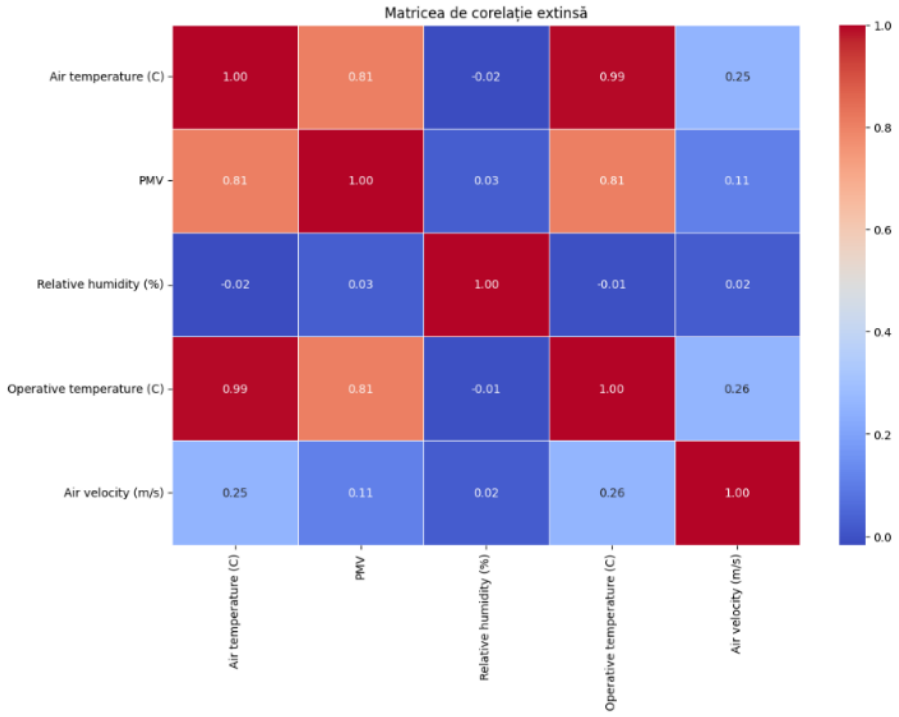

Figura 2.7.2 Matricea de corelație extinsă

Se observă o corelație foarte puternică între Air temperature (C) și Operative temperature (C) (coeficient de corelație = 0.99), ceea ce confirmă că temperaturile operative și aerului sunt aproape echivalente. În schimb, Relative humidity (%) prezintă corelații foarte slabe cu alte variabile, indicând o influență redusă asupra măsurătorilor legate de temperatură sau confort termic

Se observă o corelație foarte puternică între Air temperature (C) și Operative temperature (C) (coeficient de corelație = 0.99), ceea ce confirmă că temperaturile operative și aerului sunt aproape echivalente. În schimb, Relative humidity (%) prezintă corelații foarte slabe cu alte variabile, indicând o influență redusă asupra măsurătorilor legate de temperatură sau confort termic

3 Pre-procesarea setului de date

3.1 Introducere

Pre-procesarea datelor reprezintă o etapă esenţială în cadrul oricărui proiect de analiză sau modelare a datelor, având scopul de a asigura calitatea şi integritatea setului de date. Această etapă implică aplicarea unor metode şi tehnici pentru eliminarea zgomotului, completarea valorilor lipsă, reducerea dimensionalităţii, eliminarea valorilor aberante şi optimizarea caracteristicilor relevante.

Scopul pre-procesării este de a transforma datele brute într-un format adecvat pentru analize ulterioare şi modelare predictivă. Aceasta nu doar că îmbunătăţeşte performanţa modelelor predictive, dar şi asigură robusteţea acestora prin reducerea efectelor erorilor, tendinţelor sau valorilor extreme.

3.2 Denoising (Eliminarea zgomotului)

3.2.1 Definiţie şi scop

Denoising, sau eliminarea zgomotului, reprezintă procesul de reducere sau eliminare a variaţiilor nedorite dintr-un set de date, care pot afecta acurateţea modelelor predictive sau analizele ulterioare. Zgomotul poate apărea din cauza erorilor de măsurare, fluctuaţiilor neaşteptate sau a altor factori externi. Scopul principal al denoising-ului este de a îmbunătăţi calitatea datelor, păstrând semnalele relevante şi eliminând variaţiile inutile.

3.2.2 Metode utilizate

Eliminarea zgomotului se realizează prin aplicarea diverselor filtre, fiecare adaptat unui anumit tip de date sau zgomot. Aceste filtre sunt utile pentru a netezi valorile variabilelor numerice şi pentru a reduce incertitudinea datelor, permiţând o interpretare mai clară.

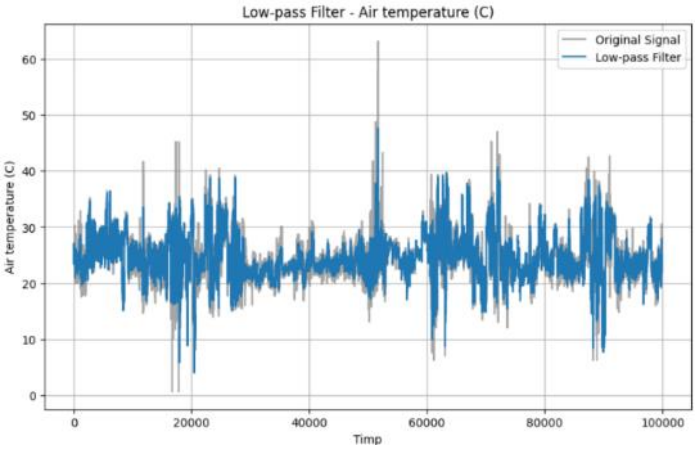

Figura 3.2.2 a). Low-pass Filter - Air temperature

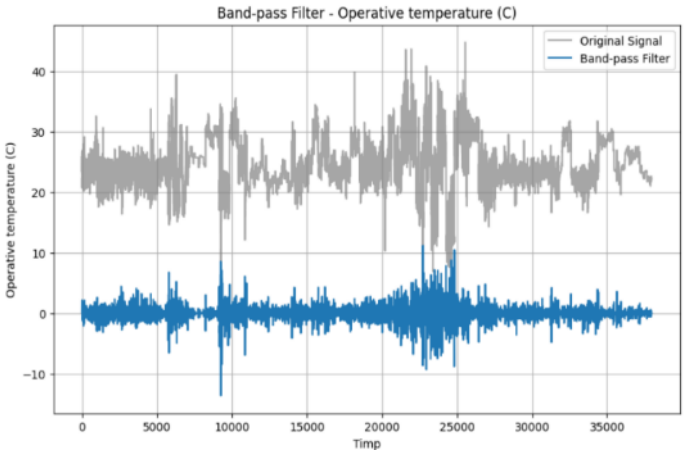

Figura 3.2.2 b). Band-pass Filter - Operative

Low-pass Filter - Air temperature (C)

Filtrul low-pass elimină variaţiile de înaltă frecvenţă pentru a evidenţia tendinţa generală a temperaturii aerului.

Semnalul filtrat arată o reducere clară a fluctuaţiilor rapide, evidenţiind variaţiile lente ale temperaturii.

Band-pass Filter - Operative temperature (C)

Filtrul band-pass păstrează doar componentele de frecvenţă dintr-un interval specific, eliminând zgomotul de joasă şi înaltă frecvenţă.

Semnalul rezultat scoate în evidenţă schimbările moderate ale temperaturii operative, eliminând zgomotul extrem.

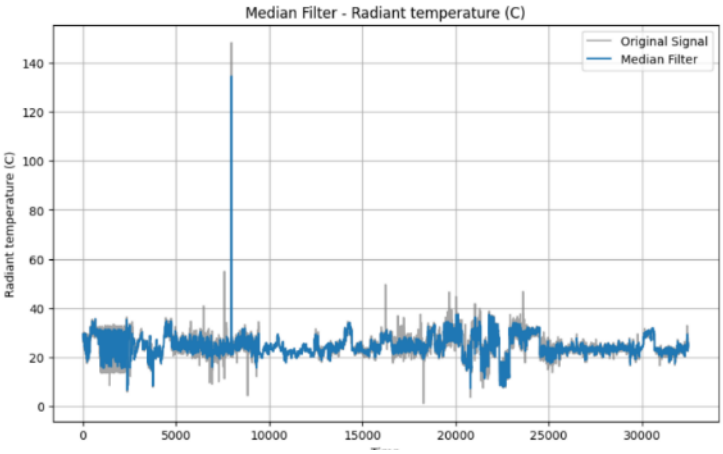

Figura 3.2.2 c). Median Filter - Radiant

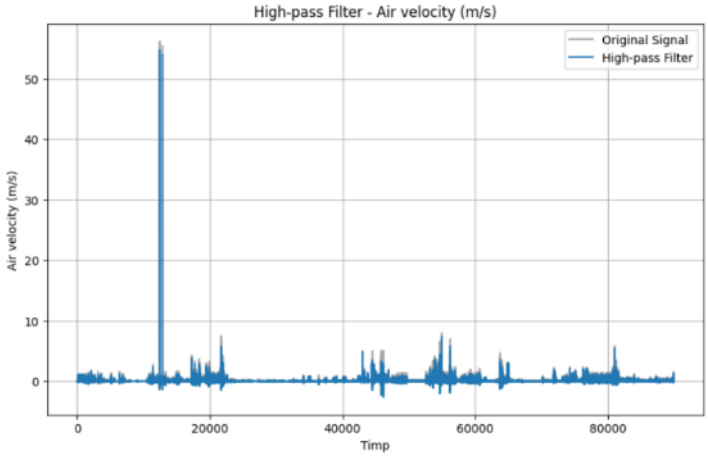

Figura 3.2.2 d). High-pass Filter - Air velocity

Median Filter - Radiant temperature (C)

Filtrul median este eficient pentru eliminarea valorilor aberante sau a zgomotului de tip "sare şi piper."

Se observă o netezire semnificativă a semnalului, în special eliminarea vârfurilor extreme.

High-pass Filter - Air velocity (m/s)

Filtrul high-pass este utilizat pentru a păstra variaţiile rapide şi a elimina tendinţele lente.

Zgomotul de frecvenţă joasă este eliminat, iar semnalul rezultat evidenţiază variaţiile rapide ale vitezei aerului.

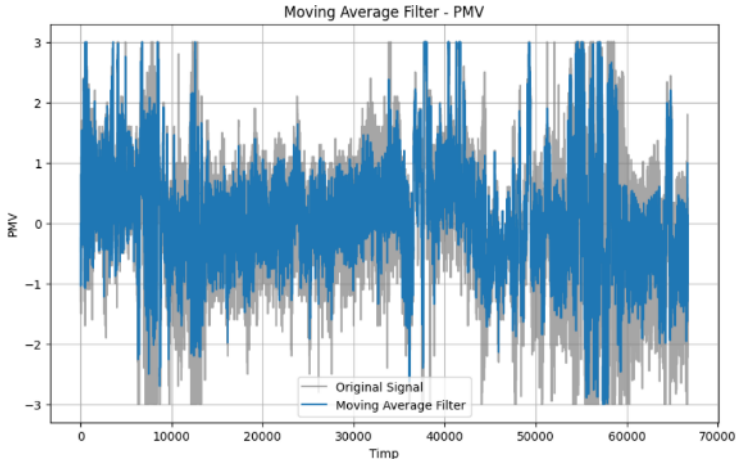

Figura 3.2.2 e). Moving Average Filter - PMV

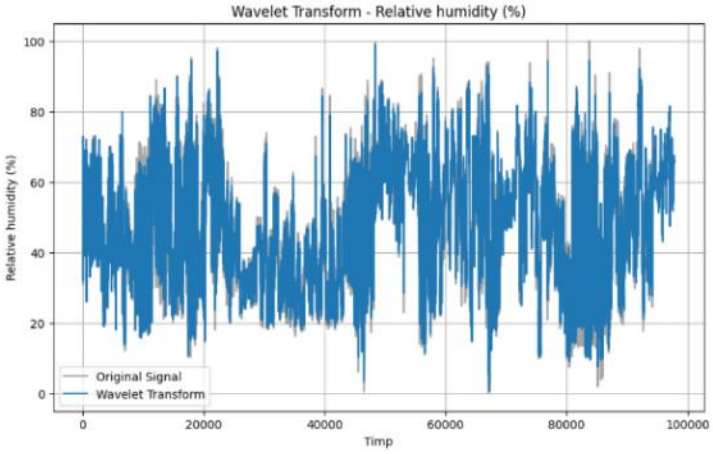

Figura 3.2.2 e). Wavelet Transform Relative humidity (%)

Moving Average Filter - PMV

Filtrul de medie mobilă netezeşte semnalul pentru a elimina fluctuaţiile mici, păstrând tendinţa generală.

Semnalul filtrat este mai neted, iar variaţiile bruşte ale valorilor PMV sunt reduse.

Wavelet Transform - Relative humidity (%)

Transformata wavelet este utilizată pentru a elimina zgomotul la mai multe scale de timp, păstrând structura semnalului.

Semnalul rezultat este netezit, iar zgomotul de înaltă frecvenţă este eliminat, evidenţiind variabilitatea umidităţii relative.

3.3 Detrending (Eliminarea tendinţelor)

3.3.1 Definiţie şi scop

Detrending-ul reprezintă procesul de eliminare a tendinţelor sau componentelor de ordin înalt dintr-un set de date, cu scopul de a analiza doar variaţiile semnalului în jurul unei valori de bază. Tendinţele pot apărea din cauza fenomenelor sezoniere, a erorilor de măsurare sau a variaţiilor de lungă durată şi pot masca informaţiile relevante din date. Prin eliminarea acestora, putem obţine un semnal mai curat, adecvat pentru analize mai detaliate.

Metodele utilizate includ substracţia mediei mobile, care elimină tendinţele de ordin inferior prin calculul unei medii pe o fereastră glisantă. Detrending-ul polinomial este folosit pentru a îndepărta componentele de trend de ordin superior, ajustând datele cu ajutorul unor polinoame de diferite grade. Detrending-ul prin diferenţiere permite extragerea variaţiilor rapide prin calculul diferenţelor dintre punctele succesive. De asemenea, detrending-ul folosind transformata Fourier ajută la identificarea şi eliminarea componentelor de frecvenţă joasă, păstrând fluctuaţiile mai rapide din date.

Air temperature (C):

Metodă utilizată: Detrending Polinomial pentru a elimina tendințele de ordin superior din datele privind temperatura aerului, a fost utilizată metoda detrending polinomial. Aceasta ajustează datele utilizând un polinom de gradul 2, eliminând fluctuațiile lente asociate cu trendurile.

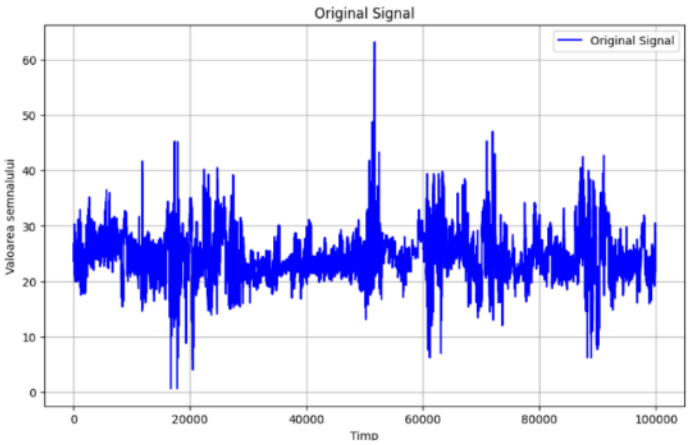

Figura 3.3 a) Semnal Original - Air temperature (C)

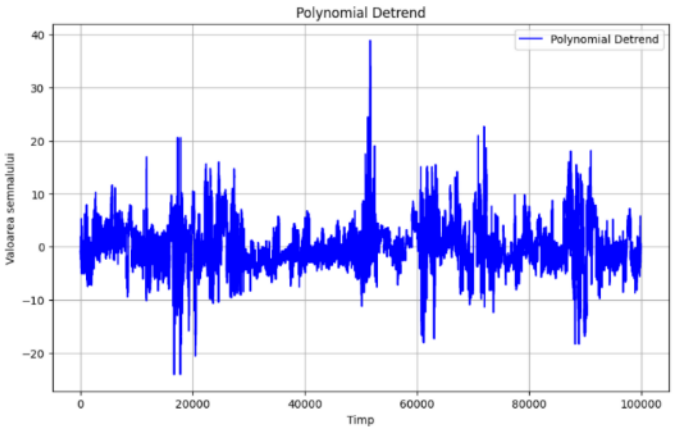

Figura 3.3 b) Detrending Polinomial - Air temperature (C)

Graficul ilustrează compararea dintre semnalul original şi semnalul rezultat după aplicarea detrending-ului polinomial. Se observă că variaţiile de ordin inferior au fost eliminate, păstrând doar fluctuaţiile rapide şi relevante.

Relative humidity (%):

Metodă utilizată: Substracţia Mediei Mobile pentru datele referitoare la umiditatea relativă, a fost aplicată metoda substracţiei mediei mobile. Aceasta elimină tendinţele de ordin inferior prin calcularea unei medii pe o fereastră glisantă şi scăderea acesteia din semnalul original.

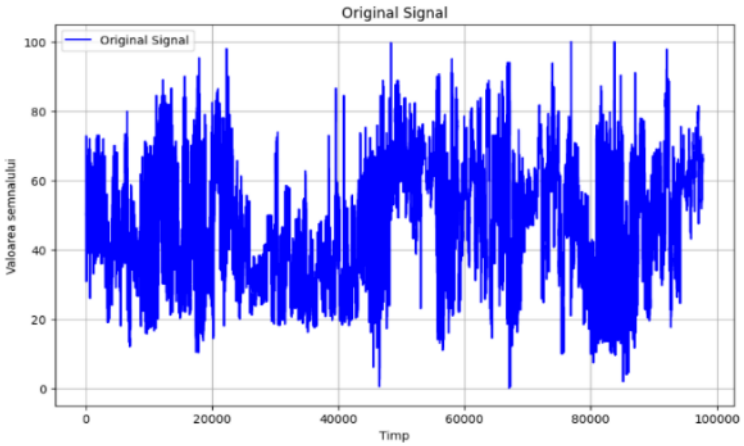

Figura 3.3 c) Semnal Original - Relative humidity (%)

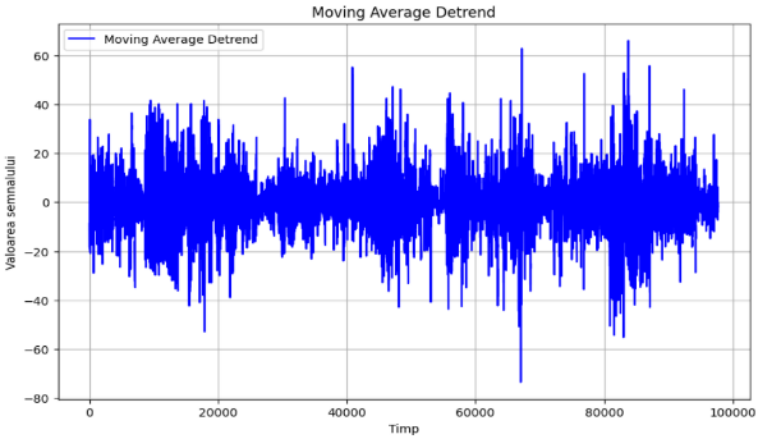

Figura 3.3 d) Detrending prin Substracția Mediei Mobile - Relative

Graficul arată cum metoda substracției mediei mobile a redus componentele lente ale semnalului, păstrând variațiile rapide. Compararea cu semnalul original evidențiază eliminarea tendințelor pe termen lung.

Operative temperature (C):

Metodă utilizată: Transformata Fourier pentru temperatura operativă, a fost aplicată metoda detrending utilizând transformata Fourier. Aceasta a identificat și eliminat componentele de frecvență joasă, lăsând doar variațiile mai rapide și relevante pentru analiză.

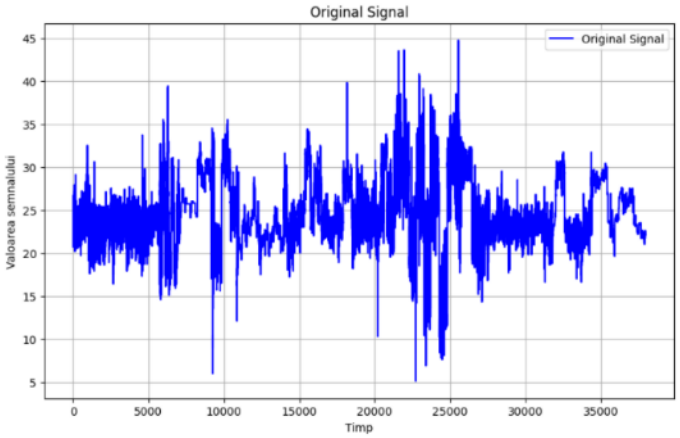

Figura 3.3 e) Semnal Original - Operative temperature (C)

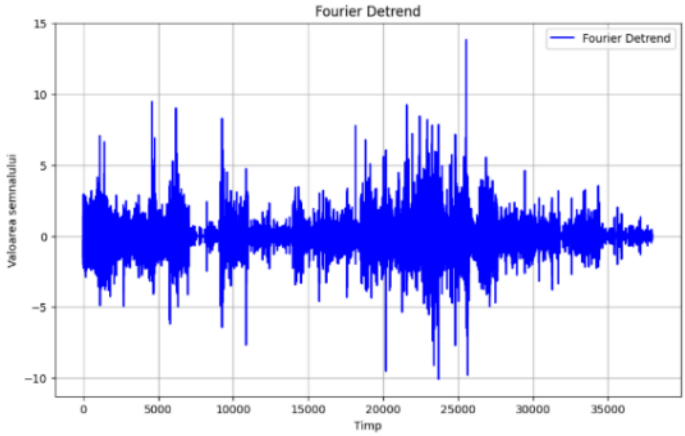

Figura 3.3 f) Detrending prin Transformata Fourier - Operative

Graficul afișează semnalul prelucrat după aplicarea metodei Fourier comparat cu semnalul original. Se observă că metoda a eliminat variațiile lente, păstrând detaliile rapide ale semnalului.

3.4 Eliminarea valorilor aberante (Outlier Removal)

3.4.1 Definiţie şi scop

Eliminarea valorilor aberante reprezintă un pas esenţial în pre-procesarea datelor, având ca scop identificarea şi eliminarea punctelor de date care diferă semnificativ de restul setului. Aceste valori pot apărea din cauza erorilor de măsurare, a introducerii incorecte a datelor sau a altor procese anormale, şi pot afecta negativ calitatea modelelor predictive şi analiza statistică.

Scopul eliminării valorilor aberante este de a îmbunătăţi acurateţea şi robusteţea modelelor, reducând influenţa extremelor asupra rezultatelor analizei. Aceasta se realizează utilizând diverse metode statistice şi algoritmice, fiecare fiind potrivită pentru tipuri specifice de seturi de date şi structuri.

Urmează să analizăm şi să aplicăm mai multe metode utilizate pentru identificarea şi eliminarea valorilor aberante: Z-Score Method, IQR Method (Interquartile Range,MAD (Median Absolute Deviation ,DBSCAN Method ,Isolation Forest Method.

Pentru fiecare metodă, vom analiza implementarea şi efectele asupra unui subset de date selectate, incluzând vizualizări grafice şi interpretări ale rezultatelor.

3.4.2 Metoda Z-Score (Z-Score Method)

Metoda Z-Score este o tehnică utilizată pentru identificarea valorilor aberante (outliers) prin calcularea distanţei fiecărui punct de date faţă de media dataset-ului, exprimată în termeni de deviaţii standard. Aceasta este potrivită pentru seturi de date care urmează o distribuţie normală şi permite detectarea valorilor extreme care se abat semnificativ de la majoritatea datelor..

Cum funcţionează:

Se calculează Z-Score-ul pentru fiecare punct de date utilizând formula**:**

$$Z = \frac{x-\mu}{\sigma} \tag{3.4.2}$$

Unde:

x :Valoarea punctului de date.

μ: Media dataset-ului.

σ: Deviaţia standard a dataset-ului.

Un punct de date este considerat aberant dacă valoarea absolută a scorului Z depăşeşte un prag prestabilit (de exemplu, $Z > 3$ sau $Z < -3$).

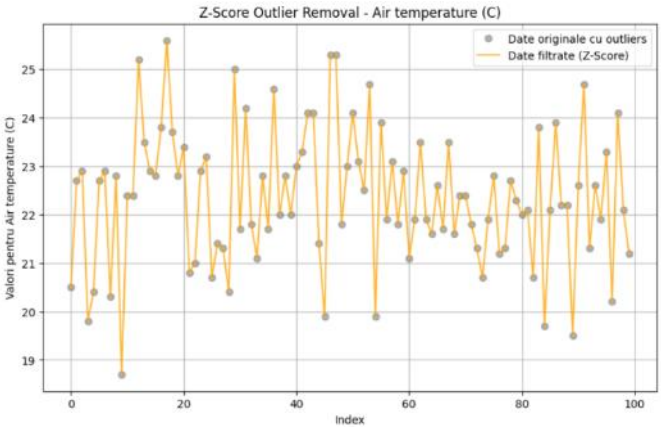

Figura 3.4.2 a) Eliminarea valorilor aberante folosind metoda Z-Score - Temperatura aerului (°C)

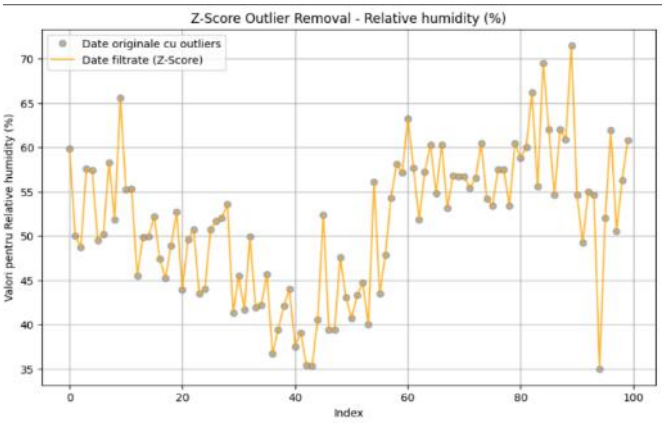

Figura 3.4.2 b) Eliminarea valorilor aberante folosind metoda Z-Score - Umiditatea relativă (%)

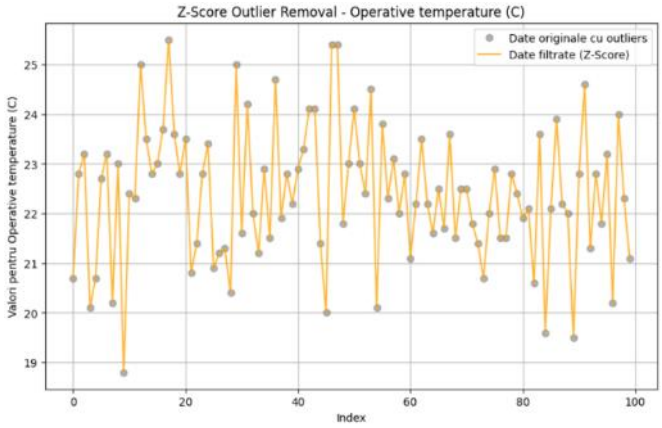

Figura 3.4.2 c) Eliminarea valorilor aberante folosind metoda Z-Score - Temperatura operativă (°C)

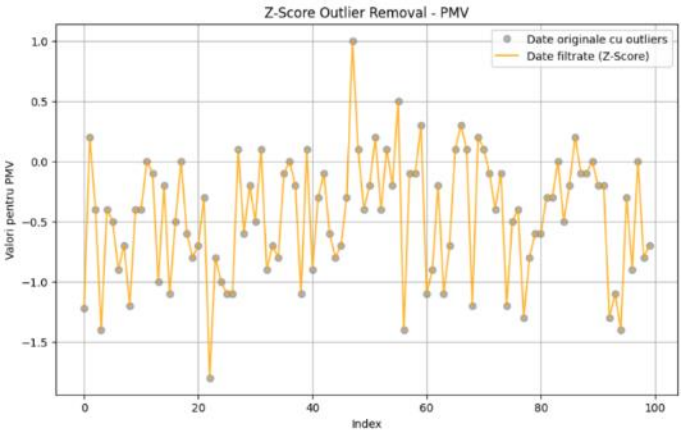

Figura 3.4.2 d) Eliminarea valorilor aberante folosind metoda Z-Score - PMV (Predicted Mean Vote)

Metoda Z-Score a fost aplicată pe cele patru coloane selectate, iar graficele evidențiază clar eliminarea valorilor aberante. Datele originale conțineau puncte care deviau semnificativ de la medie, în timp ce datele filtrate au eliminat aceste valori, rezultând un set de date mai curat și mai reprezentativ. Această abordare este ideală pentru date distribuite normal, iar rezultatele au demonstrat eficiența metodei în acest context.

3.4.3 Metoda IQR (Interquartile Range)

Metoda IQR (Interval Intercuartil) este utilizată pentru identificarea și eliminarea valorilor aberante dintr-un set de date. Aceasta se bazează pe calcularea intervalului dintre percentila 25 (Q1) și percentila 75 (Q3), iar valorile care depășesc acest interval extins cu un factor multiplicativ (1.5 * IQR, de obicei) sunt considerate valori aberante.

Cum funcţionează:

Se calculează percentila 25 (Q1) și percentila 75 (Q3).

Intervalul intercuartil (IQR) se determină astfel: $IQR = Q3 - Q1$ (3.4.3.1)

Limitele pentru detectarea valorilor aberante sunt:

$Lower\ Bound = Q1 - 1.5 * IQR$ (3.4.3.2)

$Upper\ Bound = Q1 + 1.5 * IQR$ (3.4.3.3)

Orice punct de date care se află sub limita inferioară sau peste limita superioară este considerat o valoare aberantă.

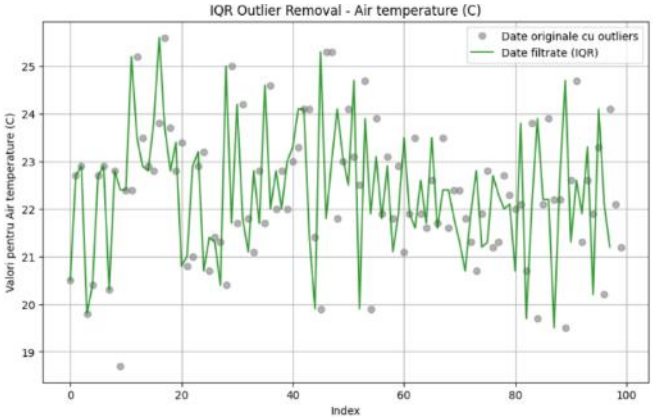

Figura 3.4.3 a) IQR Outlier Removal - Air Temperature (C)

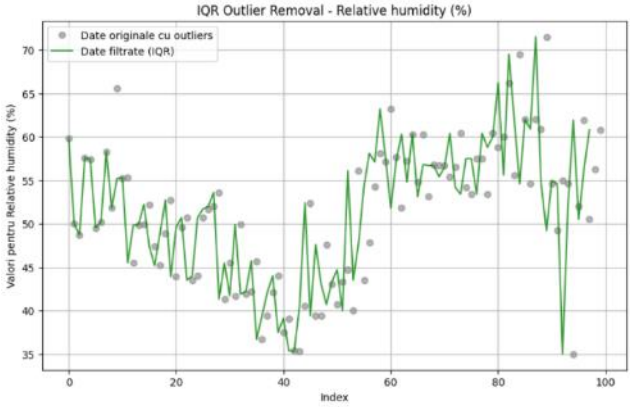

Figura 3.4.3 b) IQR Outlier Removal - Relative Humidity (%)

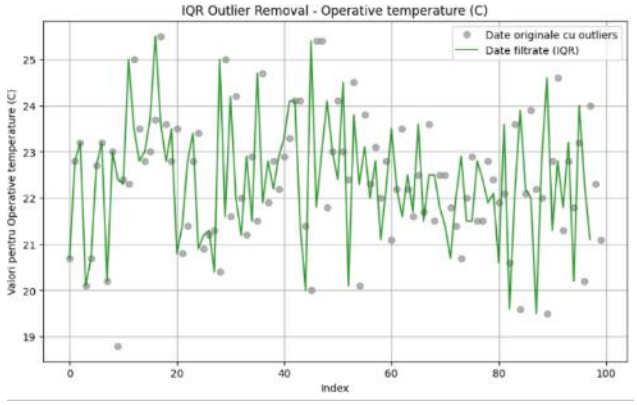

Figura 3.4.3 c) IQR Outlier Removal - Operative Temperature (C)

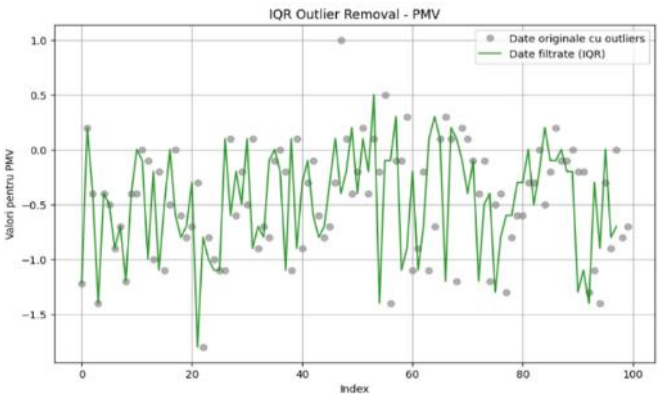

Figura 3.4.3 d) IQR Outlier Removal - PMV

Metoda IQR (Interquartile Range) a fost aplicată pentru cele patru coloane analizate, cu scopul de a elimina valorile aberante care se află în afara intervalului [Q1 - 1.5 * IQR, Q3 + 1.5 * IQR]. Graficele arată clar diferenţele dintre datele originale şi cele filtrate, evidenţiind eliminarea punctelor aberante. Această metodă este deosebit de utilă pentru datele cu distribuţii nesimetrice sau date cu extreme multiple, oferind un set de date mai curat şi mai robust pentru analizele ulterioare.

3.4.4 MAD (Median Absolute Deviation)

Metoda MAD (Median Absolute Deviation) este utilizată pentru identificarea valorilor aberante într-un mod robust, fiind mai puţin sensibilă la extreme faţă de metodele bazate pe media aritmetică. Aceasta este ideală pentru seturi de date care prezintă distribuţii asimetrice sau cu cozi lungi.

Cum Funcţionează:

Se calculează mediana absolută a deviaţiilor faţă de mediana datelor:

$$MAD = median(\ |x_i - M|) \qquad (3.4.4.1)$$

unde M este mediana setului de date.

Se defineşte o valoare de prag kkk (de obicei 1.5 sau 3) pentru detectarea valorilor aberante. Un punct este considerat aberant dacă se află în afara intervalului:

$$[\ M - k * MAD, M + k * MAD] \qquad (3.4.4.2)$$

Această metodă este robustă deoarece medianele sunt mai puţin afectate de valorile extreme, comparativ cu metodele care folosesc media.

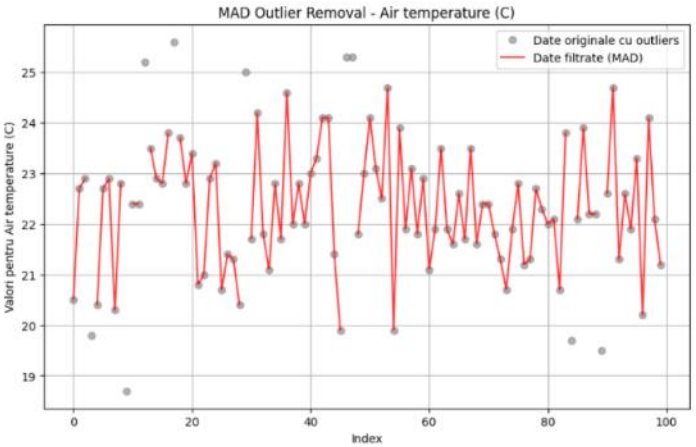

Figura 3.4.4 a) MAD Outlier Removal - Air temperature (C)

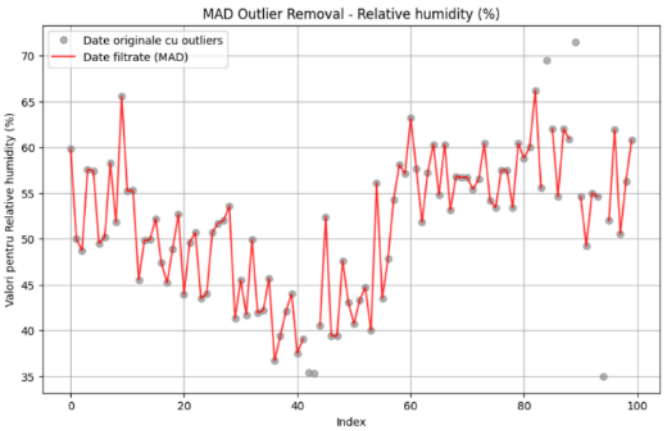

Figura 3.4.4 b) MAD Outlier Removal - Relative humidity (%)

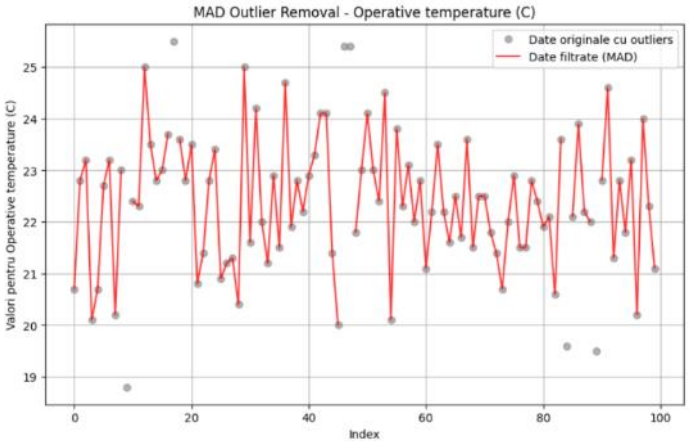

Figura 3.4.4 c) MAD Outlier Removal - Operative temperature (C)

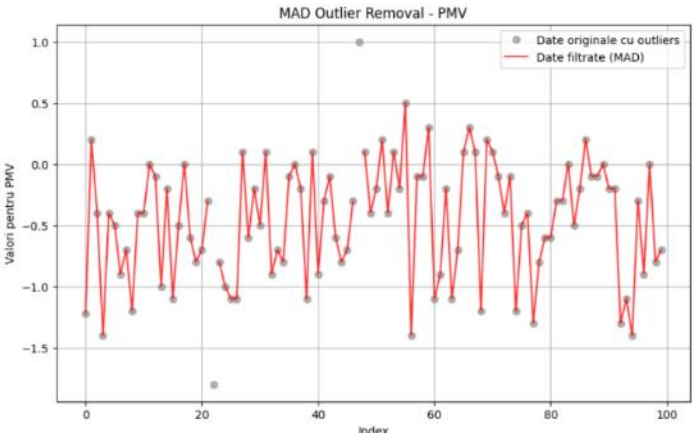

Figura 3.4.4 d) MAD Outlier Removal - PMV

Metoda MAD (Median Absolute Deviation) a fost aplicată pentru cele patru coloane selectate, demonstrând eficienţa sa în identificarea şi eliminarea valorilor aberante.

Această metodă este robustă împotriva valorilor extreme şi funcţionează bine chiar şi în cazul distribuţiilor asimetrice. În graficele prezentate, se observă eliminarea outlier-ilor, rezultând un set de date mai curat, potrivit pentru analize ulterioare.

3.4.5 DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) este o metodă de clasificare bazată pe densitate care identifică grupurile de puncte dense în date şi consideră punctele din regiunile de densitate scăzută ca fiind valori aberante (outliers). Această metodă este utilă pentru datele care prezintă structuri spaţiale sau clustere naturale.

Cum funcţionează:

Puncte de bază (core points): Punctele care au un număr minim de vecini într-un anumit radius definit de parametrul epsilon.

Vecini direcţi: Punctele care sunt în vecinătatea imediată a unui punct de bază.

Outliers (noise points): Punctele care nu aparţin niciunui cluster din cauza densităţii scăzute.

Parametri principali:

epsilon (raza maximă): Defineşte distanţa maximă dintre două puncte pentru a fi considerate vecine.

minPts (numărul minim de puncte): Numărul minim de puncte necesare pentru a forma un cluster.

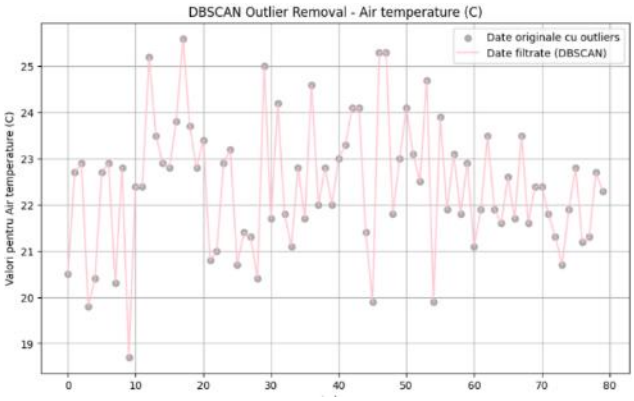

Figura 3.4.5 aDBSCAN Outlier Removal - Air temperature (C)

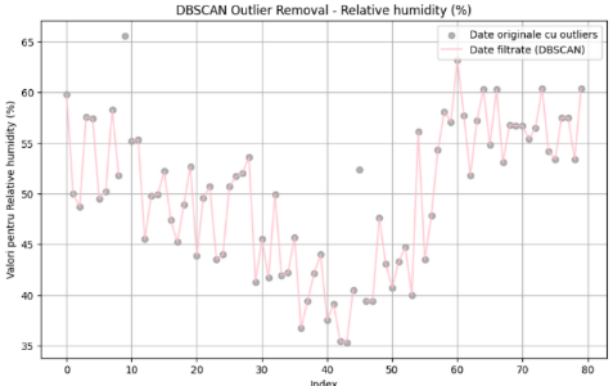

Figura 3.4.5 b) DBSCAN Outlier Removal - Relative humidity (%)

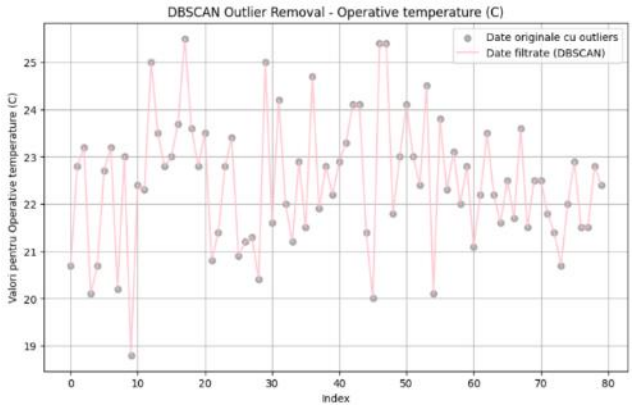

Figura 3.4.5 c) DBSCAN Outlier Removal - Operative temperature (C)

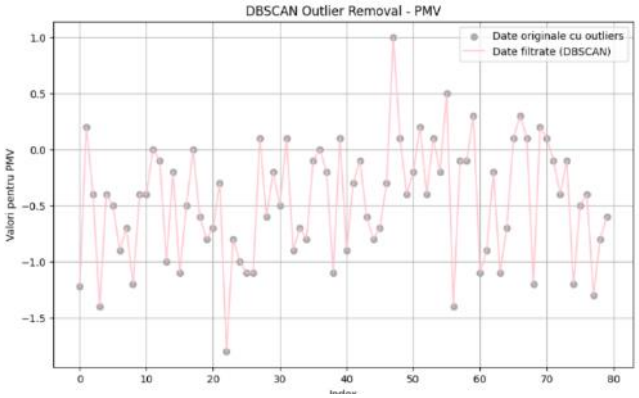

Figura 3.4.5 d) DBSCAN Outlier Removal - PMV

Metoda **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)** a fost aplicată pentru identificarea şi eliminarea valorilor aberante din cele patru coloane analizate. Graficele evidenţiază cum această tehnică a identificat şi filtrat valorile care se află în regiuni cu densitate scăzută, fiind considerate puncte aberante. Această metodă este deosebit de utilă pentru datele cu grupări naturale şi distribuţii neuniforme, oferind o abordare adaptivă pentru detecţia valorilor aberante în funcţie de parametrii densităţii (eps) şi numărul minim de puncte (min_samples). Rezultatele arată un set de date mai curat, potrivit pentru analize ulterioare

3.4.6 Isolation Forest Method

Isolation Forest Method este o tehnică bazată pe învăţarea ansamblului (ensemble learning) utilizată pentru detectarea valorilor aberante în seturi de date complexe. Această metodă izolează în mod iterativ punctele de date prin divizarea spaţiului de date, utilizând arbori de decizie construiţi aleatoriu.

Cum funcţionează

Izolare prin partajare: Datele sunt partajate în mod repetat utilizând arbori de decizie, iar punctele care sunt izolate rapid (apar în ramuri mai îndepărtate) sunt considerate valori aberante.

Evaluarea scorului de anomalie: Fiecare punct primeşte un scor bazat pe cât de rapid este izolat în structura arborilor. Punctele cu un scor ridicat sunt marcate ca valori aberante.

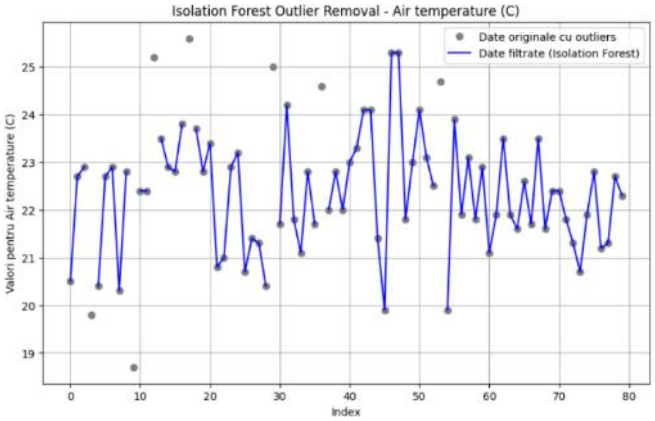

Figura 3.4.6 a) Isolation Forest Outlier Removal - Air temperature (C)

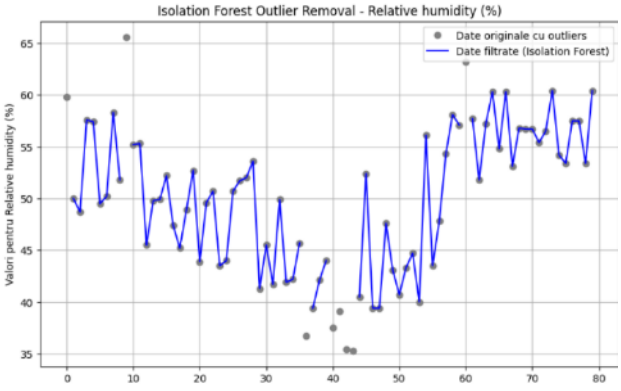

Figura 3.4.6 a) Isolation Forest Outlier Removal - Air temperature (C)

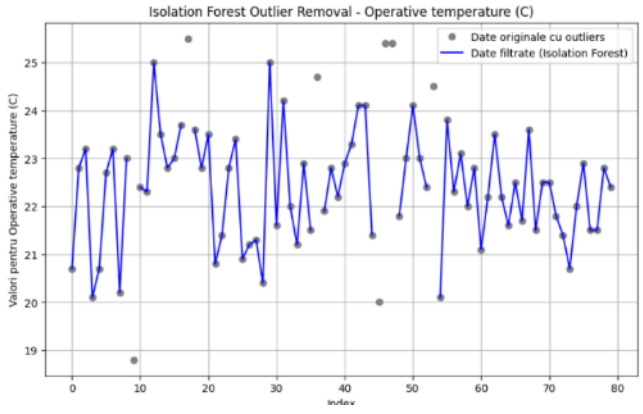

Figura 3.4.6 a) Isolation Forest Outlier Removal - Air temperature (C)

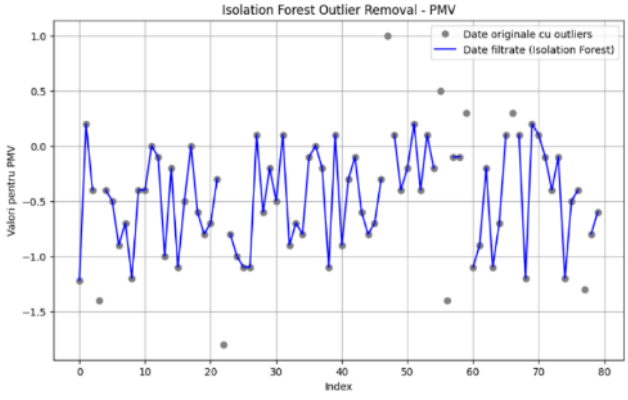

Figura 3.4.6 a) Isolation Forest Outlier Removal - Air temperature (C)

Graficele evidențiază diferențele între datele originale, care conțineau puncte izolate identificate drept outlieri, și datele filtrate, care prezintă un set de valori mai uniform. Această metodă, fiind bazată pe un algoritm de tip ensemble ce utilizează păduri de decizie, este eficientă pentru seturi de date multidimensionale și poate identifica cu acuratețe punctele aberante. Rezultatele demonstrează capacitatea acestei tehnici de a îmbunătăți calitatea datelor pentru analize ulterioare.

3.5 Interpolare

3.5.1 Definiție și importanță

Interpolarea reprezintă procesul de estimare a valorilor lipsă într-un set de date, asigurând continuitatea și prevenind impactul negativ al datelor lipsă asupra analizei sau modelării.

Datele lipsă pot perturba analiza, mai ales în serii temporale sau seturi de date spațiale. Interpolarea completează aceste lacune, facilitând o analiză mai fluidă și prevenind introducerea de erori sau biasuri.

3.5.2 Interpolare liniară

Această metodă conectează două puncte cunoscute cu o linie dreaptă, estimând valorile lipsă de-a lungul acestei linii.

Formulă utilizată: $$y = y_1 + \frac{x - x_1}{x_2 - x_1} \times (y_2 - y_1) \quad (3.5.2)$$

Unde:

(x_1, y_1) și (x_2, y_2) sunt punctele cunoscute

x este poziția valorii lipsă.

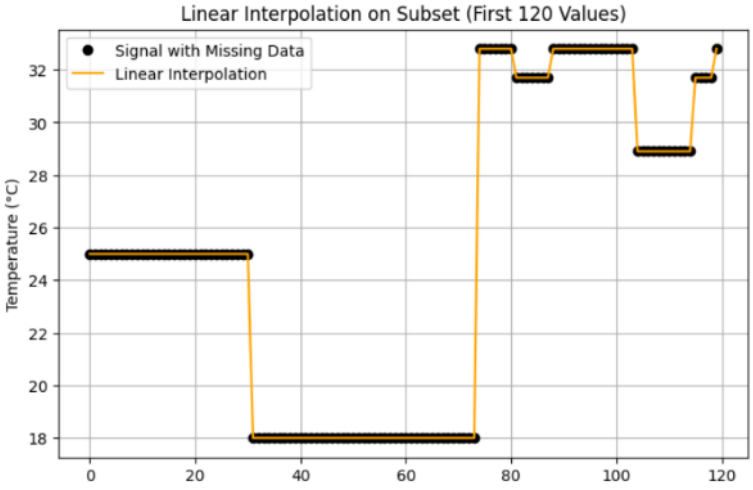

Figura 3.5.2 Linear Interpolation on Subset (First 120 Values)

Graficul arată tranziţii directe şi netede între punctele lipsă. Este simplu şi eficient pentru serii temporale cu variaţii reduse

3.5.3 Interpolare cubică

Interpolarea cubică ajustează o funcţie polinomială cubică între punctele cunoscute, creând o curbă continuă.

Formulă utilizată:

$$S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \quad (3.5.3)$$

Unde coeficienţii (a_i, b_i, c_i, d_i) asigură continuitatea derivatelor şi curburii.

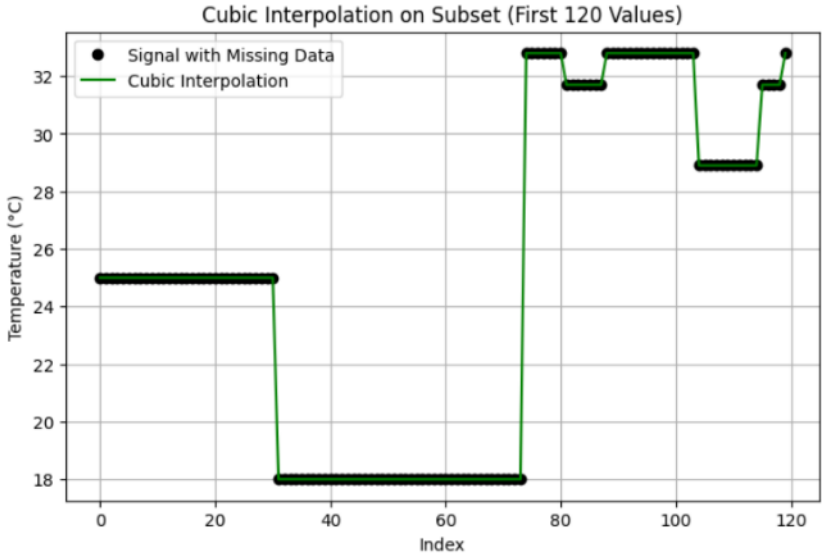

Figura 3.5.3 Cubic Interpolation on Subset (First 120 Values)

Rezultatele sunt similare interpolării liniare datorită variaţiilor reduse din setul de date.

3.5.4 LOESS (Locally Estimated Scatterplot Smoothing)

LOESS aplică o regresie locală utilizând punctele dintr-un interval apropiat. Este flexibilă pentru date cu modele non-liniare.

Formulă utilizată: $w_i = e^{-\frac{d(x,x_i)^2}{2h^2}}$ (3.4.5)

Unde h este parametrul care controlează lăţimea vecinătăţii.

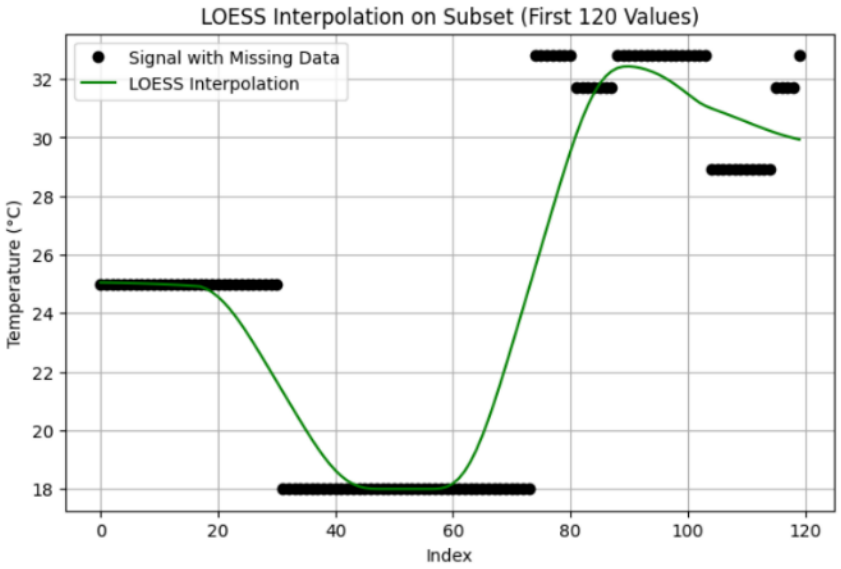

Figura 3.5.4 LOESS Interpolation on Subset (First 120 Values)

LOESS oferă un rezultat neted şi flexibil, adaptându-se la variaţiile locale ale datelor.

Analiza graficelor rezultate pentru metodele de interpolare evidențiază următoarele aspecte:

Interpolarea liniară şi **interpolarea cubică** oferă rezultate similare în cazul datelor utilizate. Acest lucru se datorează caracterului constant sau variaţiilor reduse ale datelor originale. În astfel de situaţii, curbele generate de aceste metode nu aduc diferenţe semnificative, ceea ce face ca utilizarea ambelor să fie redundantă.

LOESS, în schimb, se remarcă prin flexibilitatea sa în adaptarea la variaţiile locale şi prin capacitatea de a genera o curbă mai netedă şi mai realistă. Această metodă este mai potrivită pentru datele analizate, deoarece evidenţiază diferenţele subtile şi oferă o interpolare mai naturală.

Astfel, pentru acest set de date cu variaţii reduse şi valori constante în anumite intervale, **LOESS** reprezintă cea mai relevantă metodă de interpolare, fiind capabilă să captureze subtilităţile datelor, spre deosebire de metodele mai simple.

4 Modelarea sistemului

4.1 Introducere

Scopul acestei secţiuni este de a dezvolta modele predictive care să reliefeze relaţia dintre variabilele climatice şi PMV (Predicted Mean Vote), o măsură esenţială în evaluarea confortului termic. În cadrul proiectului, au fost utilizate diverse tehnici de regresie şi modele bazate pe arbori de decizie, având ca obiectiv principal identificarea celui mai performant model din punct de vedere al preciziei şi capacităţii de *generalizare.*

4.1.1 Importanţa modelării corecte în contextul aplicaţiei HVAC

Modelarea precisă a confortului termic este crucială pentru aplicaţiile HVAC (Heating, Ventilation, and Air Conditioning), deoarece acestea trebuie să asigure condiţii optime de confort termic în timp ce optimizează consumul de energie.

Un model precis permite sistemelor HVAC să anticipeze necesităţile utilizatorilor şi să ia decizii informate pentru controlul variabilelor climatice, cum ar fi temperatura, umiditatea sau fluxul de aer. Acest lucru nu doar că îmbunătăţeşte satisfacţia utilizatorilor, dar contribuie şi la sustenabilitatea energetică a clădirilor.

4.1.2 Utilitatea datelor preprocesate pentru îmbunătăţirea performanţei modelului

Datele preprocesate reprezintă o bază solidă pentru obţinerea unor modele performante. Eliminarea outlierilor, interpolarea valorilor lipsă şi normalizarea caracteristicilor au contribuit la crearea unui set de date consistent şi uniform, reducând riscul de erori şi asigurând condiţii ideale pentru antrenarea modelelor.

În acest fel, modelele predictive dezvoltate au fost capabile să capteze mai bine relaţiile dintre variabilele climatice şi confortul termic, maximizând acurateţea predicţiilor.

Această secţiune prezintă metodologia, tehnicile aplicate, şi rezultatele obţinute pentru fiecare model, oferind o bază solidă pentru alegerea soluţiei optime.

4.2 Metodologie și tehnici aplicate

4.2.1 Regresie liniară

Regresia liniară este o tehnică de modelare statistică utilizată pentru a estima relația dintre o variabilă y și una sau mai multe variabile independente x.

Modelul presupune că relația este liniară și poate fi reprezentată prin ecuația:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \qquad (4.4.2)$$

unde:

- y: variabila dependentă (PMV în acest caz),
- β_0: interceptul (valoarea yyy când toate variabilele independente sunt 0),
- $\beta_1, \beta_2 \ldots \beta_n$,: coeficienții care indică ponderea fiecărei variabile independente,
- $x_1, x_2 \ldots x_n$: variabilele independente.

Scopul regresiei este să găsească valorile coeficienților care minimizează eroarea pătratică medie (MSE) între valorile observate și cele prezise.

4.2.1.1 Rezultate:

- **MSE (Mean Squared Error):** 0.0554
- **R^2 Score:** 0.9391

4.2.1.2 Interpretarea rezultatelor

Un **R^2 Score** de 0.939 sugerează că modelul explică aproximativ 93.9% din variația datelor, ceea ce indică o performanță bună.

Valoarea relativ mică a **MSE (0.0554)** arată că modelul are erori de predicție reduse.

Deși modelul oferă o predicție precisă, este important să fie comparat cu alte metode pentru a verifica dacă există soluții mai performante.

4.2.2 Regresie logistică

Regresia logistică este o metodă utilizată pentru clasificare binară, unde variabila dependentă y poate avea două clase (ex. confortabil/disconfortabil). Modelul prezice probabilitatea ca un punct de date să aparțină unei clase folosind funcția logistică:

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}} \qquad (4.4.2)$$

Unde: $P(y = 1)$reprezintă probabilitatea ca observația să aparțină clasei 1. Predicțiile sunt transformate în clase (0 sau 1) utilizând un prag (de obicei 0.5).

4.2.2.1 Rezultate:

- **Acurateţe:** 99.01%
- **Matricea de clasificare:**
- Clasa 0: Precision = 1.00, Recall = 0.98, F1-Score = 0.99
- Clasa 1: Precision = 0.98, Recall = 1.00, F1-Score = 0.99
- **F1-Score mediu ponderat:** 0.99

4.2.2.2 Interpretarea rezultatelor

Acurateţea de 99.01% arată că modelul clasifică foarte bine datele.

Precision şi Recall sunt aproape 1.00, indicând un echilibru bun între clase şi o clasificare precisă a confortului termic.

Regresia logistică este un model eficient pentru clasificarea binară a PMV în clase de confort termic. Performanţa excelentă sugerează că este o soluţie solidă pentru probleme de clasificare.

4.2.3 Regresii Ridge şi Lasso

4.2.3.1 Ridge Regression

Utilizează regularizarea L2, adăugând o penalizare la suma pătratelor coeficienţilor în funcţia de cost:

$$Cost = \sum(y_i - \hat{y}_i)^2 + \lambda \sum(\beta_j)^2 \qquad (4.2.3.1)$$

Aceasta ajută la prevenirea suprapotrivirii prin reducerea magnitudinii coeficienţilor, păstrând toate caracteristicile.

4.2.3.2 Lasso Regression

Utilizează regularizarea L1, adăugând o penalizare la suma valorilor absolute ale coeficienţilor:

$$Cost = \sum(y_i - \hat{y}_i)^2 + \lambda \sum|\beta_j| \qquad (4.2.3.2)$$

Această metodă poate elimina coeficienţi, fiind utilă pentru selecţia caracteristicilor.

4.2.3.3 Rezultate

Ridge Regression:

- MSE: 0.0554
- R^2 Score: 0.9391

Lasso Regression:

- MSE: 0.0563
- R^2 Score: 0.9381

4.2.3.4 Interpretarea rezultatelor

Ridge Regression: Performanţe similare cu regresia liniară, dar mai stabile datorită regularizării L2.

Lasso Regression: Uşor mai slabă decât Ridge (MSE puţin mai mare, R^2 mai mic), dar a fost utilă pentru identificarea caracteristicilor mai puţin relevante.

Ridge Regression este preferabilă pentru modele generale stabile, în timp ce Lasso este mai potrivită când selecţia automată a caracteristicilor este necesară.

4.2.4 Random Forest

Random Forest este o tehnică de învăţare automată bazată pe arbori de decizie, utilizând un ansamblu de arbori pentru a îmbunătăţi precizia modelului şi a reduce overfitting-ul.

Metoda principală din spatele Random Forest este Bagging (Bootstrap Aggregating), care presupune:generarea de subeşantioane aleatorii din datele de antrenament,construirea arborilor de decizie individuali pe aceste subeşantioane,agregarea predicţiilor arborilor prin:media predicţiilor pentru regresie,votul majoritar pentru clasificare.

4.2.4.2 Rezultate

- MSE (Mean Squared Error): 0.0041
- R^2 Score: 0.9955

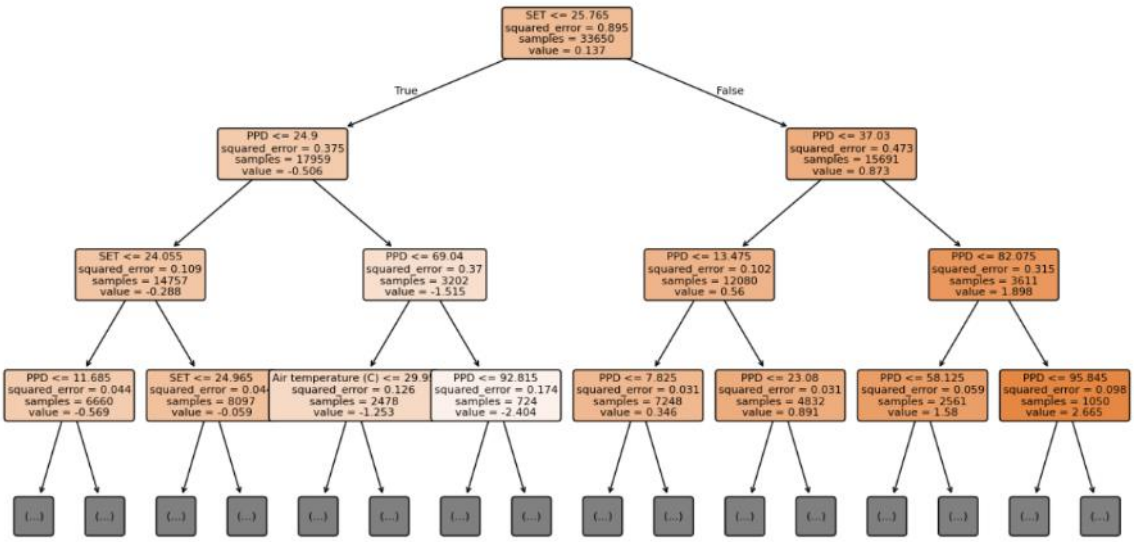

Figura 4.2.4 a)Vizualizare arobore nr.0 (max_depth=3)

Analiza importanței caracteristicilor:

- Air temperature (C): 0.008638
- Relative humidity (%): 0.000691
- Air velocity (m/s): 0.001731
- PPD: 0.365071
- SET: 0.623870

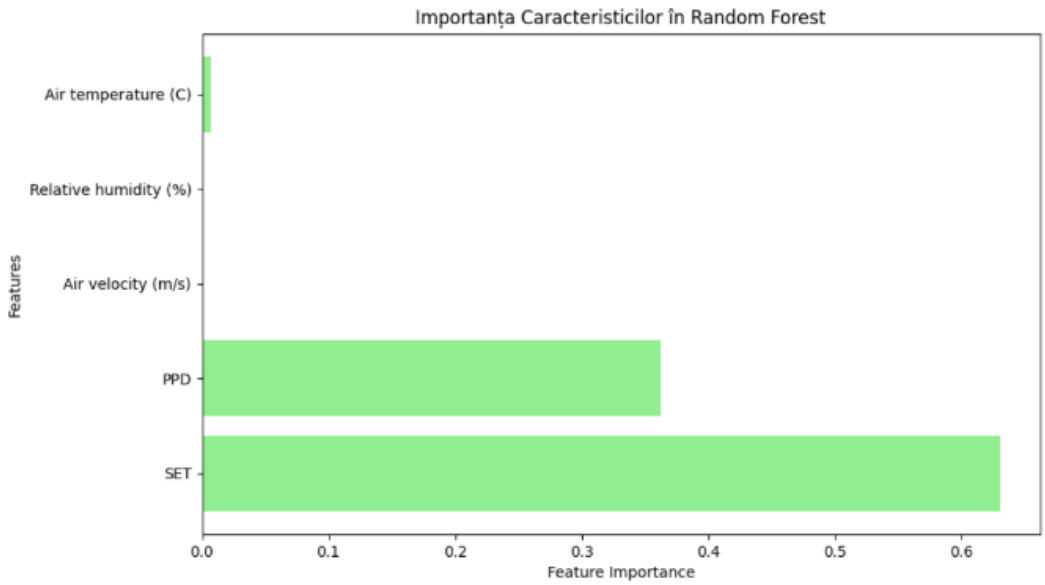

Figura 4.2.4 b) Importanța Caracteristicilor în Random Forest

4.2.4.2 Interpretarea rezultatelor

Performanță ridicată: Cu un **MSE de 0.0041** și un **R^2 Score de 0.9955**, Random Forest este unul dintre cele mai precise modele utilizate.

Feature Importance: Caracteristica SET este dominantă, contribuind cu peste 62% la predicțiile modelului, urmată de PPD cu 36.51%. Restul caracteristicilor (temperatura aerului, umiditatea relativă și viteza aerului) au importanțe semnificativ mai mici, ceea ce sugerează că modelul se bazează în principal pe factori integrați pentru a prezice PMV.

4.2.5 XGBoost

XGBoost (Extreme Gradient Boosting) este o variantă îmbunătățită a algoritmului Gradient Boosting, optimizată pentru performanță și eficiență. **Gradient Boosting** construiește arbori succesivi, fiecare încercând să corecteze erorile arborelui anterior, minimizând o funcție de cost prin gradient descent.

Avantaje ale XGBoost:

- Include regularizare L1 și L2 pentru prevenirea overfitting-ului.
- Construirea arborilor este optimizată pentru a utiliza eficient resursele hardware.
- Utilizează tehnici precum subsampling și shrinkage (înmulțirea predicțiilor arborilor cu un factor niu) pentru a îmbunătăți generalizarea.

4.2.5.1 Rezultate

- **MSE (Mean Squared Error):** 0.0039
- **R^2 Score:** 0.995
-

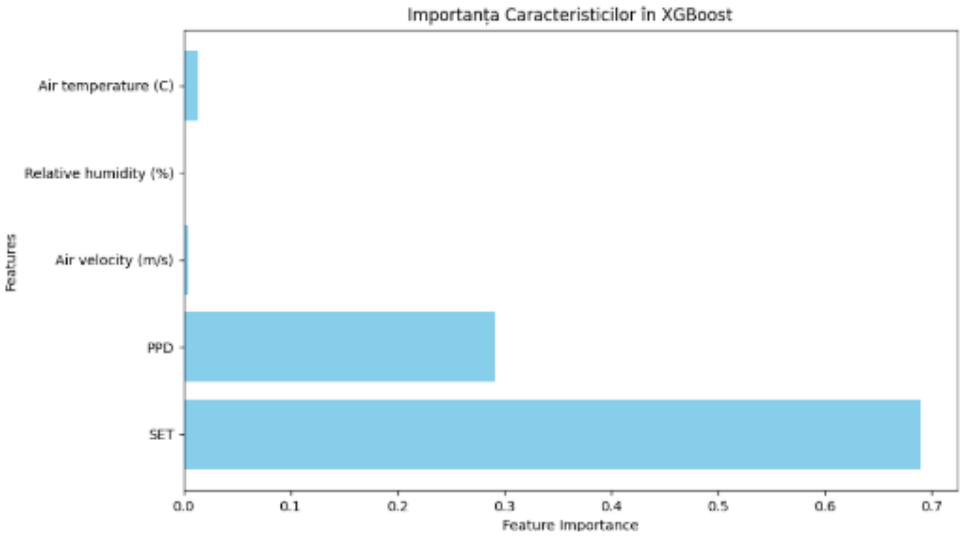

Figura 4.2.5 Importanta Caracteristicilor în XGBoost

Analiza importanţei caracteristicilor:

- Air temperature (C): 0.013489
- Relative humidity (%): 0.000389
- Air velocity (m/s): 0.004639
- PPD: 0.291973
- SET: 0.689510

4.2.5.2 Interpretarea rezultatelor

Performanţă excelentă: Cu un **MSE de 0.0039** şi un **R^2 Score de 0.9956**, XGBoost depăşeşte uşor Random Forest în precizie.

Feature Importance: Similar cu Random Forest, SET este cea mai importantă caracteristică, contribuind cu aproape 69% la predicţii, urmată de PPD (29.20%). Celelalte caracteristici au importanţe neglijabile, confirmând relevanţa dominantă a factorilor integraţi.

4.3 Implementarea

4.3.1 Mediul de implementare

Pentru acest proiect, implementarea s-a realizat utilizând Python , iar codul complet al proiectului este salvat într-un fişier separat, care conţine toate etapele necesare: de la preprocesarea datelor, la construirea şi evaluarea modelelor. Acest fişier poate fi accesat şi rulat pe orice sistem compatibil cu Python, asigurând reproducerea completă a lucrării.

4.3.2 Cum se poate replica lucrarea

Pentru a replica rezultatele acestui proiect, urmaţi paşii următori:

Pregătirea mediului:

Asiguraţi-vă că aveţi instalate toate bibliotecile necesare:

- *scikit-learn*: pentru regresie şi evaluarea modelelor.
- *xgboost*: pentru implementarea Gradient Boosting.
- *pandas*: pentru manipularea şi analiza datelor.
- *numpy*: pentru operaţiuni numerice.
- *matplotlib*: pentru vizualizarea rezultatelor.

Fişierul principal al proiectului:

Toate etapele proiectului, inclusiv codul pentru modele, sunt salvate într-un fişier Python

Acest fişier include: preprocesarea datelor,divizarea datelor în seturi de antrenament şi testare,construirea şi antrenarea modelelor,evaluarea performanţei fiecărui model,vizualizarea rezultatelor şi analiza importanţei caracteristicilor.

4.4 Evaluarea modelelor

4.4.1 Metode şi metrici de evaluare

Evaluarea performanţei modelelor utilizate în acest proiect s-a bazat pe metrici specifice pentru regresie şi clasificare, fiecare fiind alese pentru a măsura acurateţea şi eficienţa modelelor în diferite aspecte.

Pentru modelele de **regresie**, s-au utilizat următoarele metrici:

MSE (Mean Squared Error): Această metrică măsoară media pătratelor diferenţelor dintre valorile reale şi cele prezise de model. Un MSE mic indică o eroare de predicţie redusă şi, implicit, un model mai precis.

R^2 Score (Coeficientul de determinare): Această metrică arată proporţia din variaţia totală a datelor care este explicată de model. Valori apropiate de 1 indică un model care explică bine datele, în timp ce valori apropiate de 0 sugerează că modelul nu capturează suficient de bine relaţiile dintre variabile.

Pentru **clasificare binară**, s-au utilizat metrici precum:

Acurateţea (Accuracy): Proporţia predicţiilor corecte în totalul datelor testate. Este o metrică globală, dar poate fi insuficientă în cazul unui dezechilibru între clase.

Precizia (Precision): Măsoară procentul de predicţii corecte dintre toate predicţiile pozitive făcute de model.

Recall-ul: Măsoară capacitatea modelului de a identifica toate cazurile pozitive din setul de date.

F1-score: Combinaţia dintre precizie şi recall, fiind o medie armonică a celor două. Este utilă în special când există un dezechilibru între clase, oferind o imagine mai completă a performanţei modelului.

În acest proiect, **MSE** şi **R^2 Score** au fost folosite pentru a evalua modelele de regresie, în timp ce **F1-score**, acurateţea, precizia şi recall-ul au fost utilizate pentru evaluarea clasificării binare.

4.4.2 Analiza rezultatelor

Performanţa fiecărui model utilizat a fost comparată şi analizată în detaliu, utilizând metricile menţionate mai sus. Tabelul de mai jos sintetizează rezultatele obţinute pentru fiecare model:

Tabel 4.4.2 Performanţa modelelor utilizate pentru analiza PMV

Model	MSE	R^2 Score / Acurateţe
Linear Regression	0.0554	0.9391
Logistic Regression	N/A	0.9901 (Acurateţe)
Ridge Regression	0.0554	0.9391
Lasso Regression	0.0563	0.9381
Random Forest Regressor	0.0041	0.9955
XGBoost Regressor	0.0039	0.9956

4.4.3 Interpretarea rezultatelor

Modelele **XGBoost** şi **Random Forest** au obţinut cele mai bune rezultate pentru regresie, cu valori extrem de mici ale MSE (0.0039 şi 0.0041, respectiv) şi R^2 Score foarte apropiate de 1 (0.9956 şi 0.9955). Aceste performanţe superioare se datorează capacităţii acestor algoritmi de a captura relaţiile non-liniare dintre variabile şi de a gestiona interacţiuni complexe între caracteristici.

Modelele liniare, cum ar fi **Linear Regression**, **Ridge** şi **Lasso**, au obţinut un R^2 Score de aproximativ 0.939, ceea ce arată că acestea pot explica ~93.9% din variaţia datelor. Totuşi, aceste rezultate sunt mai slabe comparativ cu modelele bazate pe arbori de decizie, ceea ce sugerează că modelele liniare nu sunt la fel de eficiente în capturarea relaţiilor complexe.

Pentru clasificare, **Logistic Regression** a demonstrat o performanţă excelentă, cu o acurateţe de 99.01% şi un F1-score aproape perfect (0.99). Acest lucru arată că modelul poate clasifica eficient confortul termic (PMV_binary) în funcţie de variabilele de intrare.

4.4.4 Concluzii privind performanţa modelelor

Modele bazate pe arbori de decizie: XGBoost şi Random Forest s-au dovedit a fi cele mai performante modele, având capacitatea de a gestiona relaţii complexe şi de a minimiza erorile de predicţie. De asemenea, analiza importanţei caracteristicilor a confirmat că variabilele compuse (SET şi PPD) au fost dominante, contribuind semnificativ la predicţii.

Modele liniare: Deşi performanţele lor au fost bune, acestea nu au reuşit să concureze cu modelele avansate din punct de vedere al preciziei. Totuşi, aceste modele sunt mai simple şi mai rapide, fiind utile în scenarii unde interpretabilitatea este esenţială.

Clasificare logistică: Acest model a demonstrat o capacitate remarcabilă de a clasifica datele, fiind un instrument eficient pentru problemele de clasificare binară.

Această analiză arată că alegerea unui model depinde de natura problemei și de complexitatea datelor, iar modelele bazate pe arbori oferă cel mai bun echilibru între precizie și capacitatea de generalizare.

4.5 Concluzii

În concluzie, analiza performanței fiecărui model a evidențiat diferențe semnificative între metodele aplicate. Modelele liniare, precum regresia liniară, Ridge și Lasso, au oferit rezultate decente, cu un R^2 de aproximativ 93.9%, dar s-au dovedit mai puțin eficiente în capturarea relațiilor complexe dintre variabile. Pe de altă parte, Random Forest și XGBoost au excelat, cu scoruri R^2 de 0.9955 și 0.9956 și MSE-uri foarte mici, indicând o capacitate ridicată de predicție.

Modelul **XGBoost** a fost selectat ca model final datorită performanței superioare, reflectată de cel mai mic MSE și cel mai mare R^2 Score. Analiza importanței caracteristicilor a confirmat relevanța măsurilor integrate precum **SET** (62.39%) și **PPD** (36.51%), subliniind contribuția majoră a acestor variabile la predicția confortului termic.

Pentru îmbunătățirea viitoare a modelului, ar putea fi utile tehnici suplimentare de optimizare a hiperparametrilor, cum ar fi utilizarea căutării Bayesiene sau a rețelelor de căutare, pentru a ajusta performanța XGBoost și Random Forest. De asemenea, explorarea altor modele avansate, precum rețelele neuronale, ar putea aduce beneficii suplimentare în cazul unor seturi de date mai complexe.

5 Concluzii

5.1 Rezultate obținute

Proiectul a avut ca scop modelarea și analiza confortului termic utilizând date climatice și diverse tehnici de modelare, pentru a identifica cea mai potrivită metodologie.

Am parcurs etape clare, începând cu preprocesarea datelor, continuând cu aplicarea diferitelor modele și evaluarea rezultatelor, și încheind cu interpretarea performanțelor fiecărui model.

În etapa de preprocesare, am eliminat datele necorespunzătoare prin tehnici precum **Isolation Forest**, iar valorile lipsă au fost completate folosind metode diverse de interpolare, inclusiv LOESS, care s-a dovedit cea mai adaptată naturii datelor noastre. Acest proces a îmbunătățit calitatea setului de date, reducând impactul erorilor asupra performanței modelelor.

Pentru etapa de modelare, am aplicat mai multe metode, inclusiv **regresia liniară, Ridge și Lasso**, precum și modele avansate bazate pe arbori de decizie, cum ar fi **Random Forest** și **XGBoost**.

Rezultatele au demonstrat că XGBoost a avut cea mai bună performanță, cu un **MSE de 0.0039** și un **R^2 Score de 0.9956**, depășind semnificativ modelele liniare în capacitatea sa de a captura relațiile non-liniare din date.

Contribuțiile proiectului includ:

- Implementarea unui flux complet de preprocesare, modelare și evaluare, adaptat datelor climatice.
- Selectarea și validarea modelelor avansate pentru predicția PMV, subliniind relevanța caracteristicilor integrate precum SET și PPD.
- Oferirea unei metodologii detaliate și reproductibile, care poate fi aplicată în alte studii sau proiecte HVAC.

Comparativ cu alte lucrări din domeniu, proiectul nostru se remarcă prin integrarea mai multor modele, evaluarea detaliată a performanței și o atenție sporită acordată preprocesării datelor, aspect care a contribuit semnificativ la acuratețea rezultatelor.

5.2 Direcții de dezvoltare

Pentru a continua dezvoltarea proiectului, există mai multe direcții posibile:

Optimizarea hiperparametrilor: Tehnici precum căutarea Bayesiană sau optimizarea cu rețele neuronale pot îmbunătăți performanțele modelelor XGBoost și Random Forest.

Extinderea setului de date: Adăugarea mai multor caracteristici, precum date geografice sau istorice, ar putea îmbunătăți predicțiile și ar permite o analiză mai detaliată.

Explorarea modelelor avansate: Implementarea unor arhitecturi de rețele neuronale, cum ar fi rețelele neuronale convoluționale (CNN) sau rețelele recurente (RNN), ar putea deschide noi oportunități pentru analiza datelor climatice.

Implementarea în timp real: Integrarea unui sistem în timp real pe o platformă hardware, precum Raspberry Pi, pentru a monitoriza și controla confortul termic într-un mediu HVAC.

Validare pe seturi externe de date: Testarea modelelor pe alte seturi de date climatice, din diferite locații geografice, pentru a verifica robustețea soluțiilor propuse.

Proiectul oferă o bază solidă pentru analiza și predicția confortului termic, dar extinderea cercetării și aplicarea metodelor în medii practice ar putea aduce beneficii semnificative în domeniul sistemelor HVAC inteligente.

6 Bibliografie

1] **V. Moroșanu**, "Cercetări privind realizarea de sisteme integrate pentru clădiri cu consum redus de energie," Teză de doctorat, Universitatea Tehnică "Gheorghe Asachi" din Iași, 2024. [Online].

Disponibil:https://doctorat.tuiasi.ro/wp-content/uploads/2024/09/Rezumat-Morosanu-Vlad.pdf

2] **A. M. Măgurean**, "Analiza performanței energetice a clădirilor nerezidențiale prin tehnici de modelare numerică și Inteligență Artificială aplicată," *Teză de doctorat*, Universitatea Tehnică din Cluj-Napoca, 2017. [Online].
Disponibil:https://www.researchgate.net/publication/387504020_Analiza_performantei_energetice_a_cladirilor_nerezidentiale_prin_tehnici_de_modelare_numerica_si_Inteligenta_Artificiala_aplicata_Teza_de_doctorat

3] **IEEE Citation Reference**, 2009. [Online].

Disponibil:https://journals.ieeeauthorcenter.ieee.org/wp-content/uploads/sites/7/IEEE_Reference_Guide.pdf

[4] **IEEE Editorial Style Manual**, 2016. [Online].

Disponibil:https://moodle.hs-augsburg.de/pluginfile.php/283344/mod_resource/content/2/IEEE-Editorial-Style_Manual.pdf

6.1 Specificaţii pentru bibliografie

Bibliografia proiectului a fost selectată pentru a include surse relevante din domeniul modelării sistemelor HVAC, al eficienţei energetice şi al metodologiilor utilizate în analiza şi optimizarea performanţei acestora. Referinţele alese acoperă atât cercetări academice de actualitate, cât şi standarde utilizate pentru redactarea documentaţiei tehnice şi ştiinţifice.

[1] V. Moroşanu, "Cercetări privind realizarea de sisteme integrate pentru clădiri cu consum redus de energie," Teză de doctorat, Universitatea Tehnică "Gheorghe Asachi" din Iaşi, 2024. [Online]. Disponibil:
Acest document oferă o perspectivă detaliată asupra strategiilor de integrare a sistemelor HVAC în clădiri cu consum redus de energie, subliniind metodele moderne de optimizare energetică. Informaţiile incluse sunt utile pentru înţelegerea modului în care tehnologiile avansate pot contribui la eficienţa termică a clădirilor şi reducerea consumului de resurse.

[2] A. M. Măgurean, "Analiza performanţei energetice a clădirilor nerezidenţiale prin tehnici de modelare numerică şi Inteligenţă Artificială aplicată," Teză de doctorat, Universitatea Tehnică din Cluj-Napoca, 2017. [Online]. Disponibil:
Această lucrare explorează aplicarea tehnicilor de modelare numerică şi a algoritmilor de inteligenţă artificială în analiza performanţei energetice a clădirilor. Studiul este relevant pentru proiectul actual, deoarece abordează utilizarea modelelor predictive pentru optimizarea consumului energetic în sistemele HVAC.

[3] IEEE Citation Reference, 2009. [Online]. Disponibil:
Acest ghid oferă regulile oficiale de citare conform standardelor IEEE. A fost utilizat pentru a asigura o structură corectă şi uniformă a referinţelor bibliografice din proiect.

[4] IEEE Editorial Style Manual, 2016. [Online]. Disponibil:
Manualul de stil IEEE a fost utilizat pentru redactarea documentului, asigurând respectarea standardelor de formatare, numerotare şi organizare a conţinutului tehnic.