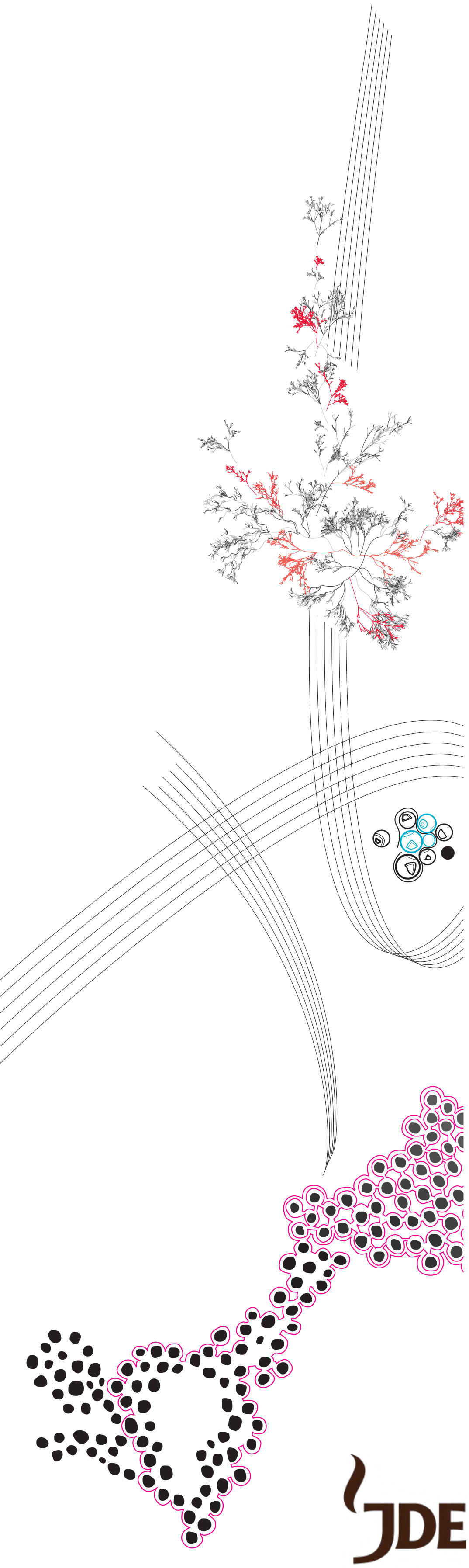


Moisture Optimization and Heating Process Automation of Freeze-Dried Coffee Production: A Case Study at Jacobs Douwe Egberts Peet’s

by

Michael
s2767708
michael@student.utwente.nl

A Master’s thesis submitted to the
Faculty of Electrical Engineering, Mathematics
& Computer Science (EEMCS)
in partial fulfilment of the requirements for the degree of

MSc in Business Information Technology -
Data Science & Business

Faculty of Electrical Engineering, Mathematics
& Computer Science (EEMCS)
University of Twente
Drienerlolaan 5
7522 NB Enschede
The Netherlands

JDE
JACOBS DOUWE EGBERTS

September 2023

UNIVERSITY OF TWENTE.

UNIVERSITY OF TWENTE

FACULTY OF ELECTRICAL ENGINEERING, MATHEMATICS & COMPUTER SCIENCE (EEMCS)

Moisture Optimization and Heating Process Automation of Freeze-Dried Coffee Production: A Case Study at Jacobs Douwe Egberts Peet's

Author

Michael
Data Science & Business
Faculty of Electrical Engineering, Mathematics & Computer Science (EEMCS)
University of Twente

University Supervisors

Marcos R. Machado
Industrial Engineering and Business Information Systems
Faculty of Behavioural, Management and Social Sciences (BMS)
University of Twente

Faiza A. Bukhsh
Data Management & Biometrics
Faculty of Electrical Engineering, Mathematics & Computer Science (EEMCS)
University of Twente

Company Supervisor

Murat Celik
Global Tech. CoE Manager
Mfg, StP, and R&D
Jacobs Douwe Egberts Peet's

September 20, 2023

Abstract

In the realm of freeze-dried coffee production, moisture plays a pivotal role due to its profound impact on product quality. Maintaining a consistent moisture level is of paramount importance for coffee manufacturers like Jacobs Douwe Egberts Peet's (JDE) to uphold stringent quality standards. However, existing manual procedures exhibit substantial moisture variation between batches and operational inefficiencies for engineers. This research endeavors to create an enhanced automated solution for JDE's freeze-dried coffee production, harmonizing quality and efficiency. It encompasses an extensive exploration of machine learning algorithms, with a focus on the effectiveness of non-linear models such as the eXtreme Gradient Boosting (XGBoost) Regression Model in moisture prediction. Simultaneously, optimization methods are scrutinized for dynamically adjusting temperature settings based on moisture forecasts. Among these, Adaptive Neuro-Fuzzy Inference System (ANFIS) emerges as the prime optimization method, showcasing its prowess in capturing intricate, nonlinear data relationships. Notably, ANFIS reduces the need for extensive manual engineering of linguistic variables and membership functions while maintaining a reasonable level of generalization. The culmination of this study is the successful deployment of the automated solution, resulting in moisture levels that are better aligned with the target, while accomplishing a significant reduction in its standard deviation.

Keywords: Moisture Optimization, Heating Process Automation, Freeze-Dried Coffee, Data Science, Machine Learning, Regression, Optimization Method, Fuzzy Logic, ANFIS.

Author's Declaration

I solemnly declare that the thesis titled *Moisture Optimization and Heating Process Automation of Freeze-Dried Coffee Production: A Case Study at Jacobs Douwe Egberts Peet's* presented herein is the culmination of my original work. The ideas, analyses, findings, and conclusions presented in this thesis are the result of my independent research and authorship, conducted during my thesis internship at Jacobs Douwe Egberts Peet's as part of my Master study program at the University of Twente.

I affirm that the content of this thesis is based on my own intellectual contributions, supported by references to the works of others duly cited and acknowledged. The data, experimental results, simulations, and methodologies presented in this thesis are authentic and have not been plagiarized or fabricated. Any contributions from collaborators or individuals have been appropriately credited.

I hereby grant the University of Twente the permission to lend this thesis to other academic institutions or individuals for the purpose of scholarly research. Additionally, I authorize the University of Twente to reproduce this thesis, either in its entirety or in part, by means such as photocopying or electronic reproduction, as requested by other academic institutions or individuals for scholarly research purposes.

Furthermore, I acknowledge that the University of Twente may choose to make this thesis electronically available to the public. This will facilitate wider access to the knowledge and insights contained within this research, contributing to the scholarly community and promoting further academic discourse.

I understand the significance of academic integrity and the consequences of any breaches thereof. I affirm that this thesis adheres to the ethical guidelines and academic standards set forth by the University of Twente and Jacobs Douwe Egberts Peet's.

Michael

ACKNOWLEDGEMENTS

In contemplating the decision to leave my beloved home country, Indonesia, to embark on a transformative journey pursuing my Master's degree, I could not help but reflect upon the profound words of Jeremiah 29:11, *"For I know the plans I have for you, plans to prosper you and not to harm you, plans to give you hope and a future".* This verse has been a steadfast beacon of faith throughout my life's journey. I wish to express my deepest gratitude to my Lord and Savior, Jesus Christ, for guiding me through this path and illuminating my way.

To my esteemed academic supervisors, Marcos and Faiza, whose unwavering support and scholarly guidance have been instrumental in shaping my research. Your mentorship has enriched my academic experience and opened new horizons of knowledge. I am equally indebted to my dedicated industry mentors in Jacobs Douwe Egberts Peet's, Murat, Ahmet, and Jorge, whose practical insights and encouragement have bridged the gap between theory and real-world application. My heartfelt appreciation extends to Andrey, the entire GIS Team, and my colleagues at the workplace, who fostered a collaborative and innovative work environment.

To my beloved family, I owe an immense debt of gratitude. To my dad, who is always my example and inspiration, words cannot describe the gratitude for your unwavering belief in my potential and the invaluable life lessons you've shared. To my mom, your constant reminders to stay close to God and walk in His path have will always be my guiding light in life. Thank you for visiting me during the peak moment in my thesis, providing me with unwavering emotional support and delicious meals every day. I am equally inspired by my grandmother's daily morning messages, a testament to the constancy of love.

To my dear friends, Vriendenkring (Yesaya, Raka, Aldi, Silvi, Irine, Afif) and Karang Taruna Enschede, words cannot express my gratitude for the countless unforgettable moments we've shared, from adventurous journeys across the country to joyous laughter-filled gatherings. A special mention to ASAF (Yesaya, Raka, Aldi, Afif, Daniel) for our time together playing music, and I will never forget our unforgettable performance in Utrecht. I am profoundly grateful to my brothers and sisters in Christ at International Christian Fellowship Enschede (ICF-E) and the Choir family, where I found a "home" away from home. A heartfelt thank you goes to the Indonesian Christian Fellowships for the enriching komsel gatherings and the opportunity to praise, worship, and grow together in Christ. To my Boomerang buddies, Tifani, I deeply appreciate your delicious but spicy food and advice, to Raef and Sam, for our fun game nights and togetherness in Maarssen. Lastly, to all my friends back in my hometown, whose bonds remain etched deep in my heart, and to the new friends I have made in the Netherlands, your friendship has been a continuous source of joy and support.

This journey has been a tapestry woven with the threads of faith, knowledge, and the warmth of cherished relationships. I am profoundly grateful for each individual and experience that has contributed to the journey of my life. You have all played an integral role in my life's story, and I look forward to the chapters yet to be written.

Maarssen, 20 September 2023
Michael

Contents

List of Figures

LIST OF TABLES

ABBREVIATIONS

R^2	R-squared.
AIC	Akaike Information Criterion.
ANFIS	Adaptive Neuro-Fuzzy Inference System.
ANN	Artificial Neural Network.
ARIMA	AutoRegressive Integrated Moving Average.
ARX	AutoRegressive models with eXogenous inputs.
AutoML	Automated Machine Learning.
CNN	Convolutional Neural Network.
CRISP-DM	CRoss-Industry Standard Process for Data Mining.
DT	Decision Tree.
EDA	Exploratory Data Analysis.
FD	Freeze Dryer.
FIS	Fuzzy Inference System.
FL	Fuzzy Logic.
FLC	Fuzzy Logic Controller.
GAM	Generalized Additive Models.
IA	Implementation & Automation.
ICO	International Coffee Organization.
JDE	Jacobs Douwe Egberts Peet's.
KDE	Kernel Density Estimate.
KNN	K-Nearest Neighbors.
LightGBM	Light Gradient-Boosting Machine.
LinReg	Linear Regression.
LSTM	Long Short-Term Memory.
MAE	Mean Absolute Error.
ML	Machine Learning.
MLFFNN	MultiLayer Feed-Forward Neural Network.
MLPNN	MultiLayer Perceptron Neural Network.
MRA	Multiple Regression Analysis.
NB	Naive Bayes.
OLS	Ordinary Least Square.
OM	Optimization Method.
RBFNN	Radial Basis Function Neural Network.
RF	Random Forest.

RMSE	Root Mean Squared Error.
RNN	Recurrent Neural Network.
SHAP	SHapley Additive exPlanations.
SLR	Systematic Literature Review.
SVM	Support Vector Machine.
SVR	Support Vector Regression.
XGBoost	eXtreme Gradient Boosting.

1
Introduction

1.1. Research Background

Coffee, being one of the most widely consumed and highly popular beverages worldwide, has experienced a substantial increase in its consumption trend. According to information gathered by the International Coffee Organization (ICO), the consumption and production of coffee worldwide have exhibited a steady and continuous growth pattern over the course of several years [4]. Among the various forms of coffee, instant coffee, specifically freeze-dried coffee stands out for its ability to capture and preserve the rich flavor and enticing aroma of freshly brewed coffee. This trend can be attributed to their convenience and long shelf life, offering coffee enthusiasts a quick and hassle-free way to enjoy their favorite beverage [5]. Moreover, freeze-dried coffee is one of the flagship products of Jacobs Douwe Egberts Peet's (JDE). In order to meet the rising demand for high-quality freeze-dried coffee products and maintain their competitive edge in the market, coffee manufacturers like JDE need to continuously innovate and refine their production processes to ensure consistent and superior coffee experiences for consumers.

The process of freeze-dried coffee production involves a unique preservation technique that ensures the retention of the coffee's desirable qualities. Appendix A depicts the end-to-end production of freeze-dried coffee. It begins by subjecting brewed coffee extract to freezing temperatures, significantly below the freezing point of water. This freezing stage is crucial as it facilitates the transformation of liquid water into ice within the coffee solution. Subsequently, the frozen coffee undergoes sublimation, a process in which the ice converts directly into water vapor by heating it, but without transitioning into a liquid state. Through this sublimation process, the water content is effectively removed, leaving behind coffee particles with enhanced stability and prolonged shelf life. At JDE, the heating process within the freeze-drying system plays a pivotal role in achieving the desired moisture levels in the final freeze-dried coffee.

The control of moisture levels throughout the freeze-drying process is of utmost importance in guaranteeing the desired quality and characteristics of the final freeze-dried coffee product. Any variations or deviations in moisture content can significantly impact the taste, aroma, solubility, and overall quality of the end product [6]. Consequently, it is crucial for JDE to employ effective techniques and technologies to optimize moisture levels and automate the heating process of freeze-dried coffee production.

1.2. PROBLEM CONTEXT

Currently, in JDE, the temperature control in the heating zones of the Freeze Dryer (FD) machine is either carried out manually by skilled machine engineers or relies on a simplistic regression model. These approaches, while serving their purpose to some extent, have limitations in terms of ensuring consistent and precise moisture optimization. Relying solely on manual control introduces significant variability in moisture levels across different batches. This variability is starkly illustrated in Figure 1.1, which showcases the moisture levels of trays within the same FD machine. The moisture level in the y-axis is omitted for confidentiality purposes. It is evident that even within a single machine run, there exists considerable fluctuation in the moisture content of the product in the same trays. The regions marked in red represent trays exceeding acceptable moisture criteria, which must then undergo a complete restart of the freeze-drying process, resulting in reduced freeze-dried coffee production. Furthermore, this variability in tray-level moisture propagates into the average moisture content of the completed stacks, shown in Figure 1.2, contributing to a notable disparity in the final product's quality, texture, and overall consistency. In the long term, this will not only lead to increased reject rates, but will also undermine the brand's reputation due to the unpredictable variations in moisture level of the product.

To address the challenge of maintaining consistent and precise moisture optimization in the freeze-drying process, certain operational practices are currently employed within JDE. Engineers are tasked with the continuous monitoring and adjustment of temperature settings to achieve the desired moisture levels, demanding a significant allocation of their time and resources. These labor-intensive tasks divert the engineers' attention from more value-added activities, such as process optimization, quality assurance, and innovation. Consequently, a substantial portion of their productive work hours are dedicated to the manual supervision of the freeze-dryer machine, detracting from their ability to focus on strategic tasks that could significantly enhance overall production efficiency and product quality.

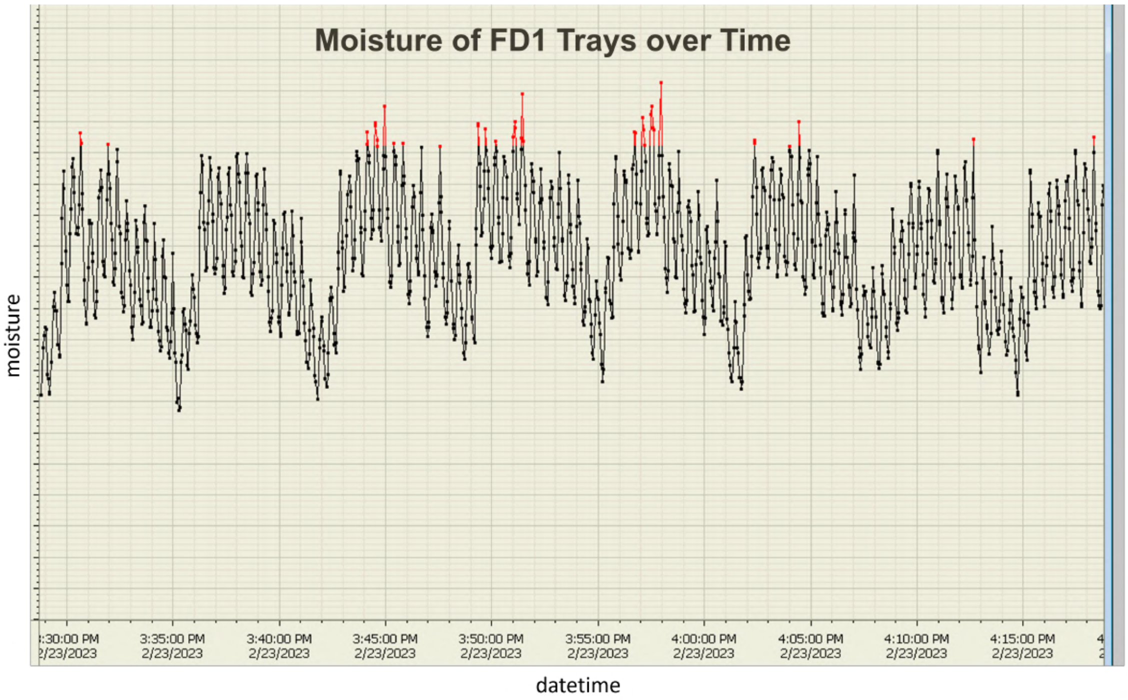


Figure 1.1: Moisture of Trays in Freeze Dryer 1 over Time

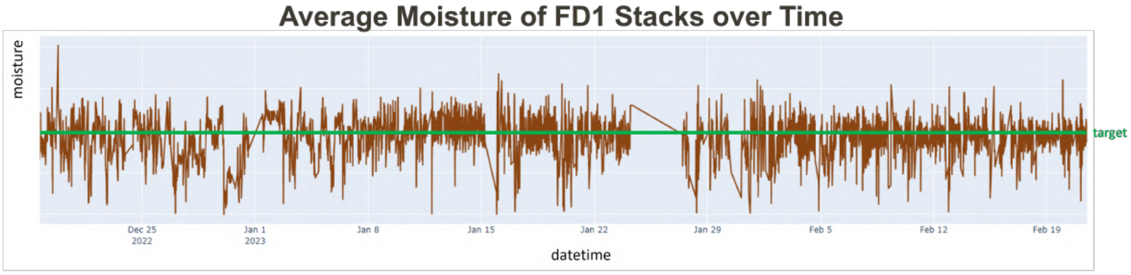


Figure 1.2: Average Moisture of Stacks in Freeze Dryer 1 over Time

1.3. RESEARCH QUESTIONS

In order to address the research goal effectively, a set of research questions has been formulated. These research questions serve as a framework for exploring the related existing literature and developing the solutions to the research problems.

Main Research Question:
How can the heating process of freeze-dried coffee at Jacobs Douwe Egberts Peet's (JDE) be optimized and automated to reduce the variability of its target moisture level?

Sub-Research Question:

1. **Sub-RQ1 (knowledge question)**: What are the underlying motivations and drivers for conducting research on the moisture optimization and heating process automation of freeze-dried coffee production?

Motivation: This question aims to understand the core motivations and drivers behind this research contextualizing the significance of optimizing and automating the heating process in freeze-dried coffee production at JDE. By identifying the underlying reasons, the research can lay a strong foundation for its potential impact on the freeze-dried coffee industry.

2. **Sub-RQ2 (knowledge question)**: What are the current methods and techniques used for the moisture optimization and heating process automation in the freeze-dried coffee production?
   Motivation: Gaining insights into current methods and techniques used for moisture optimization and heating process automation is essential for identifying existing practices. Understanding prevailing approaches helps advancing state-of-the-art techniques into a more efficient heating process.

3. **Sub-RQ3 (knowledge question)**: What are the challenges and limitations associated with the current approaches to the moisture optimization and heating process automation in the freeze-dried coffee production?
   Motivation: The obeective of this question is to recognize the challenges and limitations associated with current approaches, which is fundamental for devising solutions that address these critical issues. Assessing shortcomings allows formulating innovative strategies to overcome challenges and achieve superior moisture optimization and automation.

4. **Sub-RQ4 (knowledge question)**: What are the emerging trends of techniques, methods and future directions in the moisture optimization and heating process automation in the freeze-dried coffee production?
   Motivation: This question helps to identify emerging trends and future directions in moisture optimization and heating process automation. It is essential to stay at the forefront of technological advancements by proposing forward-thinking solutions that are aligned with industry trends, ensuring long-term relevance.

5. **Sub-RQ5 (design question)**: Which machine learning algorithm demonstrates the best performance for predicting the moisture levels of freeze-dried coffee?
   Motivation: Evaluating and selecting the most suitable machine learning algorithm based on the right metrics is critical for accurate moisture prediction. Rigorously testing various algorithms ensures precise and reliable moisture level predictions, enhancing product quality and consistency.

6. **Sub-RQ6 (design question)**: What optimization methods are the most suitable to automatically adapt the temperature settings in the heating zones based on the predicted moisture levels of freeze-dried coffee?
   Motivation: This question aims to find the most appropriate optimization methods to dynamically adjust temperature settings based on predicted moisture levels. Investigating

different optimization techniques ensures an adaptive heating process and improving production efficiency by reducing manual workloads of the machine operators.

7. **Sub-RQ7 (validation question)**: How does the optimized and automated heating process perform in comparison to the current processes?
   Motivation: Conducting a comprehensive performance evaluation is crucial for validating the effectiveness of the developed solution. Comparing the optimized and automated approach with current processes provides concrete evidence of its potential benefits for freeze-dried coffee production.

1.4. RESEARCH GOAL

The main goal of this research is to develop an optimized and automated approach for the heating process in freeze-dried coffee production at JDE. The initial hypothesis underlying this research is that the application of machine learning algorithms and optimization methods will reduce the variability in target moisture levels, leading to improved product quality with consistent moisture levels and enhanced efficiency by replacing manual processes with automated schemes. Initially, this study conducts a comprehensive examination of the current state of knowledge regarding predictive techniques, optimization methods, and the implementation of process automation across diverse applications and disciplines. By examining a wide range of academic articles and research papers, this thesis aims to first identify the existing gaps and research opportunities in the field of machine learning. Subsequently, it seeks to provide insights and motivation for further research on developing an optimized and automated approach for the heating process of the freeze-dried coffee at JDE.

To accomplish the main research goal, this research endeavor will focus on three specific objectives. First, it seeks to investigate various machine learning algorithms that can fittingly predict moisture levels, based on applications in similar fields such as agriculture forecast and climate prediction. The best-performing algorithm according to the pre-defined metrics will be deployed into production to generate the moisture prediction. Second, the study aims to implement the most suitable optimization methods that can adjust the temperature settings in freeze-dried coffee production based on the moisture prediction result. Lastly, this study will explore implementation approaches to replace manual processes with automated schemes, identifying advancements in automation technologies and techniques that can be applied to the automation process of freeze-dried coffee production at JDE.

1.5. RESEARCH STRUCTURE

This thesis consists of eight chapters, which are organized as follows. Chapter 1 sets the stage by outlining the research background, defining the problem context, formulating some research questions, and highlighting the research goal. Chapter 2 conducts a Systematic Literature Review (SLR) to address sub-research questions 1-4. This chapter also synthesizes existing theoretical knowledge on pertinent subjects and finding existing research gaps. Next, Chapter 3 introduces the primary methodology, delving into the theories underlying machine learning algorithms and optimization methods central to the study's execution. Chapter 4 then provides insights into data sources, Exploratory Data Analysis (EDA), and outlines the minimum functional and non-functional requirements of the solution. Furthermore, Chapters 5 and 6 focus on data preparation, modeling, and evaluation. Chapter 5 specifically answers sub-research question 5 by exploring machine learning models for moisture level prediction, while Chapter 6 provides answers to sub-research question 6 by delving into optimization methods for dynamically adjusting temperature settings. In Chapter 7, attention shifts to practical implementation, providing a comprehensive solution overview and presenting tangible outcomes from deployment efforts, where sub-research question 7 is addressed. Finally, Chapter 8 concludes the study, highlighting academic and practical contributions of this research, critically examining its limitations, and paving the way for future research endeavors. Figure 1.3 provides a visual overview of the research structure and indicating the specific sub-research questions addressed in each corresponding chapter, if applicable.

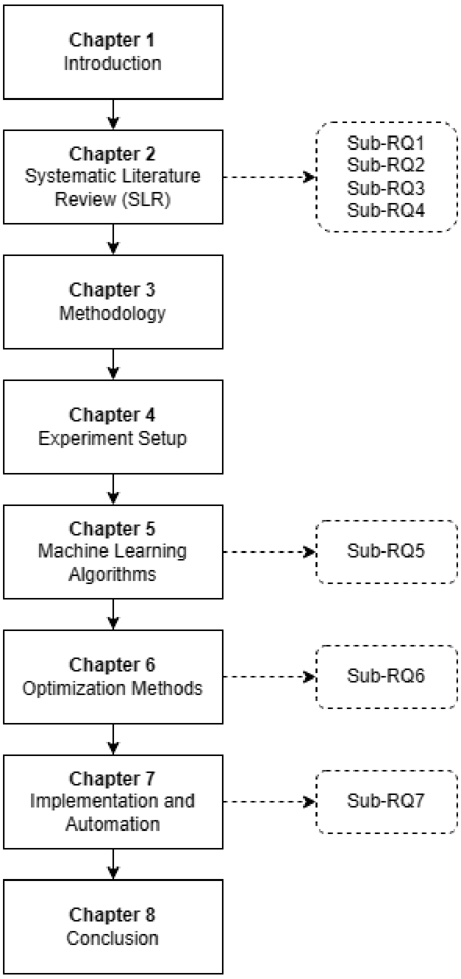


Figure 1.3: Research Structure

1.6. CHAPTER SUMMARY

In this chapter, the significance of freeze-dried coffee production is emphasized, with a specific focus on the critical role of the heating process in maintaining product quality. The chapter also sheds light on the prevailing manual control practices employed at JDE and the associated challenges. Then, a set of formulated research questions is introduced to guide the study's trajectory, encompassing motivations and practical implementation. The research objective is defined, which is formulating an optimized and automated heating process to enhance both product quality and production efficiency. Furthermore, three specific objectives are outlined, entailing an investigation into machine learning algorithms, optimization methods, and innovative deployment approaches, all devised to address the intricate challenges posed by freeze-dried coffee production.

2

Systematic Literature Review

This chapter presents the framework adopted to conduct the Systematic Literature Review (SLR) in this study, which is based on the guidelines proposed by Kitchenham & Charters [7]. The steps of planning and conducting the review are adapted from their framework, ensuring a rigorous and structured approach to the literature review process. The planning phase is the first step in the SLR, which involves determining research questions, selecting scientific databases, formulating search queries, and establishing inclusion and exclusion criteria. This is followed by conducting the review, which focuses on the selection of articles and the subsequent data extraction process. These steps are essential for identifying and including studies that are relevant to the research questions and objectives of this study, as well as for extracting pertinent information from the selected articles. Furthermore, this chapter provides a comprehensive analysis and interpretation of the findings from the SLR, aiming to uncover the trends in the literature, discuss the implications of the research questions, and identify the research gaps.

2.1. Planning the Review

2.1.1. Scientific Databases and Search Query Formulation

In this research, two main scientific databases were chosen for the retrieval of relevant articles, namely Scopus[1] and IEEE[2]. Scopus was chosen as it offers extensive coverage of scholarly publications from various fields, including peer-reviewed journals and conference papers. Its inclusion allows for a comprehensive analysis of the research landscape related to a particular topic [8]. In addition, the selection of IEEE, a renowned professional association, provides a specialized database focusing on computer science, decision science, and related disciplines. This choice emphasizes relevant studies and advancements in machine learning, automation,

[1] https://www.scopus.com/search/form.uri?display=advanced

[2] https://ieeexplore.ieee.org/search/advanced

and process optimization, which are essential to this research.

Formulating the search query is a crucial step in SLR, aiming to retrieve relevant works related to the main and sub-research questions. The search query consists of keywords and terminologies that are closely aligned with the research objectives. To enhance the comprehensiveness of the search, four main groups of queries were constructed, each comprising the main query and its corresponding synonyms. By incorporating synonyms, the search queries aim to encompass a broader range of relevant literature. Table 2.1 presents a comprehensive list of keywords associated with the main query, which will be executed on the selected scientific databases to retrieve pertinent literature for the review.

Table 2.1: Search Query Keywords

Machine Learning	**Optimization & Automation**	**Freeze-Dried Coffee**	**Artefact**
"Machine Learning"	"Fuzzy Logic"	Coffee	Model
Regression	Minimization	Moisture	MLflow
		Temperature	Algorithm

From Table 2.1, the search queries were developed by formulating logical relationships between keywords and their synonyms. Each keyword and its synonym were connected using the logical operator "OR," while groups of keywords were connected using the logical operator "AND." The queries were tailored to adhere to the format requirements of each scientific database, focusing on specific fields such as title, abstract, and keywords. The retrieval of relevant documentation was facilitated by utilizing the advanced search bar, which allowed for the input of multiple keywords and the creation of customized queries. The queries used on the aforementioned databases can be seen as below:

Scopus:

```
TITLE-ABS-KEY(
("machine learning" OR regression)
AND
("fuzzy logic" OR minimization)
AND
(coffee OR moisture OR temperature)
AND
(algorithm OR MLflow OR model))
```

IEEE:

```
(("machine learning" OR regression)
AND
("fuzzy logic" OR minimization)
AND
```

(coffee OR moisture OR temperature)
AND
(algorithm OR MLflow OR model))

2.1.2. INCLUSION AND EXCLUSION CRITERIA

The establishment of inclusion and exclusion criteria is crucial in conducting a SLR, as emphasized by Kitchenham and Charters [7]. These criteria play a vital role in refining the search results and ensuring that only relevant studies are considered for the review. By adhering to these predefined standards, the integrity and validity of the review are upheld, while potential biases arising from the inclusion of irrelevant studies are mitigated. Table 2.2 presents the inclusion and exclusion criteria, which outline the specific protocols for study selection.

Table 2.2: Inclusion and Exclusion Criteria

Inclusion Criteria	**Exclusion Criteria**
Literature was written in English	Duplicate literatures
Literature was published in the last 10 years	Full-text literature unavailability
Literature was a journal or conference proceedings in the Computer Science, Mathematics, Decision Science, or Business and Management areas	Irrelevant literatures based on its abstract to this study's defined research questions

The inclusion criteria employed in this SLR entail several key aspects. First, the selected literature must be written in English, ensuring accessibility and comprehension for the research team. Additionally, publications within the last 10 years were considered to maintain a contemporary perspective on the subject matter. Furthermore, specific disciplinary boundaries were established, encompassing Computer Science, Mathematics, Decision Science, and Business and Management. To adhere to scholarly rigor, preference was given to literature in the form of peer-reviewed journals or conference proceedings, which are recognized avenues for scholarly discourse.

To ensure the integrity and efficiency of the review process, steps were taken to eliminate redundant and unavailable literature. Duplicate sources were identified and removed from the pool of results obtained from the selected databases. Moreover, literature that was not accessible in its full-text form was excluded, as comprehensive analysis and interpretation require access to complete publications. In order to further streamline the review process and focus on pertinent research, a manual assessment was conducted to identify and exclude literature that was deemed irrelevant based on its abstract. By aligning with the defined research questions of this study, these assessments aimed to refine the selection and retain literature that directly contributes to the research objectives.

2.2. CONDUCTING THE REVIEW

2.2.1. SELECTION

The literature selection process was initiated by executing the formulated search query in the selected scientific databases. Since the search results may contain irrelevant studies, the application of the previously established inclusion and exclusion criteria becomes crucial. This step not only ensures the quality of the research collection but also streamlines the subsequent data extraction process by focusing on the inclusion of relevant or highly relevant studies. Figure 2.1 depicts a more detailed illustration of the selection process.

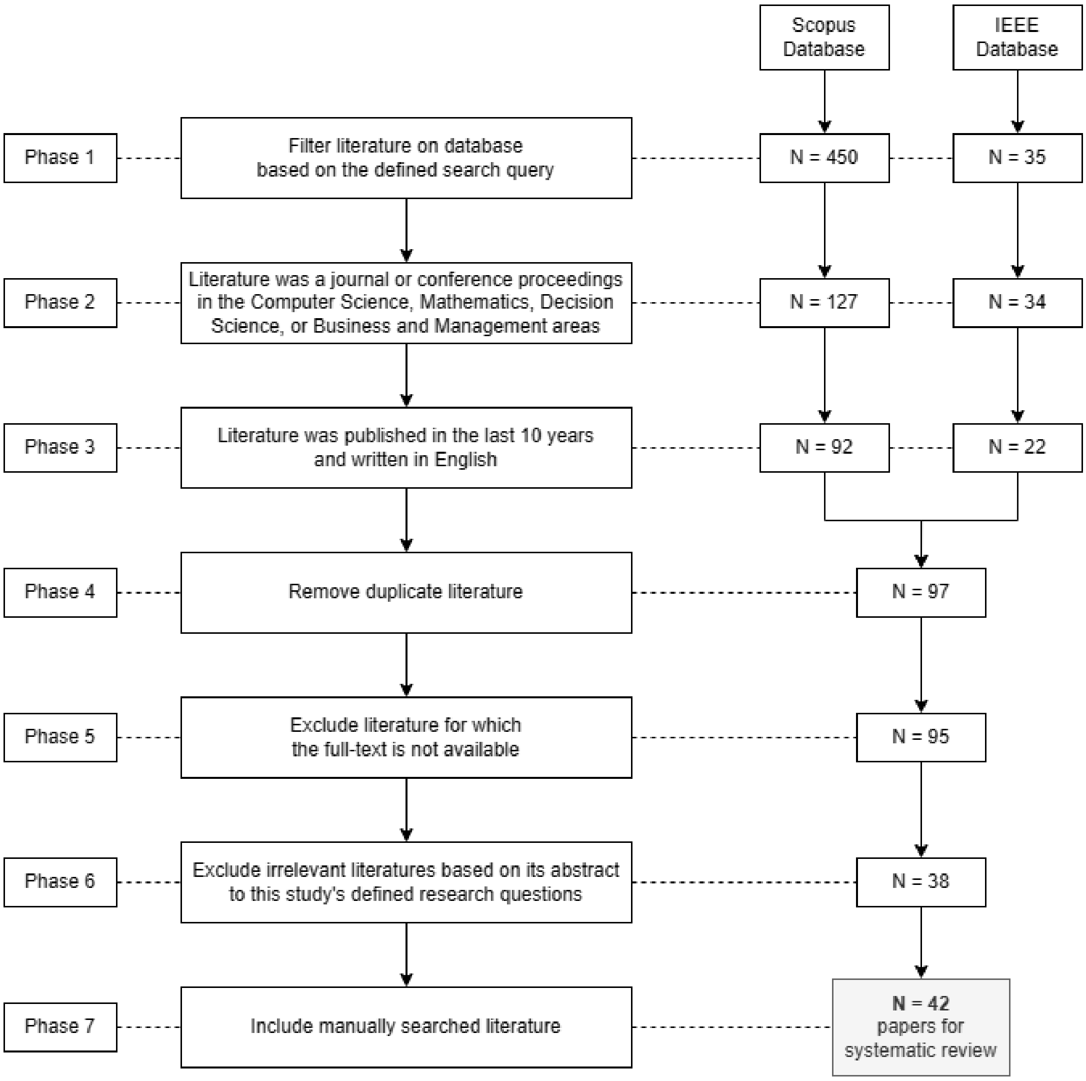

Scopus Database
IEEE Database
Phase 1
Filter literature on database based on the defined search query
N = 450
N = 35
Phase 2
Literature was a journal or conference proceedings in the Computer Science, Mathematics, Decision Science, or Business and Management areas
N = 127
N = 34
Phase 3
Literature was published in the last 10 years and written in English
N = 92
N = 22
Phase 4
Remove duplicate literature
N = 97
Phase 5
Exclude literature for which the full-text is not available
N = 95
Phase 6
Exclude irrelevant literatures based on its abstract to this study's defined research questions
N = 38
Phase 7
Include manually searched literature
N = 42 papers for systematic review

Figure 2.1: Literature Selection Phases

The literature selection process involved several phases, guided by the inclusion and exclusion criteria outlined in Table 2.2. After these phases, a total of 38 relevant articles were selected by applying the criteria, with phase 6 requiring manual assessment. Subsequently, four additional relevant articles were included, resulting in a final set of 42 papers for the SLR.

2.2.2. Data Extraction

Once the literature were selected, the data extraction process was initiated. In this research, the data extraction was divided into qualitative and quantitative analysis methods. These approaches were employed to gain a comprehensive understanding of the existing literature and extract meaningful insights from the selected articles.

The quantitative analysis conducted in this SLR encompasses the examination of three distinct domains: Machine Learning (ML), Optimization Method (OM), and Implementation & Automation (IA). The ML domain focuses on identifying whether the literature incorporates any machine learning methods or techniques. The OM domain aims to determine if the literature discusses the optimization of multivariate objective functions by finding optimal values for the variables within a defined set of constraints. Lastly, the IA domain evaluates whether the literature addresses real-case implementations and automation rather than solely theoretical or analytical discussions. By categorizing the literature according to these domains, the analysis provides insights into the prevalence and application of ML, OM, and IA approaches within the research landscape. Table 2.3 summarizes the findings, offering a comprehensive overview of the distribution and thematic emphasis of the included studies across these domains.

The qualitative analysis approaches implemented within the SLR encompass an extensive and in-depth exploration of the literature, delving into its intricacies. This analysis is organized to facilitate the understanding and synthesis of the gathered insights. Table B.1 in Appendix B provides valuable insights into the identified research purposes, current challenges, and potential future directions as discussed within the reviewed literature. For instance, Yuan [9] emphasized the benefits of integrating expert knowledge with machine learning approaches to leverage both the algorithmic predictive power and domain-specific insights. Sabrina [10] highlighted the utilization of Fuzzy Logic and Adaptive Neuro-Fuzzy Inference System (ANFIS) for temperature optimization based on various factors such as sensor types, crop types, and environmental conditions. Additionally, several studies, including Harsawardana [11], Imammuddien [12], Isikdemir [13], Mohapatra [14], and Patel [15], demonstrated the integration of Fuzzy Logic into the current systems, enabling precise control and autonomous decision-making processes.

Table 2.3: Quantitative Analysis of the Literature

Literature	Subject Area	Machine Learning	Optimization Methods	Implementation & Automation
M. Abbaspour-Gilandeh and Y. Abbaspour-Gilandeh (2019) [16]	Agriculture	FL, ANFIS	–	–
S. A. Abdul-Wahab, A. S. M. Omer, K. Yetilmezsoy and M. Bahramian (2020) [17]	Energy	FL, MRA	–	–
F. Al-Shanableh, M. Bilin, A. Evcil and M. A. Savaş (2020) [18]	Agriculture	FL, Multiple LinReg	–	–
K. N. Amrutha, Y. K. Bharath and J. Jayanthi (2019) [19]	Transportation	MRA, ANN, FL	–	–
K. Boma and S. Palizdar (2016) [20]	Energy	LinReg, FL, Fuzzy-Wavelet, ANN	–	–
K. T. T. Bui, D. Tien Bui, J. Zou, C. Van Doan and I. Revhaug (2018) [21]	Energy	SONFIS, SVR, MLPNN, Gaussian processes, RF, Different evolution-based neural FIS	–	The SONFIS model is constructed autonomously where the optimized antecedent and consequent parameters of the model were found autonomously with the use of the PSO algorithm
A. Choudhary, D. Pandey and S. Bhardwaj (2020) [22]	Energy	ANN	–	–
M. El Midaoui, M. Qbadou and K. Mansouri (2022) [23]	Healthcare	ARIMA, ANN, Transfer-learning FL	–	–
G. Ellina, G. Papaschinopoulos and B. K. Papadopoulos (2020) [24]	Agriculture	Fuzzy LinReg	–	–
M. Fauziyah, S. Adhisuwignjo, M. Rifai and D. Dewatama (2018) [25]	Engineering	–	FL	–
C. G. Gay and B. O. Bastien (2014) [26]	Climate	FL, FIS	–	–
M. K. Goyal, B. Bharti, J. Quilty, J. Adamowski and A. Pandey (2014) [27]	Climate	ANN, Least-squares SVR, FL, ANFIS	–	–
M. Gustin, R. S. McLeod and K. J. Lomas (2019) [28]	Climate	GAM, ARX	Minimization of AIC, Backward stepwise regression	–
Harsawardana, B. Samodro, B. Mahesworo, T. Suparyanto, S. Atmaja and B. Pardamean (2020) [11]	Engineering	–	FL	FL-based control system of the developed prototype

A. M. Imammuddien, S. Wirayoga and M. D. Muliono (2022) [12]	Engineering	–	FL	A block diagram system of the classification of coffee beans roasting maturity levels
Y. E. Isikdemir, G. Erturk, H. Ates and M. O. Tas (2022) [13]	Energy	RF, LinReg with various regularizations, SVR	FIS	Fuzzy inference-based controller
A. Khosravi, R. N. N. Koury, L. Machado and J. J. G. Pabon (2018) [29]	Energy	MLFFNN, RBFNN, SVR, FIS, ANFIS	–	–
J. Y. Kim (2022) [30]	Agriculture	RF	–	–
C. E. Lachouri, K. Mansouri and M. M. Lafifi (2022) [31]	Climate	ANFIS	–	–
T. L. Lam (2021) [32]	Engineering	SVR, RNN	FL (adjusting the temperature)	–
C. K. Leung, J. D. Elias, S. M. Minuk, A. R. R. d. Jesus and A. Cuzzocrea (2020) [33]	Transportation	Mean rule algorithm, RF, FL	–	–
S. Li (2019) [34]	Healthcare	KNN, DT, RF, ANN, FL	–	–
J. Liang, X. Liu and K. Liao (2018) [35]	Agriculture	FL, ANFIS, RF, ANN	–	–
D. M. Minhas, R. R. Khalid and G. Frey (2017) [36]	Energy	FL, ANN, LinReg	–	An implementation flow of the Hybrid Adaptive Fuzzy Neural System (HAFNS) algorithm
F. Mirzaei, M. Delavar, I. Alzoubi and B. Nadjar Arrabi (2018) [37]	Agriculture	Artificial bee colony algorithm (ABN-ANN), Multiple LinReg, ANFIS	Hybrid Mamdani FIS, Hybrid Sugeno FIS, Backpropagation Mamdani FIS, Backpropagation Sugeno FIS	–
A. G. Mohapatra and S. K. Lenka (2016) [14]	Agriculture	Partial least-square regression, ANN	–	An architecture of the complete real-time soil moisture content prediction methodology, and a block diagram of soil MC prediction model along with DSS for irrigation. FL is used to send SMS accordingly.
S. K. Mousavi Mashhadi, H. Yadollahi and A. Marvian Mashhad (2016) [38]	Engineering	–	–	PID fuzzy controller, which replaces the manual PID controller.
H. Neog, P. E. Dutta and N. Medhi (2022) [39]	Healthcare	Seasonal ARIMA, LSTM-Markov, LinReg, DT, KNN, K-means clustering	FL	–
A. H. Orta, I. Kayabasi and M. Tunc (2018) [40]	Energy	Directional Equivalent Plant Power Curve, Regression with different regularizations, SVR, ANN, Non-linear ARX, ANFIS	–	–
P. Patel, Y. Patel, U. Patel, V. Patel, N. Patel, P. Oza, et al. (2022) [15]	Agriculture	CNN, Various deep learning methods	FL (determining the exact time to irrigate the crop)	A model flowchart for the implementation of the project
V. K. Patil and V. R. Pawar (2022) [41]	Psychology	ANN, FL, K-means clustering, LinReg	–	–

B. Petković, D. Petković, B. Kuzman, M. Milovančević, K. Wakil, L. S. Ho, et al. (2020) [42]	Agriculture	ANFIS	–	–
M. R. C. Qazani, V. Pourmostaghimi, M. Moayyedian and S. Pedrammehr (2022) [43]	Engineering	ANFIS	Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Grey Wolf Optimization (GWO), Bayesian Optimization (B)	–
J. Refonaa and M. Lakshmi (2021) [44]	Climate	ANN, FL, Big data assisted Integrated Routing and Surplus Memory (BIRSM)	–	–
F. Sabrina, S. Sohail, F. Farid, S. Jahan, F. Ahamed and S. Gordon (2022) [10]	Agriculture	FL, SVM, KNN, NB	–	–
S. Sharma, R. K. Agrawal and M. M. Tripathi (2020) [45]	Energy	FL, RNN, SVM, ANFIS, Generic Regression Neural Network (GRNN)	–	–
A. Shastry, H. A. Sanjay and M. Hegde (2015) [46]	Agriculture	FL, ANFIS, Multiple LinReg	–	–
J. M. Siqueira, T. A. Paço, J. C. Silvestre, F. L. Santos, A. O. Falcão and L. S. Pereira (2014) [47]	Agriculture	FL	–	–
Q. T. T. Tran, K. Davies and L. Roose (2021) [48]	Energy	FL, SVM	–	–
V. Vivekanandhan, S. Sakthivel and M. Manikandan (2022) [49]	Agriculture	ANFIS, Variable Neighborhood Seach (VNS), Support Vector Clustering (SVC), Symbolic Aggregate approXimation & Vector Space Model (SAX-VSM)	–	–
H. Yuan, M. Tan and Y. Chen (2014) [9]	Agriculture	Radial Basis Function (RBF), SVM, LinReg, Non-linear Regression Model (NRM)	–	–
G. Zhang, S. S. Band, S. Ardabili, K. W. Chau and A. Mosavi (2022) [50]	Climate	ANFIS, Bilayered Neural Network (BNN)	–	–

2.3. TRENDS IN LITERATURE

2.3.1. YEAR-BASED TREND

Valuable insights into the evolving trends and patterns within the academic landscape are gained by analyzing the count of literature grouped by year. The results of this analysis, as depicted in the Figure 2.2, provide an academic perspective on the distribution and growth of research output over the years.

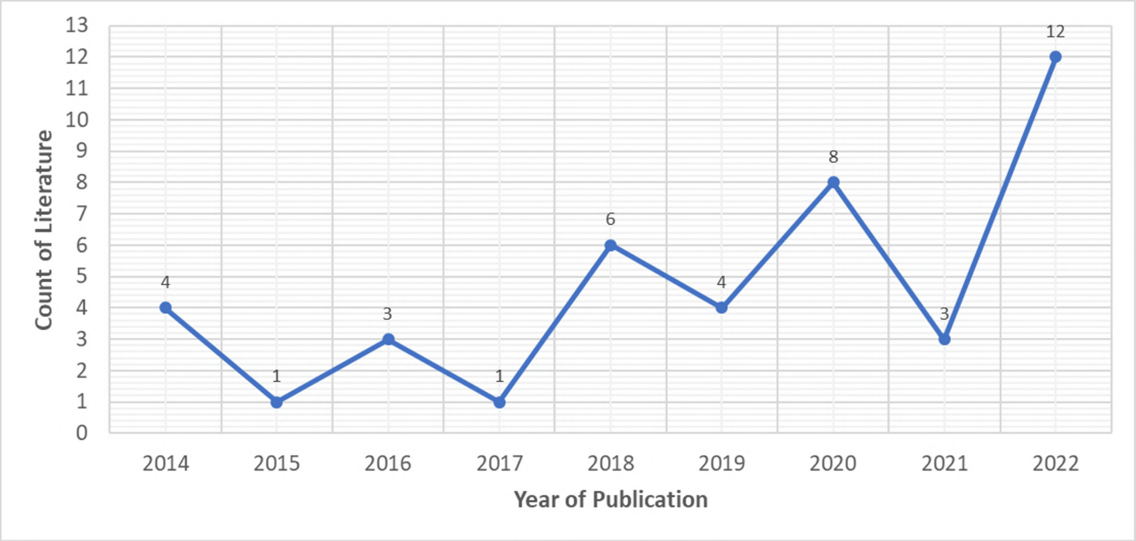


Figure 2.2: Year-based Trends of the Reviewed Literature

The data in Figure 2.2 reveals a notable increase in the number of publications from 2014 to 2022, indicating a growing interest in the particular research topic. Starting with a modest count of 1 to 4 publications between 2014 and 2017, the trend demonstrates a gradual rise with intermittent variations. The subsequent years show a mix of fluctuations, with varying publication counts. However, it is important to note that in 2018, there is a substantial surge in the number of publications, indicating a possible turning point or increased scholarly attention during that period on the search query formulation of this study.

Furthermore, the line graph shows an upward trend from 2019 to 2021, with a relatively high number of publications each year. However, the most significant surge in research output is observed in 2022, where the count reaches its peak at 12 publications. This notable increase suggests a heightened level of academic activity and interest in the research area, potentially grounded on the advancement of computational power, big data, and data science tools.

2.3.2. KEYWORDS-BASED TREND

This subsection examines the trends based on the keywords extracted from the reviewed literature. A word cloud visualization as depicted by Figure 2.3 is used to represent the distribution and prominence of keywords within the literature. The word cloud depicts the relative frequency of keywords derived from the literature analysis. It provides a visual representation where the size of each word corresponds to its frequency of occurrence.

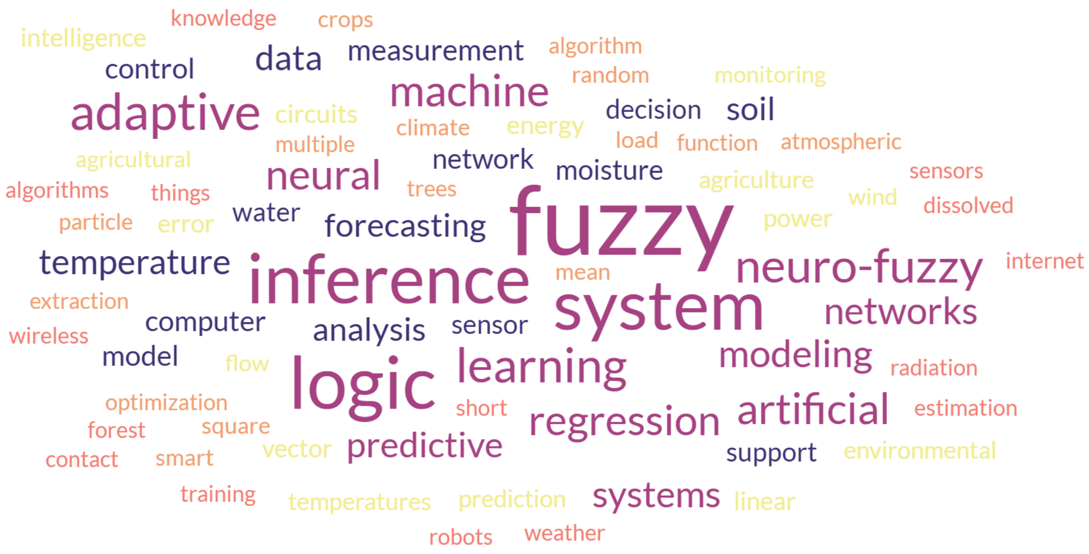


Figure 2.3: Word Cloud depicting the Keywords from the Reviewed Literature

The prominence of some keywords in the literature indicates their significant contribution to research and applications within the field. Fuzzy Logic emerges as a central concept in the literature, with a notable frequency in the word cloud. Fuzzy Logic is widely recognized for its ability to handle uncertainty and its explainability in decision-making processes. The mention of the Adaptive Neuro-Fuzzy Inference System (ANFIS) in the word cloud also highlights its relevance and utilization in the research domain. ANFIS combines the adaptive capabilities of neural networks with the interpretability of Fuzzy Logic, making it an attractive choice for modeling complex systems. Moreover, machine learning techniques, including Artificial Neural Network (ANN) and regression models, play a vital role in the research landscape, as indicated by their appearance in the word cloud. Machine learning algorithms enable the extraction of valuable insights from data and the development of predictive models. Furthermore, the presence of subject-specific keywords such as environmental, agriculture, and climate highlights the importance of these domains within the research landscape. Researchers in the field especially recognize the impact of environmental factors and agricultural practices, as well as the study of climate-related phenomena.

2.3.3. SUBJECT AREA-BASED TREND

The subject area-based trend analysis reveals interesting patterns and highlights potential research area gaps that can be explored in the context of this study. Figure 2.4 illustrates the distribution of the reviewed literature in this study based on the subject area.

The analysis reveals that Agriculture and Energy are the two primary subject areas that have received significant attention, with 14 and 10 publications, respectively. Within the Agriculture domain, notable research areas that can be grouped together include soil quality analysis [14, 16, 35, 37, 49] and crop-specific irrigation management [10, 14, 15, 46, 49]. In the Energy

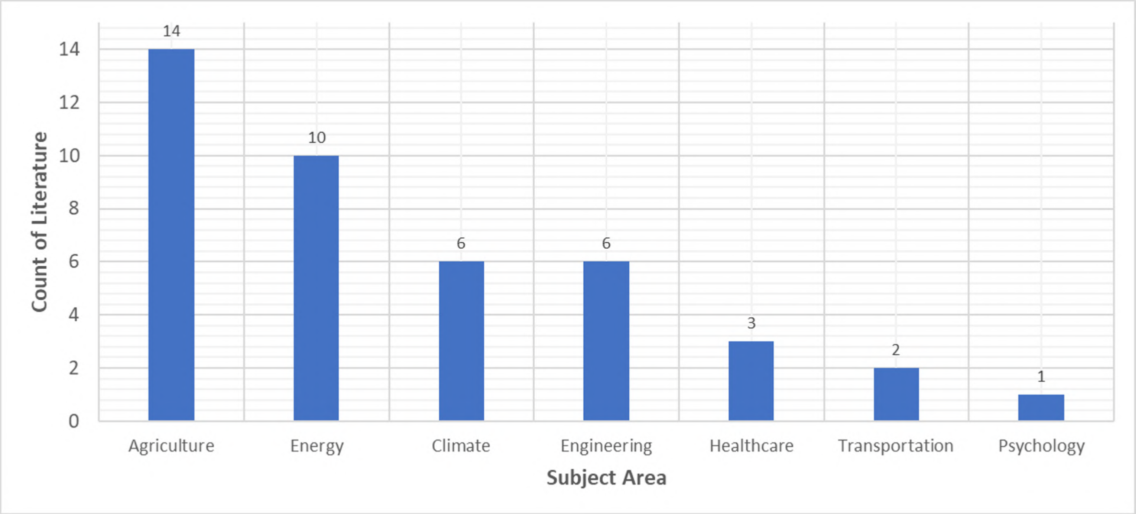


Figure 2.4: Subject Area-based Trends of the Reviewed Literature

domain, research topics such as solar radiation estimation [22, 29, 42] and electricity load forecasting [20, 36] have been explored. Additionally, other subject areas like weather forecasting [27, 28, 31, 44, 50] and coffee roasting process control [11, 12, 25, 30] have also received attention in a few studies.

2.4. DISCUSSION ON RESEARCH QUESTIONS

2.4.1. SUB-RQ1: MOTIVATIONS AND DRIVERS IN FREEZE-DRIED COFFEE PRODUCTION

The first sub-research question seeks to explore the underlying motivations and drivers for conducting research on the moisture optimization and heating process automation of freeze-dried coffee production. Several factors contribute to the motivation for investigating moisture optimization and heating process automation in freeze-dried coffee production.

Firstly, the demand for freeze-dried coffee products has been steadily increasing due to their convenience and extended shelf life [5]. Consumers are seeking high-quality instant coffee that closely resembles the flavor and aroma of freshly brewed coffee. In particular, achieving the desired moisture levels in freeze-dried coffee is crucial for preserving the taste and sensory characteristics of the coffee. Any variations or deviations in moisture content can significantly impact the quality and overall experience of the end product [6]. Thus, the optimization of moisture levels and the automation of the heating process are of paramount importance in meeting consumer expectations and maintaining product consistency.

Secondly, the current approaches to moisture optimization and heating process control in freeze-dried coffee production are limited in their ability to ensure consistent and precise moisture levels. According to a discussion with a system engineer in one of the JDE factories, manual temperature control by skilled machine engineers or simplistic regression models introduce variability in moisture levels across different batches. This variability can lead to inconsistent product quality and pose challenges for manufacturers in meeting quality standards and customer expectations. Therefore, there is a need to develop optimized and automated approaches that can minimize variability, enhance efficiency, and improve product quality.

Moreover, advancements in machine learning and automation technologies offer promising opportunities for optimizing the moisture levels and automating the heating process in freeze-dried coffee production. Machine learning algorithms can analyze complex data patterns and make accurate predictions, enabling real-time adjustments to temperature settings based on moisture level predictions [15]. By leveraging these technologies, coffee manufacturers can achieve more precise control over the moisture optimization process, reducing variability, and ensuring the production of high-quality freeze-dried coffee consistently.

2.4.2. SUB-RQ2: METHODS AND TECHNIQUES IN MOISTURE OPTIMIZATION AND HEATING PROCESS AUTOMATION

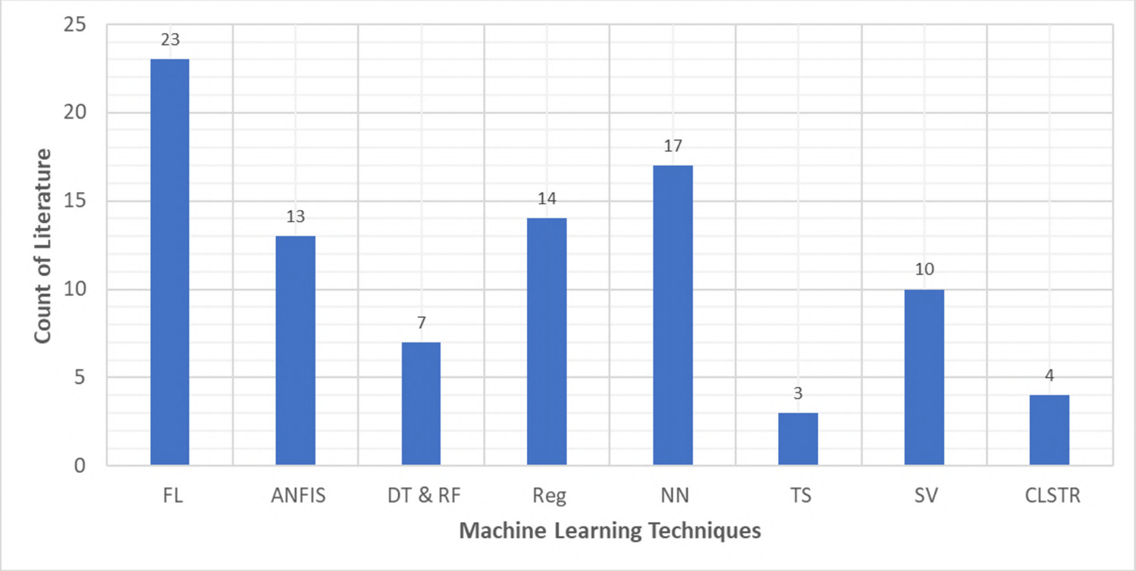


Figure 2.5: Implemented Machine Learning Techniques in the Reviewed Literature

Various machine learning models[3] have been employed in the studies, as shown in Figure 2.5. Fuzzy Logic emerged as the most used technique, constituting over 50% of the literature reviewed, which is justified by the aforementioned word cloud in Figure 2.3 [10, 13, 16–20, 23, 24, 26, 27, 29, 32–36, 41, 44–48]. Moreover, ANFIS [16, 21, 29, 31, 35, 37, 40, 42, 43, 45, 46, 49, 50], tree-based models such as Decision Tree and Random Forest [13, 21, 30, 33–35, 39], Regression models (Reg) [9, 13, 14, 17–20, 24, 36, 37, 39–41, 46], deep learning algorithms such as Neural Networks (NN) [14, 15, 19–23, 29, 32, 34–36, 40, 41, 44, 45, 50], Time Series models (TS) [23, 28, 39], Support Vector (SV) [9, 10, 13, 21, 29, 32, 40, 45, 48, 49], and Clustering algorithms (CLSTR) [10, 34, 39, 41] have also been utilized to address different machine learning tasks, such as agriculture prediction, climate forecasting, and medical prognosis. These diverse ap-

[3] FL = Fuzzy Logic; ANFIS = Adaptive Neuro-Fuzzy Inference System; DT&RF = Decision Tree and Random Forest; Reg = Regression models; NN = Neural Networks; TS = Time Series models; SV = Support Vector; CLSTR = Clustering algorithms

proaches reflect the efforts to explore and leverage the predictive potential of various machine learning techniques.

Optimization methods have played a crucial role in refining the performance of machine learning models, which aim to discover optimal values for variables that maximize or minimize a multivariate objective function, while taking into account certain limitations or conditions. However, the reviewed literature indicate a relatively limited application of optimization methods compared to machine learning techniques, with Fuzzy Logic being the most commonly used optimization approach in the reviewed literature [11, 13, 15, 25, 32, 37, 39]. Other optimization methods, such as minimization of Akaike Information Criterion (AIC) [28], backward stepwise regression [28], and optimization algorithm in Fuzzy Inference System [43], have been utilized to a lesser extent in the different subject areas.

Furthermore, it is worth noting that only less than 20% the reviewed studies address real-case implementations and automation, for instance, advanced PID controllers have been utilized in certain studies to enhance the performance and efficiency of the decision-making processes [21, 36, 38], also the block diagrams and flowcharts in some studies illustrate the integration of Fuzzy Logic into the system, enabling precise control and autonomous decision-making processes [11–15]. These studies go beyond theoretical or analytical discussions and provide practical insights into the application of the proposed techniques in real-world scenarios.

2.4.3. Sub-RQ3: Challenges and Limitations in Current Approaches

Implementing machine learning models faces numerous challenges and limitations, as highlighted in the literature. One of the primary challenges is related to the availability and quality of data, which has a significant impact on the performance and accuracy of machine learning models [17, 23, 27, 28, 32, 34]. The requirement for extensive data collection and regression poses practical challenges in implementing certain machine learning methods. Additionally, poor-quality data can limit the model's ability to generalize and make accurate predictions. Lam [32] highlighted the shortcoming of their proposed method, which may necessitate data collection and regression for every new product to maintain high-temperature approximation accuracy. This requirement can be time-consuming and resource-intensive, making it a limitation for practical implementation. Researchers have proposed various strategies to tackle this challenge. For instance, Abdul-Wahab et al. [17] suggested the introduction of additional model components and the specification of new features and functions to improve the performance of machine learning models. Similarly, El Midaoui [23] and Li [34] emphasized the need for additional experimental data from the literature to enhance the validity of implemented deep learning strategies.

Another challenge in machine learning implementation is the limited availability of input variables due to the lack of data. Using only a few input variables, such as minimum and maximum temperatures, may result in poor model estimates [27]. Researchers propose incorporating state-of-the-art machine learning methods, employing proper pre-processing techniques, and creating model ensembles to address this limitation. Additionally, forecasting beyond the ranges for which the models were originally trained can introduce increasing uncertainty and amplified errors [28].

The complexity of processes and the lack of domain expertise involvement pose significant challenges in machine learning implementation. These challenges have prompted researchers to explore various strategies to overcome them. For example, in power load forecasting, Minhas et al. [36] found that errors, especially for weekends, could be reduced by developing a hybrid adaptive fuzzy neural system (HAFNS). The HAFNS incorporates probabilistic and stochastic approaches based on day-ahead profiles, thereby improving forecasting accuracy. Similarly, in domains such as fisheries, researchers have addressed the challenge of domain expertise to further enhance model accuracy [9].

Moreover, the explainability of machine learning models is also a topic of concern. While more complex models may exhibit higher performance in training and validation datasets, their interpretability may be reduced. For example, Bui et al. [21] found that using a novel hybrid artificial intelligent approach in forecasting the horizontal displacement of hydropower dams resulted in high performance but perplexing interpretability. Similarly, the combination of multiple machine learning models can lead to improved predictions, at the cost of explainability. Patil & Pawar [41] achieved significant improvements in accuracy and reliability in emotion recognition by integrating clustering and regression algorithms. Sharma et al. [45] also reported higher accuracy in load forecasting by synergistically using Fuzzy Logic with an RNN model. However, both studies concluded that such combinations resulted in solutions that are difficult to interpret.

Integration of machine learning models with existing systems also poses challenges. While numerous machine learning techniques and optimization methods have been proposed in the literature, their implementation and integration into real-world systems remain limited [11, 12, 15]. The few instances where implementation has occurred often lack automation, suggesting the need for further development in this area.

Addressing cost and resource constraints is a crucial consideration in developing practical machine learning solutions. Researchers have proposed cost-effective approaches, such as non-contact temperature approximation and control systems, to provide low-cost and accurate solutions for specific applications [32]. However, these methods may require frequent data collection and regression for new products to maintain high-temperature approximation accuracy, which poses a potential limitation [32]. Similarly, in fields like manufacturing, advanced machine learning techniques can be time-consuming and expensive experiments, highlighting the needs to optimize cost and resource utilization [43].

2.4.4. Sub-RQ4: Emerging Trends and Future Directions

While current approaches to implementing machine learning models face various challenges and limitations, researchers are actively exploring emerging trends and future directions to overcome these hurdles and advance the field. One important aspect that is gaining attention is the need for model interpretability and explainability. Research has shown that more complex non-linear models do not necessarily result in better forecasts [28]. Instead, the focus is shifting towards developing models that are interpretable and easy to understand for end-users. This can foster trust and acceptance of machine learning systems, making them more practical and effective in real-world applications. The use of Fuzzy Logic and ANFIS is one approach that enhances interpretability and allows for customization based on different factors such as sensor types, crop types, and environmental conditions [10].

Incorporating domain expert knowledge, cluster analysis, function fitting, and nonlinear regression techniques have shown promise in improving model accuracy [9]. By integrating expert knowledge with machine learning approaches, models can benefit from both the computational power of algorithms and the domain-specific insights of experts. Additionally, the combination of Fuzzy Logic and machine learning is an area that presents potential opportunities for addressing complex problems across various domains [23]. This combination allows for a balance between automated decision-making and human expertise, enabling effective problem-solving in uncertain and dynamic environments.

Looking ahead, there is also a need to consider cost and resource constraints in machine learning solutions. Researchers are exploring low-cost methods, such as FL-based systems, to provide accurate solutions in a cost-effective manner [32, 48]. Optimization techniques, including Fuzzy Logic and Linear Regression, are also being employed to improve the efficiency of machine learning algorithms while minimizing the need for time-consuming and expensive experiments [43]. Furthermore, advancements in technology are shaping the future of machine learning implementation. For example, the integration of web-based user interfaces and back-end servers can facilitate remote configuration and monitoring of machine learning algorithms, enhancing their practicality and ease of use [13]. Such advancements can enable real-time data analysis and decision-making, leading to more efficient and responsive systems.

2.5. CHAPTER SUMMARY

In summary, this chapter focuses on the planning and execution of the Systematic Literature Review. The selection of scientific databases, formulation of search queries, and establishment of inclusion and exclusion criteria were described in detail. The review resulted in the identification of 42 relevant academic literature that discuss the application of machine learning techniques and optimization methods in various related fields. The chapter also analyzed the trends in literature based on year, keywords, and subject areas, providing insights into the research landscape. Furthermore, the discussion on research questions highlighted the motivations, methods, challenges, and future directions in the field, including emerging trends in machine learning and the potential for real-world implementation.

It is worth noting that the area of Manufacturing appears to be relatively underexplored, presenting a potential research gap that this study aims to address. Specifically, within the Engineering domain, which closely relates to Manufacturing, only 2 out of the 6 studies implemented Machine Learning techniques [32, 43]. Furthermore, none of these studies incorporated all three key components of the proposed future solutions, namely Machine Learning Algorithms, Optimization Methods, and Implementation & Process Automation. This indicates the opportunity to bridge this gap and contribute to the field by integrating these key elements in the context of Manufacturing. This integration can lead to the development of a decision support system that is composed of practical and efficient ML solutions that optimize manufacturing operations and enhance decision-making processes, thus improving productivity.

Moreover, the utilization of Fuzzy Logic algorithms in the context of system optimization and real-time decision-making, as highlighted by Isikdemir et al. [13] and Lam [32], represents a promising area for future exploration within the Manufacturing domain. The interpretability and adaptability of Fuzzy Logic make it well-suited for addressing complex manufacturing problems and enabling real-time adjustments based on dynamic conditions. Additionally, the integration of Fuzzy Logic with machine learning models, as proposed by El Midaoui [23], offers an avenue for developing innovative approaches to problem-solving in manufacturing. Investigating the potential of ANFIS models, which combine the advantages of Fuzzy Logic and neural networks, could be another fruitful research direction in this domain [21, 29, 35, 40, 45, 50].

3

METHODOLOGY

This chapter outlines the methodology adopted to achieve the objectives of the study. It begins by providing an overview of the CRoss-Industry Standard Process for Data Mining (CRISP-DM), a widely recognized framework for guiding the data mining process. Each step of this framework is introduced. Following this, it also briefly discusses some theories behind the machine learning algorithms and optimization methods which are utilized in the study.

3.1. CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING

The CRoss-Industry Standard Process for Data Mining (CRISP-DM) is a widely recognized and comprehensive framework for guiding the process of data science projects. It provides a structured and systematic approach to tackling complex data science tasks by breaking them down into manageable stages and activities. CRISP-DM offers a set of well-defined steps that assist researchers and practitioners in navigating the complexities of data-driven projects, ensuring that each phase is properly executed and aligned with the project's goals [51]. In the context of this study, CRISP-DM is employed as the guiding framework to structure and conduct the research. The study can ensure a consistent and well-documented process by adhering to the CRISP-DM methodology, which is essential for producing reliable and replicable results. Additionally, CRISP-DM aids in the integration of various techniques and methodologies at different steps, promoting a holistic and interdisciplinary approach to building solutions to the research problems [52]. Figure 3.1 illustrates the key steps of the CRISP-DM framework. The framework involves six major steps, each with its own distinct set of activities and objectives, which will be further explained in the following subsections.

3.1.1. BUSINESS UNDERSTANDING

The Business Understanding phase in the CRISP-DM framework is essential because it sets the direction for the data analysis process. This phase is about understanding the business or sec-

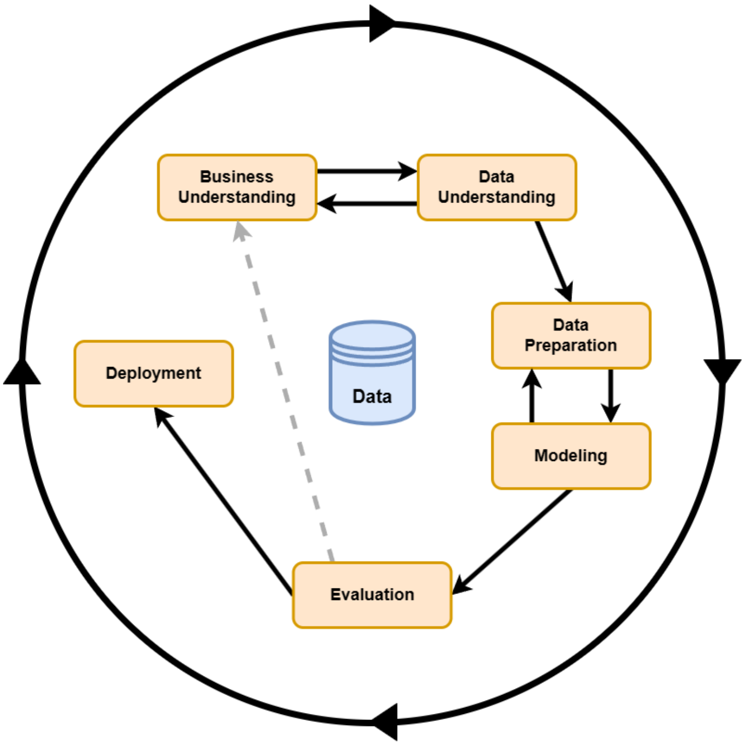


Figure 3.1: CRoss-Industry Standard Process for Data Mining (CRISP-DM) Framework (Adapted from [1])

tor being studied, clearly defining the goals and understanding the challenges. It's crucial to interact with stakeholders to determine the project's main objectives and potential hurdles. Instead of solely focusing on data, this phase highlights the broader picture, making sure that the upcoming data analysis aligns with business needs and goals, resulting in a better prepared analysis to offer practical and relevant outcomes. This phase was primarily carried out in Chapter 1.

3.1.2. Data Understanding

Continuing from the Business Understanding phase, the Data Understanding phase in the CRISP-DM framework plays a pivotal role in refining the project's direction. This phase shifts the focus towards identifying, gathering, and scrutinizing datasets that hold the potential to facilitate the accomplishment of project objectives. This phase encompasses a set of interrelated tasks, each contributing to a comprehensive understanding of the data's characteristics and its alignment with the research goals.

The first task within this phase involves the collection of initial data. This necessitates the acquisition of the relevant data from various sources and, if applicable, its integration into the chosen analysis tool. This initial data serves as the cornerstone on which subsequent analyses are built. Subsequently, the second task pertains to the description of the acquired data. During this step, the data is meticulously examined, and its surface properties are documented. Details such as the format of the data, the number of records, and the identities of various fields are captured. Moving further, the third task delves into the exploration of the data. This involves a more profound investigation, encompassing queries and visualizations to unveil underlying patterns, relationships, and insights residing within the data. This process not only provides an in-depth grasp of the data's nuances but also sets the stage for generating hypotheses and refining subsequent analyses. Lastly, the fourth task involves the verification of data quality. This task is crucial in assessing the cleanliness and integrity of the data. Any anomalies, inaccuracies, or inconsistencies within the data are documented at this stage. Addressing data quality issues is imperative for ensuring the validity and reliability of subsequent analyses and conclusions.

Looking forward, Section 4.1 will expound upon these principles by delving into the specifics of the current data environment within JDE's freeze-dried coffee production. Through Exploratory Data Analysis (EDA), this section will further reveal the dataset's characteristics and relationships, setting the stage for informed decision-making processes in the following stages of the study. This preliminary exploration stands as a vital precursor to the subsequent analytical phases, guaranteeing that the analysis is rooted in a profound comprehension of the data.

3.1.3. DATA PREPARATION

The Data Preparation, often referred to as feature engineering, phase plays a pivotal role within the CRISP-DM framework. This phase revolves around the essential task of refining the dataset to make it suitable for modeling purposes. It encompasses five primary tasks, each contributing to the transformation of raw data into a more structured and usable form. During the **Select data** task, the decision-making process revolves around choosing the datasets that will be used for analysis, with clear documentation of the rationale behind their inclusion or exclusion. Subsequently, the **Clean data** task assumes prominence, addressing issues related to data quality. This task involves solving errors, handling missing values, and removing outliers, ensuring that the subsequent analyses are not compromised by inaccurate or inconsistent data points. Furthermore, the **Construct data** task involves the creation of new attributes derived from the existing dataset. This process aims to enhance the dataset's predictive power by introducing relevant and meaningful variables. Following this, the **Integrate data** task focuses on joining data from multiple sources, thereby enriching the dataset's breadth and depth. Lastly, the **Format data** task involves standardizing the data format to facilitate consistent processing and analysis. For instance, converting text-based numeric values into actual numeric values is a common formatting step. It is important to note that not all of them are required in every project, and they may not necessarily be executed in the exact order prescribed. The specific tasks undertaken depend on the nature of the data, the project's objectives, and the insights sought from the analysis.

In addition to these tasks, a pivotal role is played by the train/test split, ensuring that a subset of the dataset (approximately 80%) is utilized for training, while the remaining portion (about 20%) is reserved for testing. This partitioning enables model training and validation on independent, unseen data, validating its generalization capabilities [53]. Another essential step is feature scaling, which ensures that independent features are standardized to a common range using the StandardScaler[1] approach, given by:

$$\text{Standardized Value} = \frac{\text{Original Value} - \text{Mean}}{\text{Standard Deviation}}$$

This not only mitigates the impact of variables with varying magnitudes but also transforms the dataset into a standardized distribution with a mean of 0 and a standard deviation of 1. This standardized distribution aids in improving the model's convergence during training, enhancing its optimization and prediction accuracy [54].

In the following chapters, Section 5.1 elaborates on the data preparation procedures undertaken for the Machine Learning aspect of the research. This section outlines the strategies used to clean, transform, and engineer features in preparation for machine learning algorithms. Similarly, Section 6.1 provides insights into the data preparation processes tailored for the Optimization Method. Here, the focus lies in structuring the data to effectively interface with optimization algorithms, ensuring that the ensuing analyses and model development are built on a robust foundation of well-prepared data.

3.1.4. MODELING

The Modeling phase in the CRISP-DM framework marks a significant transition from the preparatory stages to the actual development of data science models. This phase is characterized by the construction of predictive or descriptive models based on the refined dataset obtained through the Data Preparation phase. The primary objective of the Modeling phase is to generate models that capture meaningful patterns, relationships, and insights from the data to address the research questions or objectives. This involves selecting appropriate modeling techniques, training models on the dataset, and assessing their performance. In more detail, the steps to build a data science model include splitting the dataset into training and testing sets, selecting the appropriate algorithm, training the model on the training set, tuning hyperparameters to optimize performance, and evaluating the model on the testing set. The goal is to create models that generalize well to new, unseen data and can make accurate predictions or provide valuable insights.

In the aforemenationed hyperparameter tuning, GridSearch[2] and K-fold validation play crucial roles within the Modeling phase. GridSearch involves an exhaustive search over a predefined set of hyperparameters to determine the combination that yields the best model performance. This technique helps in optimizing the model's hyperparameters and enhancing its predictive power. In addition, K-fold validation addresses the need for robust model evaluation by partitioning the dataset into multiple subsets (folds) and iteratively training and testing the model on different combinations of these folds. This approach provides a more comprehen-

[1] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[2] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

sive understanding of the model's performance across various data partitions, thereby reducing the risk of overfitting to specific subsets of the data. Both GridSearch and K-fold validation contribute to achieving more accurate and reliable models, aligning with the objective of the Modeling phase.

In the context of model interpretation and understanding, another valuable technique that complements the Modeling phase is the use of SHapley Additive exPlanations (SHAP) [55–57]. SHAP is a method that helps explain the output of machine learning models by attributing the contribution of each input feature to the model's prediction. This technique provides insights into which features have the most significant impact on the model's outcomes and how they influence the prediction. Interpreting SHAP values often involves visualizing the contributions of individual features to the predictions. SHAP plots provide a clear graphical representation of these contributions, helping to understand how changes in feature values lead to changes in the model's prediction. In these plots, each feature's contribution to a specific prediction is displayed along a horizontal axis. The vertical position of the feature on the plot represents the feature's value for that prediction. The colors of the plot indicate whether a high or low feature value is associated with a higher or lower prediction, respectively.

Section 5.2 delves into the modeling processes related to the Machine Learning aspect of the study. It details the selection of machine learning algorithms that are best suited to address the specific research questions and dataset characteristics. The modeling section covers the training of machine learning models using the prepared dataset, including techniques to optimize the models' performance through hyperparameter tuning. Similarly, Section 6.2.1 and 6.3.1 focus on the modeling processes pertaining to the Optimization Method aspect of the study. It outlines the formulation and testing of optimization methods to enhance the decision-making process regarding the temperature in the domain of freeze-dried coffee production.

3.1.5. EVALUATION

The Evaluation phase within the CRISP-DM framework serves as a critical assessment of the models developed during the Modeling phase. This phase involves rigorously evaluating the performance and validity of the constructed models against predefined criteria. The objective is to determine how well the models generalize to new data and whether they effectively address the research questions or objectives. The evaluation process entails using appropriate metrics to measure the relevant measures of the model based on the nature of the problem. In the context of the current study, which is focused on regression tasks, several evaluation metrics are employed, including the R-squared (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE).

The R^2 metric quantifies the proportion of the variance in the dependent variable that is predictable from the independent variables [58]. It provides insight into the goodness of fit of the model by indicating how well the observed outcomes match the model's predictions. The RMSE represents the square root of the average of the squared differences between the predicted values and the actual values. It measures the typical magnitude of errors in the predictions and is particularly useful in regression tasks where the prediction errors' magnitude is significant [59]. Similarly, MAE calculates the average of the absolute differences between the predicted and actual values, providing a measure of the average prediction error magnitude.

In the context of the current study, RMSE is chosen as the main evaluation metric. This choice is grounded in the fact that RMSE gives higher weight to larger errors, making it sensitive to the prediction errors that may significantly impact the quality of the freeze-dried coffee production process [59]. Minimizing RMSE aligns well with the goal of achieving accurate and precise moisture predictions in order to optimize the freeze-dried coffee production process. Furthermore, the utilization of multiple metrics, including R^2 and MAE, complements the comprehensive evaluation of the models' performance, offering a holistic view of their effectiveness.

The evaluation of machine learning models is elaborated in Section 5.3. This section assesses the performance of the trained machine learning models, along with techniques for cross-validation to ensure robustness of the results. Additionally, Section 6.2.2 and 6.4 addresses the evaluation processes concerning the Optimization Method aspect of the study. It provides insights into how the effectiveness of optimization algorithms or methods is quantitatively measured, considering the improvements achieved in the specific process being optimized. The evaluation outcomes play a pivotal role in determining the success of the developed solutions and guiding any necessary iterations or improvements in the machine learning modeling and optimization processes. If any iterations are required, the process goes back to the Business Understanding phase.

3.1.6. DEPLOYMENT

The Deployment phase within the CRISP-DM framework marks the final step in the journey of a data science project, where the developed models are put into action to provide practical solutions to the identified problems. This phase involves the integration of the developed models and solutions into the operational environment, making them accessible for end-users or stakeholders. The primary goal of the Deployment phase is to ensure that the insights and predictions derived from the data science models can be readily utilized to drive informed decision-making and achieve the project's objectives. The success of a data science project ultimately depends on the effective deployment of the developed models and solutions in real-world scenarios.

The deployment of machine learning and optimization methods in this study will be comprehensively addressed in Section 7.1 and 7.2 respectively. These sections will delve into the practical aspects of integrating the developed machine learning models and optimization methods into the freeze-dried coffee production process. It will encompass discussions on the implementation of the models within the existing operational infrastructure, including some considerations for real-time prediction and decision-making. The deployment processes aims to bridge the gap between theoretical advancements and real-world implementation, highlighting the value of data-driven insights in optimizing manufacturing operations.

Table 3.1 below maps each of the discussed phase in the CRISP-DM framework to their respective sections in this thesis.

Table 3.1: Mapping of CRISP-DM Phases to Related Thesis Sections

Phase	Related Sections
Business Understanding	1.1, 1.2, 1.3, 1.4
Data Understanding	4.1
Data Preparation	5.1, 6.1
Modeling	5.2, 6.2.1, 6.3.1
Evaluation	5.3, 6.2.2, 6.4
Deployment	7.1, 7.2

3.2. MACHINE LEARNING ALGORITHMS

This section delves into the theoretical foundations of the machine learning algorithms utilized in this study to address the research questions posed. The selection of these algorithms was based on a rigorous analysis of the Machine Learning techniques resulted from the SLR, as depicted in Figure 2.5. These algorithms have been specifically chosen for their applicability to the freeze-dried coffee production domain and their potential to predict the moisture through regression models. Each algorithm represents a distinct approach to modeling and prediction, offering unique strengths that align with the specific objectives of the study. By exploring the principles and functionalities of these algorithms, this section aims to provide a comprehensive understanding of the methods employed to develop the machine learning models within the framework of the research.

3.2.1. ORDINARY LEAST SQUARE (OLS) REGRESSION

OLS Regression is a foundational statistical technique widely used in data analysis and machine learning to establish a linear relationship between a dependent variable and one or more independent variables [9, 14, 20, 41, 60]. The core principle of OLS Regression involves finding the optimal linear equation that best represents the relationship between the dependent variable and the independent variables. Mathematically, the linear regression model in OLS Regression can be represented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

Where:

- y is the dependent variable
- β_0 is the intercept
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the independent variables $x_1, x_2, \ldots, x_n$

- ε represents the error term

Another essential objective of OLS is to find the values of coefficients $\beta_0, \beta_1, \ldots, \beta_n$ that minimize the sum of the squared residuals ε [60, 61], which can be expressed as:

$$\min_{\beta_0,\beta_1,\ldots,\beta_n} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_n x_{in}))^2$$

Where:

- n is the number of data points
- y_i is the observed value of the dependent variable for the i-th data point
- x_{ij} is the value of the j-th independent variable for the i-th data point

By minimizing the sum of squared residuals, OLS Regression finds the coefficients that best fit the data and provide a linear equation that predicts the dependent variable based on the independent variables [60, 61]. This process is often referred to as the method of *least squares*. Applying Linear Regression to the prediction of dependent variable involves fitting the model to the available training data. Once the model is trained, it can be used to predict the dependent variable for new, unseen data points. The coefficients $\beta_0, \beta_1, \ldots, \beta_n$ provide insights into the quantitative impact of each independent variable on the predicted dependent variable [62].

While OLS Regression assumes a linear relationship between variables, it provides a solid foundation for more advanced regression techniques that can capture nonlinear patterns and interactions among variables [9, 18, 63]. The subsequent subsections will explore some of these advanced techniques, building upon the principles of linear and non-linear regression models to enhance the accuracy of moisture predictions in the freeze-dried coffee production process.

3.2.2. ElasticNet Regression

ElasticNet Regression is a regularization technique that combines both L1 (Lasso) and L2 (Ridge) regularization penalties in order to address some of the limitations of these individual techniques [64, 65]. L1 regularization encourages sparsity by adding an absolute value penalty to the coefficients, effectively driving some coefficients to zero and resulting in feature selection. On the other hand, L2 regularization adds a squared value penalty to the coefficients, which helps to control multicollinearity and stabilize the model by distributing the impact of correlated features. ElasticNet aims to strike a balance between the strengths of Lasso and Ridge regularization, combining the advantages from both methods [66]. Mathematically, the ElasticNet objective function can be represented as:

$$\min_{\beta_0,\beta_1,\ldots,\beta_n} \left[\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_n x_{in}))^2 + \alpha \left(\lambda_1 \sum_{j=1}^{n} |\beta_j| + \frac{1}{2} \lambda_2 \sum_{j=1}^{n} \beta_j^2 \right) \right]$$

Where:

- n is the number of data points
- y_i is the observed value of the dependent variable for the i-th data point
- x_{ij} is the value of the j-th independent variable for the i-th data point
- $\beta_0, \beta_1, \dots, \beta_n$ are the coefficients of the independent variables
- λ_1 and λ_2 are the hyperparameters that control the strength of L1 and L2 regularization, respectively
- α is the mixing parameter that determines the balance between L1 and L2 penalties

By tuning the hyperparameters λ_1 and λ_2, as well as the mixing parameter α, ElasticNet can be customized to emphasize either L1 or L2 regularization or a combination of both. This flexibility allows ElasticNet to handle a wide range of scenarios and achieve better performance than individual regularization techniques alone [66]. It is particularly useful when dealing with datasets that have a large number of features, as it helps to prevent overfitting and select relevant features for the model [64]. Moreover, it is also useful in situations where there are many correlated features or when the number of features is much larger than the number of data points [67]. It offers a trade-off between feature selection and feature stability, making it a powerful tool for regression tasks involving high-dimensional data.

3.2.3. Support Vector Regression (SVR)

SVR is a powerful regression technique that is based on the principles of Support Vector Machines (SVM), a widely used machine learning algorithm for regression and classification problems [9, 10, 32, 48, 49]. SVR is particularly effective when dealing with non-linear relationships between variables and can handle both linear and non-linear regression tasks. The fundamental idea behind SVR is to find a hyperplane that best captures the relationship between the input variables and the target variable. Unlike traditional regression techniques that aim to minimize the error between predicted and actual values, SVR focuses on minimizing the deviation of predicted values from a specified range (margin) around the target values [68]. Mathematically, the objective of SVR can be stated as follows:

Given a training dataset with input vectors x_i and corresponding target values y_i, SVR seeks to find the coefficients w and b that define the hyperplane $wx + b$, while minimizing the cost function [32]:

$$\min_{w,b} \left(\frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \max(0, |y_i - (wx_i + b)| - \epsilon) \right)$$

Where:

- n is the number of training examples
- w represents the weights of the hyperplane

- b is the bias term
- C is the regularization parameter that controls the trade-off between minimizing the margin error and the training error
- ϵ is the margin of tolerance that defines the error around the target values

The cost function consists of two terms: the first term ($\frac{1}{2}\|w\|^2$) aims to minimize the norm of the weight vector w, while the second term enforces that the predicted values lie within a specified range around the target values [32]. Furthermore, SVR can be extended to handle non-linear relationships by using kernel functions, which implicitly map the input data into a higher-dimensional space [69]. This allows SVR to capture complex patterns and non-linearities that might not be apparent in the original feature space.

3.2.4. RANDOM FOREST REGRESSION

Random Forest Regression is an ensemble learning technique that leverages the power of decision trees to create a robust and accurate regression model. It operates by constructing multiple decision trees during training and combining their predictions to produce a final output [30, 34, 70]. The key idea behind Random Forest is to reduce overfitting by averaging the predictions of many individual decision trees. Each decision tree in the forest is trained on a random subset of the training data and is exposed to a random subset of features. This randomness introduces diversity among the trees, reducing the risk of overfitting and improving the model's generalization ability.

Mathematically, let X represent a set of input features and y denote the target variable. The Random Forest model consists of N decision trees, each trained on a different subset of the training data. The final prediction $\hat{y}$ for a given sets of input X is the average of the predictions from all N decision trees [35, 71]. Mathematically, it can be written as:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} \text{tree}_i(X)$$

Where $\text{tree}_i(X)$ is the prediction of the i-th decision tree.

The process of building each decision tree involves recursively partitioning the feature space into subsets that are as homogeneous as possible with respect to the target variable [35]. This partitioning is achieved by selecting the best feature and threshold at each node to split the data. The tree continues to grow until a stopping criterion, such as a maximum depth or minimum number of samples per leaf, is met.

Random Forest Regression offers several advantages, including robustness to overfitting, the ability to capture complex relationships, and resistance to outliers [30, 34, 35]. It is suitable for both linear and non-linear relationships and can handle high-dimensional data effectively. Additionally, the ensemble nature of Random Forest provides built-in feature selection and ranking, helping to identify the most important features in the dataset. In practice, Random Forest Regression has gained popularity due to its simplicity, versatility, and strong performance across various domains [30, 33–35, 70]. Its ability to provide reliable predictions and effectively handle a wide range of data characteristics makes it a valuable tool for regression tasks.

3.2.5. GRADIENT BOOSTING REGRESSION

Gradient Boosting Regression is a powerful ensemble learning technique that builds a predictive model by sequentially adding weak learners, typically decision trees, and combining their predictions to create a strong overall model [72, 73]. It aims to improve upon the limitations of individual weak learners by focusing on minimizing the residual errors of the previous model in each subsequent iteration. The key idea behind Gradient Boosting is to iteratively fit new models to the negative gradient of the loss function with respect to the current model's predictions. This process effectively tunes the new models to correct the errors of the previous models, leading to a model that continually improves its predictions [72–75].

Light Gradient-Boosting Machine (LightGBM) and eXtreme Gradient Boosting (XGBoost) are two of the variations of the Gradient Boosting technique that enhance its efficiency and performance [76, 77]. LightGBM introduces a histogram-based approach to split the data and gradient-based one-sided sampling to reduce the computational cost of building decision trees. XGBoost implements a regularization term and a second-order approximation to the loss function, contributing to improved stability and performance.

Mathematically, given a dataset D with input features X and target values y, the goal of Gradient Boosting Regression is to learn a model $F(x)$ that minimizes the loss function $L(y, F(x))$. In each iteration t, a new model $h_t(x)$ is added to improve the current model $F_{t-1}(x)$ [75, 76, 78]. Mathematically, this can be represented as:

$$F_t(x) = F_{t-1}(x) + h_t(x)$$

Where $h_t(x)$ is the contribution of the new model to the ensemble.

For LightGBM, the mathematical formulation of the loss function and optimization process can be defined as [76]:

$$\text{Objective}_{\text{LightGBM}} = \sum_i l(y_i, F_t(x_i)) + \sum_k \Omega(f_k)$$

Where $l(y_i, F_t(x_i))$ is the loss function at iteration t and $\Omega(f_k)$ is the regularization term for the k-th tree.

Similarly, XGBoost introduces a regularized objective function with first-order and second-order gradients [75, 76]:

$$\text{Objective}_{\text{XGBoost}} = \sum_i l(y_i, F_t(x_i)) + \sum_k \Omega(f_k) + \gamma T$$

Where T is the number of leaves in the tree, and γ controls the regularization term.

Both LightGBM and XGBoost excel in handling large datasets, high-dimensional data, and complex relationships [72, 75, 76, 79]. Their ability to balance bias and variance and their efficiency in training make them popular choices for various regression tasks.

3.3. OPTIMIZATION METHODS

The Optimization Methods specifically aims to optimize temperature control in the final heating zones after the machine learning predictions have been made. In essence, the optimization methods automatically adjust the temperature based on the moisture prediction from the machine learning algorithm, striving to reach the target moisture content efficiently and accurately. These methods have been carefully selected based on the results of the SLR, as depicted in Figure 2.5.

3.3.1. FUZZY LOGIC

Fuzzy Logic, a computational paradigm designed for reasoning under uncertainty and imprecision, differs significantly from traditional binary logic, which deals exclusively in true or false values. Fuzzy Logic introduces the notion of "fuzziness," which quantifies the degree of membership of an element within a set, expressed as values between 0 and 1 [80]. This degree of belongingness, represented by *fuzzy sets*, extends classical (crisp) sets, replacing crisp values with fuzzy values that denote the extent of an element's membership in a set. These fuzzy values are typically characterized using membership functions, which take various shapes, such as triangular, trapezoidal, or Gaussian curves, allowing them to capture diverse levels of uncertainty and imprecision [24, 81, 82]. To harness the potential of fuzzy logic in practical applications, it is essential to comprehend how it works, in particular the key concept of fuzzy sets and their role in quantifying uncertainty.

Fuzzy Logic operates on the foundation of a set of rules organized within a *fuzzy rule-based system* [80]. These rules employ linguistic variables (e.g., "high" or "low") and their corresponding membership functions to establish connections between inputs and outputs. Typically structured as "if-then" statements, these rules utilize linguistic variables in the "if" section to outline conditions, and the "then" section specifies the resulting actions or outputs. In mathematical terms, let A denote a fuzzy set defined within a universe of discourse X. The membership function of A, denoted as $\mu_A(x)$, quantifies the degree to which x belongs to A [82]. To further convert this fuzzy output into a crisp value, various defuzzification methods are available. One widely used method is the centroid method, which computes the center of gravity of the membership function curve, resulting in a single crisp value that characterizes the fuzzy output [83]. Furthermore, it is crucial to examine the broader context of its diverse manifestations in real-world applications, such as a Fuzzy Inference System (FIS).

FIS stands as a computational framework grounded in fuzzy logic, employed to model intricate relationships between input variables and output actions, facilitating decision-making amid uncertainty and imprecision [13, 26]. Diverse types of FIS systems exist, including Mamdani, Sugeno, and Tsukamoto. The Mamdani FIS relies on fuzzy sets to define rules and accommodates linguistic variables directly [84]. In contrast, the Sugeno method employs linear combinations of input variables to generate outputs, making it suitable for modeling relationships with numerical precision [85]. Meanwhile, the Tsukamoto method employs fuzzy sets to express output values, utilizing a "smoothing" technique in output generation [86]. Additionally, Fuzzy Logic Controller (FLC)s, a popular application of fuzzy logic, excel in control systems [25, 32]. FLCs utilize linguistic rules for controlling complex and nonlinear systems, enabling effective decision-making even in the presence of uncertainty and imprecision. Key advantages of Fuzzy Control Systems encompass their capacity to handle imprecise input, adapt to dynamic conditions, and offer decision-making akin to human reasoning. However, it is worth noting that these systems may present challenges related to their interpretability [10, 87].

3.3.2. ARTIFICIAL NEURAL NETWORK (ANN)

ANNs are computational models inspired by the structure and functioning of the human brain's neural networks. ANNs consist of interconnected nodes, or "neurons," organized in layers: an input layer, one or more hidden layers, and an output layer. Each connection between neurons has an associated weight that adjusts during training to learn patterns and relationships in the data. ANNs are capable of capturing both linear and complex non-linear relationships in data, making them versatile for various applications, including regression and classification tasks [35, 36, 44, 45, 88].

Mathematically, an ANN can be represented as follows:

Let X be the input vector of features, y be the output (prediction), w be the weight matrix, and b be the bias vector. f represents the activation function applied to each neuron. The input of neuron j in layer l is denoted as $z_j^{(l)}$, and its output after activation is denoted as $a_j^{(l)}$ [89].

The output of a neuron in a hidden layer or the output layer can be calculated as:

$$z_j^{(l)} = \sum_{i=1}^{n^{(l-1)}} (w_{ij}^{(l)} \cdot a_i^{(l-1)}) + b_j^{(l)}$$

$$a_j^{(l)} = f(z_j^{(l)})$$

The weights and biases are adjusted during the training process using optimization algorithms like gradient descent to minimize the prediction error. Backpropagation, a crucial step in training ANNs, involves computing the gradient of the loss function with respect to the network's weights and using it to update the weights layer by layer [90]. The backpropagation formula for updating the weights in the hidden layers is [89]:

$$\Delta w_{ij}^{(l)} = \alpha \cdot \frac{\partial L}{\partial z_j^{(l)}} \cdot a_i^{(l-1)}$$

where α is the learning rate and L is the loss function.

Training ANNs involves iterating through a fixed number of epochs [35, 91]. Each epoch consists of a forward pass (input data passes through the network, and predictions are computed), a backward pass (gradients are computed using backpropagation), and weight updates based on the computed gradients [44]. Moreover, ANNs can be customized with different architectures, activation functions, and optimization algorithms to suit specific tasks. However, they require careful hyperparameter tuning to prevent overfitting and ensure optimal performance [92].

3.3.3. Adaptive Neuro-Fuzzy Inference System (ANFIS)

Introduced by Jang in his paper [2], ANFIS is a powerful hybrid computational framework that integrates principles from both Fuzzy Logic (FL) and ANN. It combines the interpretability of Fuzzy Logic (FL) with the learning capabilities of ANN to create a versatile model capable of handling complex and non-linear relationships in data.

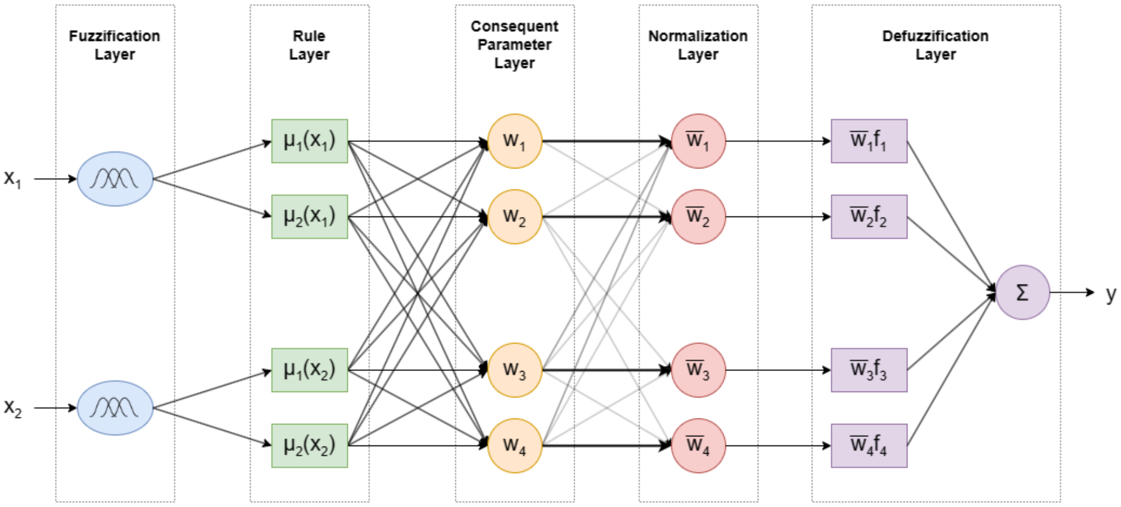


Figure 3.2: Structure of a Customized ANFIS (Adapted from [2])

Figure 3.2 depicts an example of an ANFIS structure with two inputs and one output, which is comprised of distinct layers. A typical ANFIS configuration consists of a total of five layers, each dedicated to a specific computational role [2, 31, 35, 43, 46, 49]. The **Fuzzification Layer** transforms input data into fuzzy membership degrees through membership functions, representing linguistic terms and encapsulating input uncertainty. In the **Rule Layer**, firing strengths for each rule are computed based on fuzzified inputs and rule conditions, enabling modulation of rule contributions to the ultimate output. The **Consequent Parameter Layer** determines parameters for rule consequents, often as linear combinations of input variables and firing strengths. Subsequently, the **Normalization Layer** ensures that firing strengths sum to unity. The **Defuzzification Layer** ultimately aggregates the normalized firing strengths of rule consequents, yielding the final ANFIS output. Mathematically, the ANFIS can be represented as follows [2]:

$$y = \sum_{i=1}^{R} \bar{w}_i \cdot f_i$$

where y_i is the output of rule i computed using its consequent parameters, R is the number of rules, $\bar{w}_i$ is the normalized firing strength of rule i, and f_i is the crisp function in the consequent. In addition, f_i is computed using the consequent parameters of rule i based on the Sugeno method [85]:

$$f_i = a_i \cdot x_1 + b_i \cdot x_2 + c_i$$

where x_1, x_2 represents the first and second input variable, and a_i, b_i, c_i are the parameters associated with the i-th rule's consequent part. If there are more than two input variables, the above formula should be adapted accordingly.

ANFIS employs a hybrid learning approach during training [2, 29, 35, 45]. In the forward pass, the premise parameters (related to membership functions) are modified using least squares estimation (LSE). In the backward pass, gradient descent is applied to adjust the consequent parameters. This two-phase learning process ensures that ANFIS adapts to the data efficiently and effectively captures complex relationships. In addition, ANFIS is particularly efficient when the number of inputs is limited, making it suitable for applications where interpretability is crucial and computational efficiency is desired [93]. Its ability to model complex systems using linguistic rules and neural network learning makes ANFIS a valuable tool for various engineering and scientific domains.

3.4. CHAPTER SUMMARY

This chapter introduces the methodology used to achieve the study's objectives. The CRoss-Industry Standard Process for Data Mining framework provides a structured approach, guiding the research through six major phases. Table 3.1 shows the related sections with each phase in this framework. Subsequently, the methodology involved implementing various machine learning algorithms, which were chosen based on the SLR result in Section 2.4.2 and their suitability to predict moisture values in freeze-dried coffee production. These algorithms include Linear Regression, Ordinary Least Square Regression, ElasticNet Regression, Support Vector Regression, Random Forest Regression, and eXtreme Gradient Boosting Regression. Root Mean Squared Error will be used as the main metric to evaluate the algorithms, alongside the R-squared and Mean Absolute Error. SHapley Additive exPlanations will also be applied to plot the feature importance of the machine learning models. Furthermore, optimization methods which are derived from the result of the preliminary literature study are explored, including Fuzzy Logic, Artificial Neural Network, and Adaptive Neuro-Fuzzy Inference System. These optimization methods aim to optimize and adjust temperature settings automatically in the last heating zone based on the moisture from the machine learning prediction.

4

Experiment Setup

In this chapter, the experimental setup for the research study is detailed, providing a comprehensive overview of data understanding, highlighting the sources of data and the Exploratory Data Analysis (EDA) conducted to gain insights into the dataset. The chapter also elucidates the minimal requirements for the proposed solution, both functional and non-functional, outlining the essential capabilities and qualities that it should embody.

4.1. Data Understanding

4.1.1. Data Sources

In the context of the freeze-drying process at JDE, the data flow within the system comprises a sequence of well-defined stages, each contributing to the comprehensive collection of relevant information. Figure 4.1 provides an illustrative overview of this data flow. The journey commences with the introduction of coffee granules into individual trays, followed by the stacking of trays into a complete stack. At the tray level, data is collected, capturing various attributes pertinent to the process. The frequency of data collection occurs with every tray inlet, ensuring a granular and real-time representation of the coffee granules' behavior during the initial stages.

Upon the completion of a stack, it is conveyed into the first heating zones, where it undergoes the freeze-drying process under specific temperature and time conditions, details of which are masked due to confidentiality in the recipes. Subsequent to its passage through the last heating zone, the stack's data is systematically captured at the stack level. This data collection frequency aligns with every stack outlet, providing insights into the collective behavior of the stacks as they progress through the heating zones.

Moreover, an essential aspect of the data collection process occurs during the unloading of trays onto a conveyor belt post-freeze-drying. The freeze-dried coffee is subjected to continuous moisture measurements using a moisture meter, generating data points every 0.5 seconds. This real-time monitoring contributes a granular understanding of the moisture content evolution within the freeze-drying process. The culmination of these data collection stages constructs a comprehensive dataset that reflects the intricate dynamics of the freeze-drying process, forming a critical foundation for subsequent analysis and modeling endeavors.

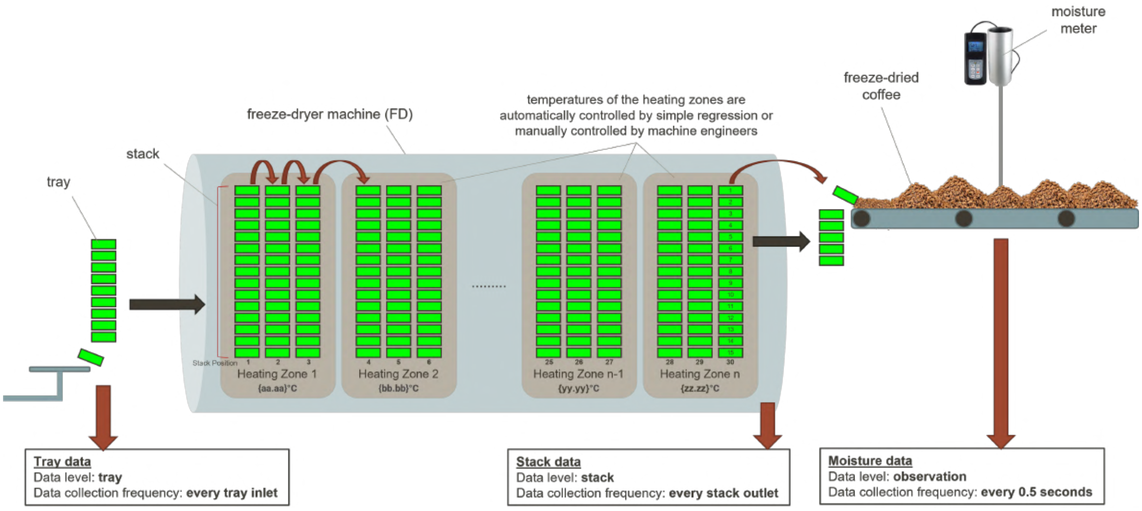


Figure 4.1: Current Freeze-Drying System Environment at JDE

The acquired data is systematically organized and stored within JDE's SQL Server infrastructure. To harness the insights from this data, a strategic workflow is employed wherein Azure Databricks is invoked to access the SQL Server database. The Databricks platform seamlessly interacts with the SQL Server, enabling the extraction of relevant data sets for further analysis and processing. Leveraging the capabilities of Python and SQL languages, Databricks facilitates data manipulation, transformation, and integration, allowing for an efficient and versatile analytical pipeline. Furthermore, the analytical exploration of the data is augmented through the integration of various open-source libraries, the details of which are elaborated in Appendix C.

4.1.2. EXPLORATORY DATA ANALYSIS (EDA)

FEATURE DESCRIPTION

EDA serves as a crucial initial step in the data mining process, enabling researchers to gain deeper insights into the characteristics of the dataset and uncover meaningful patterns. In the context of this study, EDA provides a foundational understanding of the freeze-dried coffee production data collected from four different FD machines, each processing various product IDs with distinct recipes. These recipes involve specific time and temperature settings in the heating zones of the FD machines, influencing the final product's quality. The datasets from these machines share similar features, described in Table 4.1, containing information about moisture, weight, temperature, and heating duration, among others. It is crucial to emphasize that the datasets are devoid of any missing values, thereby eliminating the necessity for imputation procedures.

Table 4.1: Features of the FD Dataset

Feature	Description
product_ID	The identifier of the coffee product
stack_ID	The identifier of the stack
moisture	Average moisture of the stack
weight	Total weight of the stack
temp_qty_n	Heating temperature at the n-th heating zone
time_qty_n	Heating duration at the n-th heating zone
stop_time_qty	Total outage duration of the stack at the heating zone

FEATURE DISTRIBUTION

A comprehensive examination of feature distributions between and within each FD machine provides valuable insights into their inherent characteristics, which are instrumental for subsequent analysis. Understanding the distribution of features is essential because it allows for a deeper grasp of the nuances within each FD machine's operation and performance. By assessing the distribution of key features, a deeper understanding of patterns, anomalies, and variations that might otherwise remain hidden can be gained. This knowledge serves as a critical foundation for the later stage of optimization and automation in the context of freeze-dried coffee production.

Figure 4.2 visually illustrates the density distribution of features across different FD machines using a Kernel Density Estimate (KDE) plot. In Figure 4.2, the x-axis represents feature values, while the y-axis signifies density, which quantifies the likelihood of observing specific values within the feature's range. The KDE plot employs two main visualization components: a continuous line and histogram bars, which serve different purposes. The continuous line represents the smooth probability density function, offering a continuous view of the data distribution. In contrast, the bars represent discrete bins or segments of the distribution, providing a more granular view of data density within specific intervals.

Upon closer examination, it becomes apparent that the moisture distributions for FD 1, 2, and 4 share a degree of similarity, indicating common moisture characteristics among these machines. Conversely, FD 3 exhibits a distinct distribution, suggesting potential disparities in operational conditions or performance. A similar trend emerges when analyzing weight distributions, as comparability is evident between FD 1 and 2, as well as between FD 3 and 4. These resemblances may be attributed to shared characteristics or processes within machine pairs. For instance, FD 1 and 2 may share similarities in their heating zones features, as do FD 3 and 4. The significance of this distribution analysis lies in its ability to highlight variations in heating temperature and duration across all FD machines. These variations underscore the necessity of accommodating these differences in subsequent modeling and analysis, as they may significantly impact the efficiency and outcome of the freeze-drying process, leading to variations in product quality. These findings suggest that the optimization and automation process must be customized for each machine based on their distinct moisture and weight distributions.

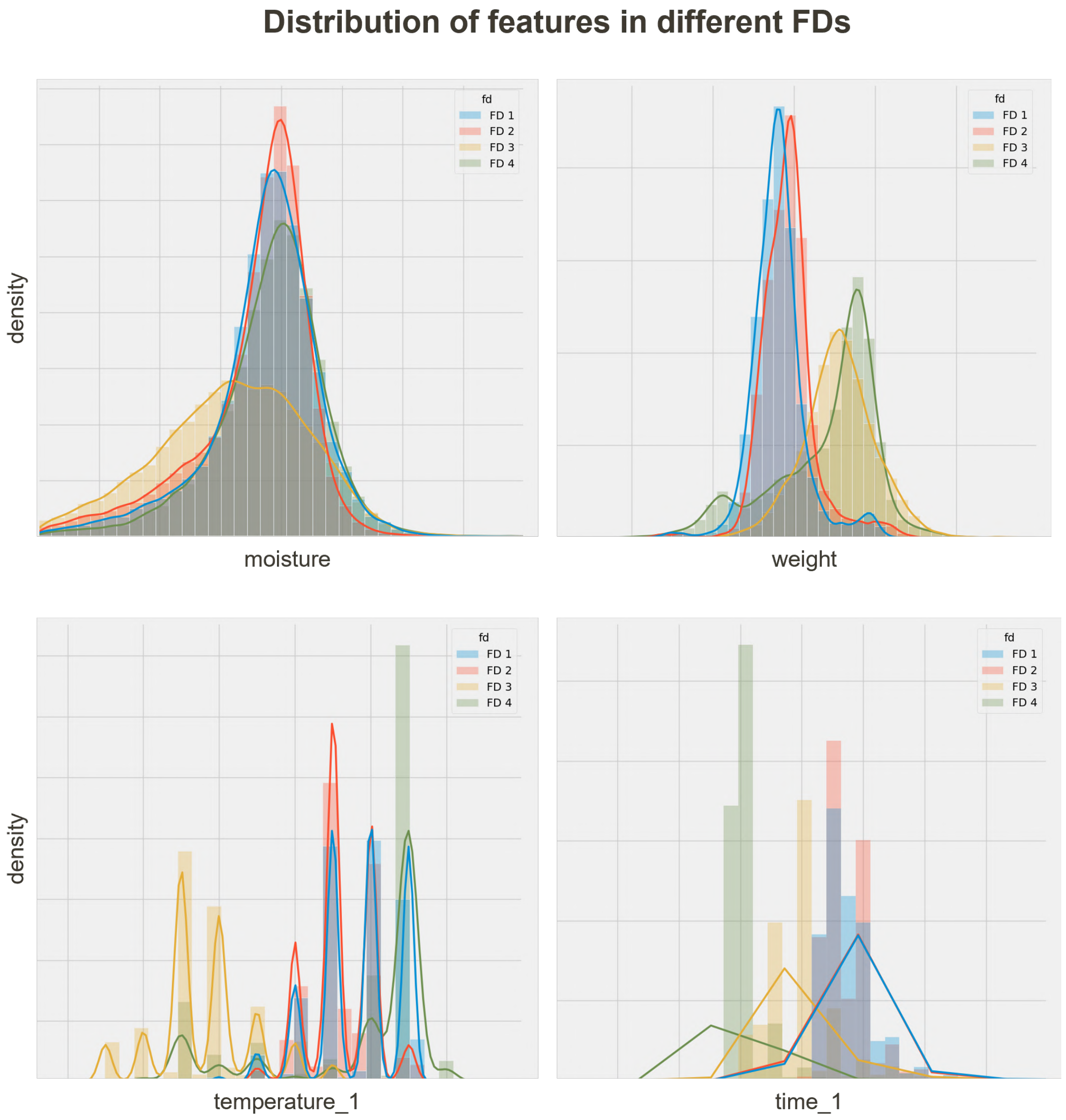


Figure 4.2: Feature Distribution in Different FD Machines

To further dissect the differences in feature distributions, it is also imperative to examine the distribution within each FD machine. For instance, Figure D.1 illustrates the distribution of features within FD Machine 1. Notably, each product processed in FD Machine 1 possesses unique feature distributions. This pattern holds true across all FD machines, as depicted in Appendix D. These variations in feature distributions within each FD machine are inherently linked to the diverse recipes, characteristics, and processes associated with individual products. Given these distinctions, it is evident that a one-size-fits-all approach to optimization and automation will not suffice. Instead, a customized optimization and automation process tailored to the specific attributes and requirements of each product within its respective FD machine is also essential for achieving optimal results in freeze-dried coffee production.

Assessing the normality of data distributions is another crucial step in statistical analysis, especially when preparing data for modeling. A normal distribution test helps determine whether a dataset follows a Gaussian distribution, which has specific statistical properties. The reason for conducting this test is to ensure that the assumptions underlying many statistical techniques, such as linear regression, are met. When data closely resembles a normal distribution, these techniques tend to perform optimally, producing reliable results [17, 94]. Various numerical tests are available to assess normality, including the Kolmogorov-Smirnov test, Anderson-Darling test, and Shapiro-Wilk test. The Shapiro-Wilk test is often regarded as the most powerful method for assessing normality, particularly for the sample sizes provided in this study [95, 96].

In the context of this analysis, the Shapiro-Wilk test was applied to the dataset, assessing the normality of features across various dimensions, including categorization by machine and product status. This determination is made based on the p-value criterion, where a p-value below the chosen significance level of 0.05 signifies non-normality. The results, as presented in the tables 4.2, indicate that none of the investigated features in each FD machine exhibit a normal distribution. In this case, all features show p-values lower than 0.05, signifying the rejection of the null hypothesis. Furthermore, Appendix E provides a detailed presentation of the Shapiro-Wilk test results for the features grouped by products in each FD machine, offering comprehensive insights into the non-normality of these features.

Table 4.2: Shapiro-Wilk Test Result of the Features in Different FD Machines

Feature	FD	p-value	Description
moisture	FD 1	Close to 0	Does not follow a normal distribution
moisture	FD 2	Close to 0	Does not follow a normal distribution
moisture	FD 3	7.35e-35	Does not follow a normal distribution
moisture	FD 4	Close to 0	Does not follow a normal distribution
weight	FD 1	Close to 0	Does not follow a normal distribution
weight	FD 2	Close to 0	Does not follow a normal distribution
weight	FD 3	8.97e-39	Does not follow a normal distribution
weight	FD 4	Close to 0	Does not follow a normal distribution
temperature_1	FD 1	Close to 0	Does not follow a normal distribution
temperature_1	FD 2	Close to 0	Does not follow a normal distribution
temperature_1	FD 3	Close to 0	Does not follow a normal distribution
temperature_1	FD 4	Close to 0	Does not follow a normal distribution
time_1	FD 1	Close to 0	Does not follow a normal distribution
time_1	FD 2	Close to 0	Does not follow a normal distribution
time_1	FD 3	Close to 0	Does not follow a normal distribution
time_1	FD 4	Close to 0	Does not follow a normal distribution

The absence of a normal distribution in the features can impact the subsequent modeling phase. Many statistical models, especially those of a parametric nature, assume that data follows a normal distribution [17, 94]. When this assumption is violated, it can potentially compromise the model's performance, leading to less precise predictions. However, it is worth noting that while certain features can be transformed to approximate a normal distribution, this transformation may have little impact on improving normality, especially in the case of linear regression [94, 97]. In such instances, it is advisable to explore alternative modeling approaches that are better suited for handling non-normally distributed features, such as tree-based models [98, 99]. These models can offer robust performance even when dealing with data that does not adhere to the normal distribution assumption.

OUTAGES

The presence of outliers is a matter of paramount concern. Outliers may signify anomalies or irregular occurrences within machine operation, prompting a deeper investigation to identify unexpected patterns or deviations from standard operating conditions. Understanding and addressing these outliers are essential steps in ensuring the reliability and quality of the freeze-drying process for coffee production. In the context of freeze-dried coffee production, the occurrence of outages in FD machines can be seen as outliers, signifying unexpected interruptions in the manufacturing process. These interruptions can result from various factors, including power interruptions, manual stops initiated for maintenance or adjustments, or mechanical failures. The identification of outages relies on a specific feature known as *stop_time_qty*, which records the duration of such interruptions. Due to confidentiality constraints, the threshold value for identifying outages cannot be disclosed.

Examining the impact of outages on feature distributions is a crucial aspect of this analysis. Figure D.5 compares the feature distributions during outages and non-outage conditions. Notably, as expected, the visualizations reveal that the distribution of features during outage conditions exhibits greater uncertainty and variability compared to non-outage conditions. This heightened variability can be attributed to the disruptive nature of outages, which introduce irregularities and deviations from standard operational patterns. These differences underscore the significance of comprehending varying feature distributions in different operational states, as they necessitate tailored strategies. This further reinforces the need for a customized optimization and automation process tailored to handle outage-specific scenarios.

In addition to identifying outages, data quality is rigorously maintained through the application of two essential filters to the dataset. Firstly, the *weight* feature must have a value greater than 0 to ensure that only valid data related to product weight is considered. Secondly, the *moisture* feature is filtered to exclude values below 1, as validated by the factory engineers. A moisture level below 1 indicates that the product will undergo reprocessing. These filters are integral to preserving the integrity and reliability of the dataset, ensuring that it accurately represents the freeze-dried coffee production process. Moreovew, it is worth noting that the volume of data retained after applying these filters aligns with the quantities observed in related studies [12, 30]. This assurance regarding the dataset's size reinforces its suitability for analysis. Specific details, such as the exact number of columns, are omitted for confidentiality reasons. Appendix F provides further insights into the rows of datasets after applying the filters.

FEATURE CORRELATION

Assessing feature correlation is a fundamental step in comprehending the intricate relationships within the dataset. Prior studies have consistently emphasized the importance of examining feature correlation before proceeding with modeling [97, 98, 100, 101]. Understanding these dependencies is essential before embarking on any modeling process, as it aids in identifying which features may influence each other and to what extent. Additionally, strong correlations between two features should be carefully considered. Strong correlations can lead to multicollinearity, a situation where two or more features in a model are highly correlated, making it challenging to distinguish their individual effects. In cases of multicollinearity, it becomes difficult to assess the precise contribution of each feature to the model, potentially leading to less reliable and interpretable results [101, 102]. Therefore, the selection of features that are strongly correlated should be done thoughtfully to avoid multicollinearity and ensure the robustness of the subsequent modeling process.

One of the commonly employed methods for assessing feature correlation is the Pearson correlation coefficient [99], which ranges from -1 to 1 and serves as a quantitative measure of these relationships. A coefficient of -1 signifies a perfect negative correlation, indicating that as one feature increases, the other decreases. Conversely, a coefficient of 1 denotes a perfect positive correlation, indicating that both features tend to increase or decrease together. A coefficient of 0 implies no linear correlation, signifying that changes in one feature do not predict changes in the other.

In the correlation matrix depicted in Figure 4.3, each cell represents the correlation between two specific features. The color intensity in each cell conveys the strength of the correlation, with darker blue indicating stronger positive correlations and darker red representing stronger negative correlations. As observed in the matrix, most of the correlations tend to be close to 0, indicating weak linear relationships between features. This observation is reassuring, as it suggests that the features considered in this study exhibit relatively independent behavior. Furthermore, analyzing the specific correlations between features can provide valuable insights into potential relationships within the freeze-drying process. For instance, a negative correlation between moisture levels and heating duration might suggest that longer heating times result in lower moisture content in the final product. Conversely, a positive correlation between weight and moisture levels may indicate that heavier product batches tend to have higher moisture content. These insights can guide the subsequent modeling process by helping to engineer relevant features based on their relationships.

4.2. MINIMAL REQUIREMENTS

This section presents the functional and non-functional requirements that should be considered for the successful implementation of the proposed solution. These requirements are defined based on related studies [13, 21], which implemented fuzzy-based systems similar to the approach considered in this research. Additionally, certain requirements are derived from the

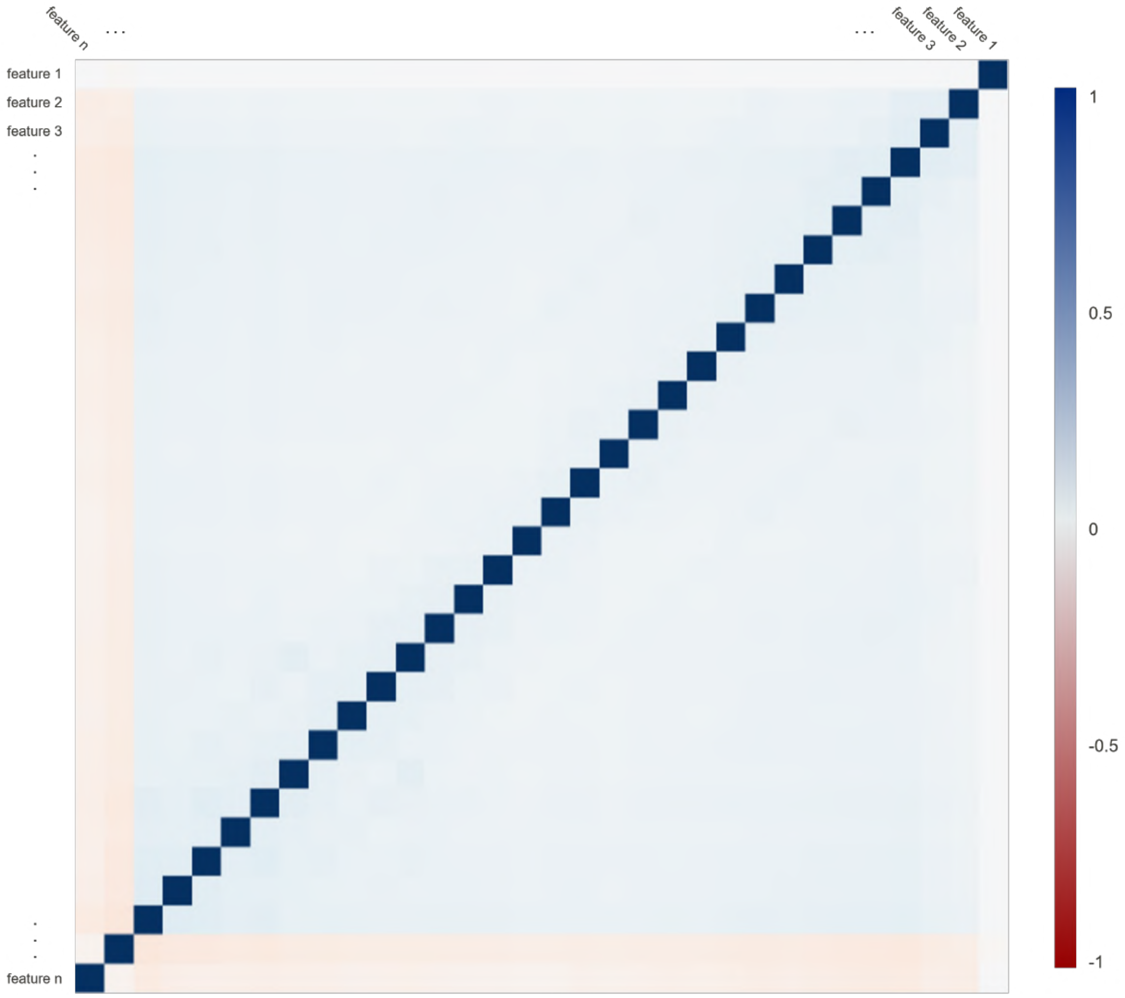


Figure 4.3: Feature Correlation in FD Dataset

specific needs and constraints of JDE, providing a comprehensive foundation for the development and implementation of the proposed solution.

4.2.1. FUNCTIONAL REQUIREMENTS

The functional requirements outline the specific capabilities and features that the proposed solution should possess. These requirements are essential for achieving the desired functionality and performance of the system. The following functional requirements are identified for the implementation of the proposed solution:

- **FR1: Data Collection and Pre-processing**
  The solution should be capable of collecting relevant data from various sources, such as sensors and databases. It should also include pre-processing mechanisms to clean, transform, and format the collected data for further analysis and model training.

- **FR2: Machine Learning Model Training**
  The solution should incorporate machine learning algorithms to train predictive mod-

els based on the collected data. It should support various machine learning techniques depending on the specific application requirements.

- **FR3: Model Evaluation and Validation**
  The solution should provide mechanisms for evaluating and validating the trained models. It should include performance metrics to assess the relevant indicators of the models' predictive capabilities.

- **FR4: Real-time Optimization and Decision-making**
  The solution should be capable of performing real-time optimization based on the trained models and optimization methods. It should be capable of handling large volumes of data and processing them within acceptable time limits to provide timely insights and recommendations.

- **FR5: Interoperability and Automation:** The solution should support interoperability with existing infrastructure and systems, such as databases, APIs, or visualization tools. It should facilitate seamless exchange of data and information to ensure smooth operation and integration with other components of the ecosystem. Additionally, the solution should exhibit automation capabilities to reduce manual intervention, enable efficient data processing, and streamline workflows.

4.2.2. NON-FUNCTIONAL REQUIREMENTS

In addition to the functional requirements, the non-functional requirements define the quality attributes and constraints that the proposed solution should adhere to. These requirements are essential for ensuring the system's performance, reliability, and usability. The following non-functional requirements are identified for the implementation of the proposed solution:

- **NFR1: Reliability and Robustness**
  The solution should be reliable, ensuring consistent and accurate predictions even in the presence of outliers or noisy data. It should also be robust, capable of handling unexpected scenarios or errors gracefully, and recovering from failures effectively.

- **NFR2: Scalability and Replicability**
  The solution should be scalable, allowing for easy replication and deployment across multiple machines in different factories or manufacturing environments.

- **NFR3: Interpretability and Usability:** The solution should ensure that the outputs generated by machine learning models and optimization methods are understandable to end-users. It should provide clear explanations, visualizations, and insights that enable effective decision-making and facilitate user acceptance and trust in the solution.

4.3. CHAPTER SUMMARY

This chapter sets the stage for the experimental setup of the research study, outlining data sources, the proposed solution, and essential requirements. The data originates from tray, stack, and moisture records of FD machines and are stored in JDE's SQL Server. With four FD machines handling distinct products, similar features in their datasets are present. Analyzing these datasets reveals varying feature distributions among products across machines, especially during outages. Additionally, based on the investigation of feature correlations, no strong correlation between features are found.

Furthermore, a general overview of the proposed solution is presented, envisioning the use of real-time predictions and adaptive optimization method to redefine freeze-dried coffee production. This involves crafting tailored machine learning models for each FD machine, product, and outage event, enhancing the solution's adaptability. The chapter closes by outlining some essential prerequisites for the solution's success. These include functional aspects such as data collection, model training, real-time optimization, and interoperability. Non-functional aspects emphasize performance, reliability, scalability, and usability. These requirements lay the foundation for the next phase: building predictive models to develop the proposed solution.

5

Machine Learning Algorithms

This chapter delves into the realm of machine learning algorithms for predicting moisture levels in the freeze-dried coffee production process. It starts by explaining the steps taken in data preparation, including feature engineering and scaling techniques. The subsequent section focuses on model building, discussing the steps that was applied along the process. Evaluation takes precedence, where model performance is assessed through optimal parameter analysis, metrics and model comparisons across different datasets. Additionally, the chapter explores Automated Machine Learning (AutoML) using Databricks[1] and presents the outcomes of the fine-tuned models.

5.1. Data Preparation

In line with the findings from Section 4.1.2, the datasets are partitioned into subsets corresponding to different FD machines, products, and outage conditions, as detailed in Appendix F. Within each of these datasets, features that bear no meaningful contribution to the machine learning model's predictive ability are eliminated. Features such as *stack_id* and *product_id*, which do not significantly inform the moisture prediction process, are removed to enhance the model's focus on relevant information. Furthermore, the multiplication of time and temperature values from each position within the heating zone culminates in the creation of a new feature, termed *heat*. These parameters jointly dictate the chemical composition of the roasted coffee [103], thus synthesizing them into a single feature streamlines their impact. Figure 5.1 provides a visual representation of the distribution of the *heat* feature across various FD machines, products, and outages.

In line with the distribution patterns observed in other features, it becomes apparent that the means and standard deviations of the *heat* feature vary significantly across the distinct sub-

[1] https://www.databricks.com/product/automl

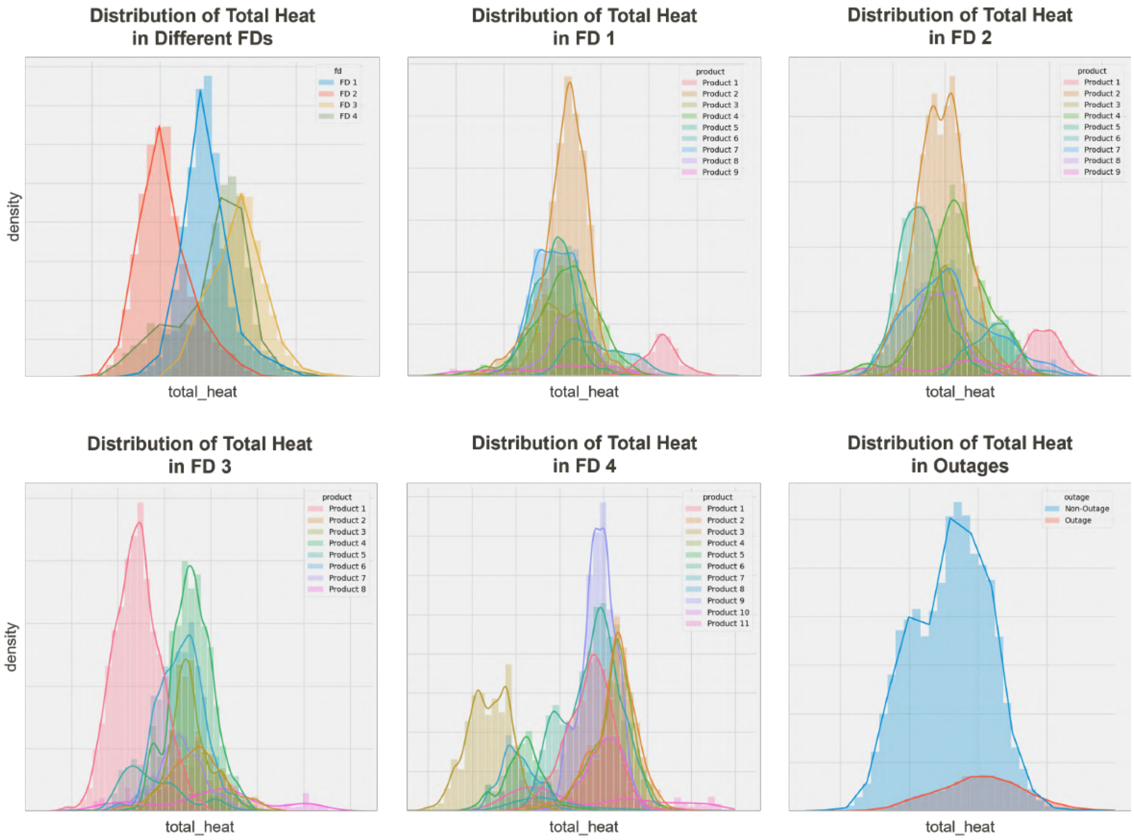


Figure 5.1: Heat Feature Distribution

sets. This observation further supports the justification that different products require their customized models, considering the unique characteristics reflected in the *heat* feature distribution. To mitigate the risk of multicollinearity resulting from the introduction of the new *heat* feature, the original time and temperature features are deliberately excluded from each dataset. Further steps involve splitting the dataset into training and testing sets, assigning 80% of the data for training and reserving 20% for testing. Concurrently, feature scaling is employed using the `StandardScaler`[2] technique to normalize the independent variables, aligning them within a standardized range for optimal model convergence during training. With the dataset prepared and the features engineered, it is important to reiterate that *moisture* serves as the dependent variable in this study. The independent variables include *weight* and the *heat_n* feature, the latter representing distinct stack positions within the heating zone. These features are harnessed by the regression algorithms to predict moisture levels in freeze-dried coffee.

5.2. MODELING

A comprehensive array of regression models are developed and scrutinized to effectively predict moisture levels in freeze-dried coffee. The primary goal is to select a model that demonstrates robust generalization capabilities while maintaining high predictive accuracy. According-

[2] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

ing to the analysis of the literature review, particularly in Section 2.4.2, six distinct regression algorithms are employed in this endeavor: Linear Regression (LinReg), Ordinary Least Square (OLS) Regression, ElasticNet Regression, Support Vector Regression (SVR), Random Forest (RF) Regression, and eXtreme Gradient Boosting (XGBoost) Regression. The time series and clustering methodologies, while valuable in their own right, are not aligned with the regression problem as the focus of this study. It is also pertinent to mention that FL, ANFIS, and ANN serve as optimization methods, a facet of this research to be comprehensively addressed in Chapter 6.

The key motivation behind employing diverse regression algorithms lies in harnessing the strengths of each approach. (OLS) serves as a fundamental baseline, providing a straightforward linear regression framework. ElasticNet Regression introduces regularization to handle multicollinearity, thereby preventing overfitting and enhancing the model's stability. SVR accommodates non-linear relationships that may exist between the predictors and the target variable, making it suitable for capturing complex patterns in the data. RF Regression exploits ensemble learning by combining the predictions of multiple decision trees, resulting in improved predictive accuracy and resilience to overfitting. Furthermore, XGBoost Regression leverages gradient boosting, a technique that sequentially builds models to correct errors made by preceding models, ultimately refining the model's predictions.

To commence this phase, each model is instantiated with its corresponding algorithm and an initial set of hyperparameters. The initial set of hyperparameters for each model, drawn from insights in similar studies [30, 73], can be found in Appendix G. The models are then subjected to a meticulous process of hyperparameter tuning through `GridSearch`[3]. In this initial phase of modeling, certain hyperparameter values are manually determined based on domain knowledge and prior research. This proactive approach is aimed at providing a starting point for the optimization process, ensuring that the models are not trapped in suboptimal parameter configurations. The `GridSearch` technique systematically explores a defined hyperparameter space, evaluating the performance of the model for each combination of hyperparameters. For instance, SVR seeks the ideal combination of *kernel*, *C*, and *epsilon*, while RF Regression undergoes a `GridSearch` to determine the optimal values for *n_estimators*, *max_depth*, *min_samples_split* and *max_features*.

Having established the current best initial hyperparameters through an iterative process, the models are then constructed and subjected to 5-fold cross-validation. This technique partitions the dataset into five distinct subsets, with one subset serving as the validation set and the remaining subsets as training sets for each fold. This procedure is executed iteratively to ensure that each subset is employed as a validation set once. The outcomes of these iterations are aggregated to discern the model with the optimal predictive performance.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

5.3. EVALUATION

The evaluation process involves the calculation of pertinent metrics, including the average RMSE, MAE, and R^2 across all folds. Figure 5.2 shows a scatter plot example of Product 7 in FD machine 2, depicting the alignment between the actual moisture levels and the moisture levels predicted by the machine learning models. Each point on the scatter plot represents a data point from the dataset, where the x-axis denotes the actual moisture levels observed in the freeze-dried coffee samples, and the y-axis represents the moisture levels predicted by the machine learning models. Additionally, each fold of the 5-fold cross-validation is distinguished by a unique color, allowing for a comparative assessment of the model's predictions across different subsets of the data. The dashed identity line represents the theoretical perfect prediction alignment, aiding in the interpretation of the scatter plot patterns.

The OLS Regression model in Figure 5.2 presents a scatter plot where the predictions tend to diverge from the identity line. The R^2 value of 0.29 implies that approximately 29% of the variance in actual moisture levels is explained by the independent variables used in the model. However, the relatively high average RMSE of 0.37 and MAE of 0.27 indicate a considerable level of prediction errors. The scatter pattern suggests that the model struggles to capture the intricate variations in moisture levels, leading to significant deviations from the actual values. Despite its simplicity and interpretability, the OLS Regression model is limited by its linear nature. It assumes a linear relationship between predictors and the target variable, which might not adequately capture the complex interactions and non-linearities present in the data. This deficiency in capturing non-linear patterns might be contributing to its relatively high prediction errors and suboptimal R^2 value compared to more advanced models.

ElasticNet Regression, characterized by regularization, offers a scatter plot in Figure 5.2 that exhibits a similar pattern to the OLS Regression model. The R^2 value of 0.30 indicates a marginal improvement in capturing the variance in actual moisture levels. However, the average RMSE of 0.36 and MAE of 0.27 remain relatively high. Despite regularization, the model appears to face challenges in achieving a tighter alignment between predicted and actual moisture levels, suggesting that linear relationships might not adequately capture the complex patterns in the data. ElasticNet Regression combines both Lasso and Ridge regularization, attempting to address multicollinearity and model overfitting. While the regularization techniques provide better generalization compared to OLS, the model might still be constrained by its linear assumptions. The trade-off between Lasso and Ridge regularization parameters could be hindering its ability to accurately capture the underlying data patterns, leading to limited performance improvements compared to other models.

The SVR model in Figure 5.2 showcases a scatter plot with data points clustered closer around the identity line, compared to the above two regressors. The considerably higher R^2 value of 0.42 signifies that the model captures a substantial portion of the variance in actual moisture levels. The relatively low average RMSE of 0.33 and MAE of 0.23 indicate a commendable predictive performance. The scatter pattern showcases a strong alignment, suggesting that the SVR

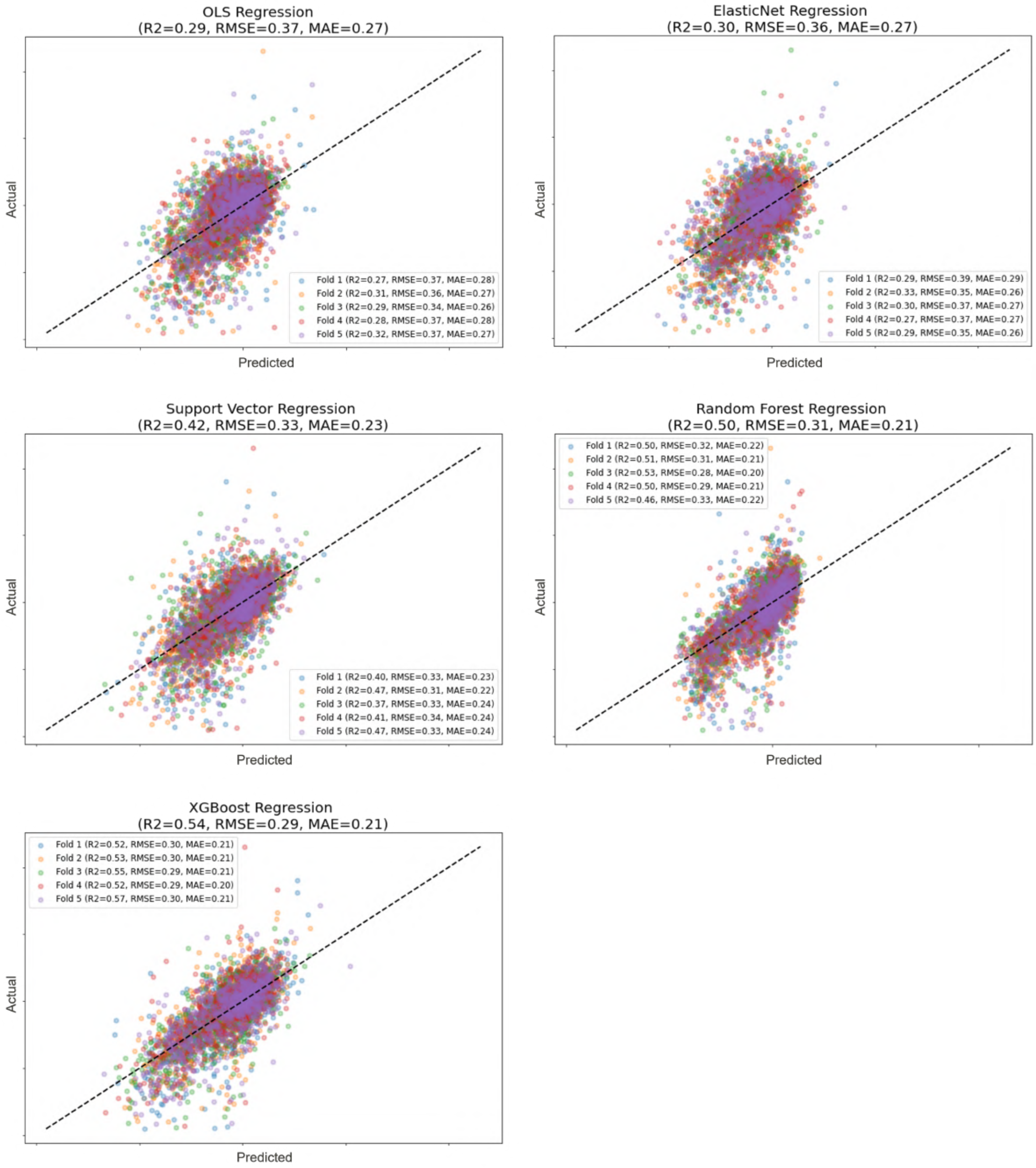


Figure 5.2: Scatter Plot of Actual vs Machine Learning Predicted Moisture in FD 2, Product 7

model effectively captures underlying trends and relationships in the data. SVR excels in capturing complex non-linear relationships by mapping the data into higher-dimensional spaces using kernel functions. This ability allows the model to capture intricate patterns that might be missed by linear models. The clustering of data points close to the identity line suggests that SVR's flexibility in fitting non-linear relationships contributes to its higher R^2 value and superior predictive performance compared to simpler linear models.

The RF Regression model in Figure 5.2 delivers a scatter plot characterized by data points closely aligned with the identity line. With an moderate R^2 value of 0.50, the model explains a significant portion of the variance in actual moisture levels. The low average RMSE of 0.31 and MAE of 0.21 underscore the model's accurate predictive capabilities. The consistent and tight clustering of data points around the identity line signifies the model's capacity to generalize well and capture complex moisture level patterns. RF Regression leverages ensemble learning, aggregating predictions from multiple decision trees. This ensemble approach reduces overfitting and enhances predictive performance. The model's ability to capture complex interactions and non-linearities, combined with its ensemble nature, contributes to its high R^2 value and superior predictive accuracy compared to simpler linear models, and even SVR in this case.

XGBoost Regression in Figure 5.2 stands out with a scatter plot that reveals data points closely aligned with the identity line. The higher R^2 value of 0.54 confirms the model's adeptness in capturing a substantial portion of the variance in actual moisture levels compared to the other models in this product. Moreover, the low average RMSE of 0.29 and MAE of 0.21 denote the model's precision in predicting moisture levels. The tight clustering of data points around the identity line across all folds demonstrates the robustness and generalization capabilities of the XGBoost model. XGBoost Regression utilizes gradient boosting to sequentially improve model predictions. It corrects errors made by preceding models, allowing for iterative refinement. This process enhances the model's ability to capture complex relationships in the data. The XGBoost model's flexibility, iterative nature, and ensemble approach contribute to its high R^2 value and predictive accuracy compared to other models.

The iterative evaluation process is applied to each product across all FD machines, ensuring a comprehensive analysis of the models' performance across different datasets. Appendix H shows the complete machine learning model comparison for all of the different scenarios. For every product, the model's predictive metrics are meticulously recorded and compared, enabling a holistic understanding of their effectiveness. Appendix I provides the comprehensive summary of these metrics. The performance evaluation is further distilled through visual aids for ease of comparison and interpretation. In the appended bar chart, shown in Figure 5.3, a condensed representation of the models' performance is presented. The chart succinctly outlines the number of occasions each machine learning algorithm secures the best metrics in terms of R^2, RMSE, and MAE.

In Figure 5.3, the machine learning algorithms are listed on the x-axis, while the counts of instances in which each algorithm achieved the best metrics for R^2, RMSE, and MAE are presented on the y-axis. Notably, XGBoost Regression consistently shines across all three metrics, securing the highest counts for each. It demonstrates the best performance among all the machine learning algorithms, achieving the best R^2, RMSE, and MAE results in 30 (73%), 33 (80%), and 33 (80%) instances respectively. RF Regression and SVR also exhibit some competitive performances, outperforming the other linear regression algorithms across various products and machines.

The prominence of non-linear models, such as XGBoost Regression and RF Regression, in achieving superior predictive performance indicates that the relationship between independent and

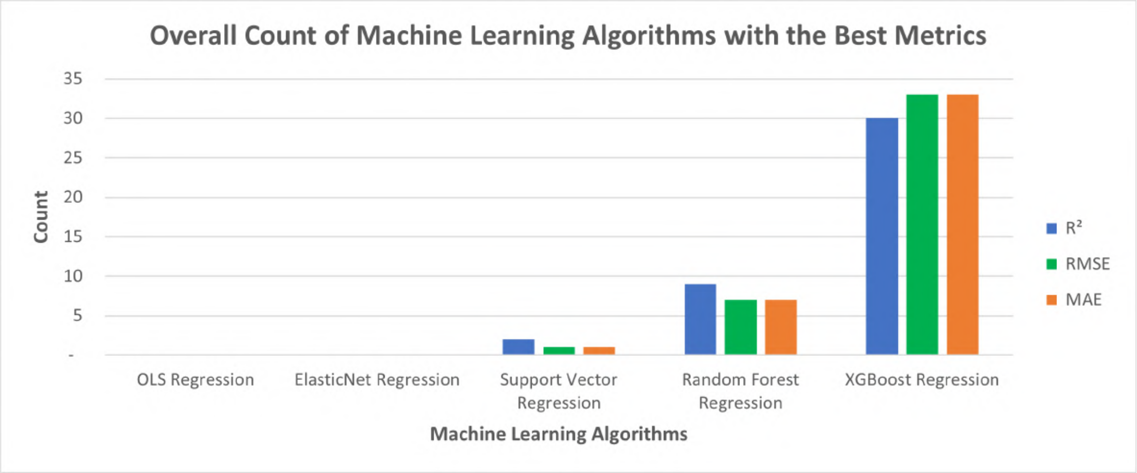


Figure 5.3: Machine Learning Algorithm Metrics Summary

dependent variables is likely not strictly linear. The consistently strong performance of these models implies that they are better equipped to capture the complex and non-linear interactions present in the data. Their ability to handle non-linear relationships and interactions between variables is a key advantage over linear models like OLS and ElasticNet Regression, which assume linear relationships. The prevalence of XGBoost and RF Regression in achieving the best metrics reinforces the notion that these models excel in capturing complex patterns that go beyond simple linear correlations.

However, the success of non-linear models does not necessarily mean that the entire relationship between predictors and the target variable is exclusively non-linear. The presence of non-linear models that perform well highlights the complexity of the data and the potential for both linear and non-linear relationships to coexist. The performance differences among the models also underscore the importance of employing a diverse range of algorithms, as different models are capable of capturing different aspects of the underlying relationships in the data. The success of models like XGBoost Regression in accurately predicting moisture levels reinforces the need to consider and accommodate non-linear patterns when developing predictive models for this application.

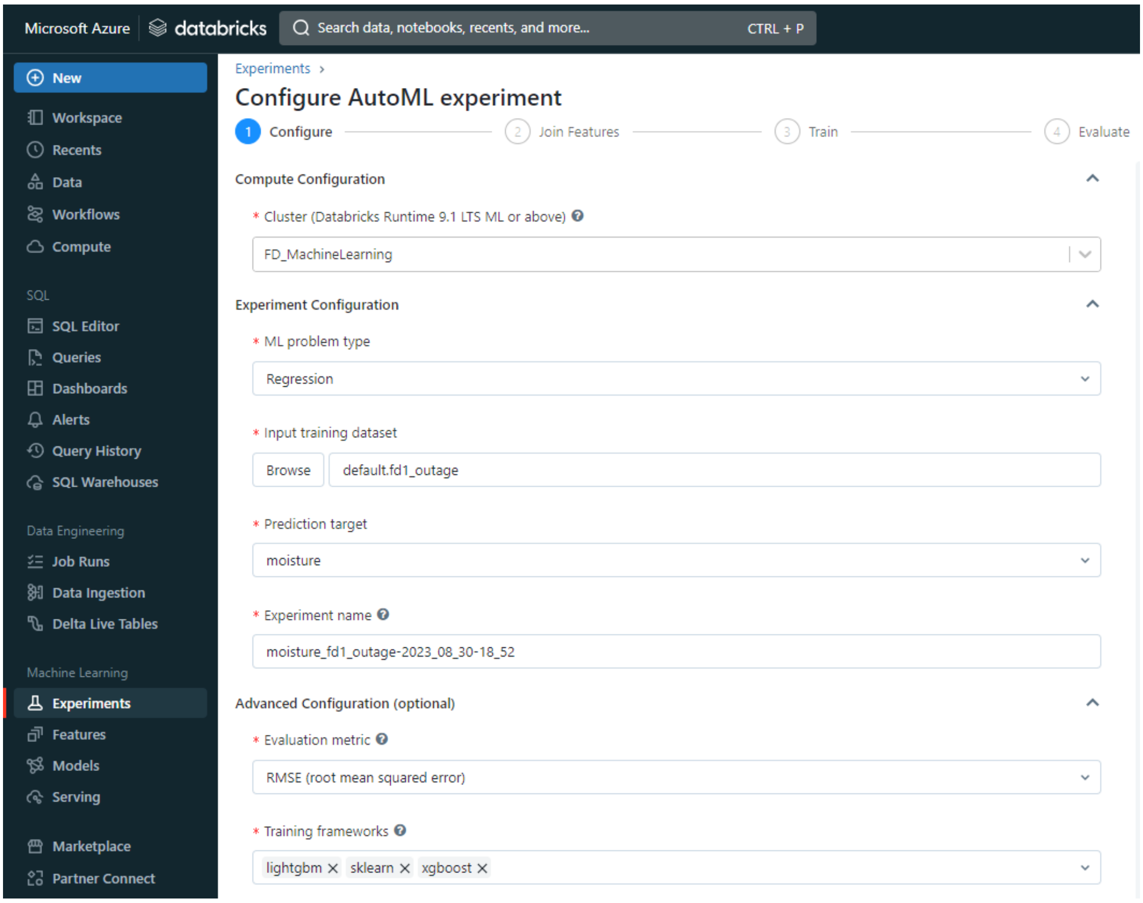
5.4. AUTOMATED MACHINE LEARNING (AUTOML)

One of the features in Databricks is AutoML[4], which is an automated machine learning solution that aims to simplify the process of developing, training, and deploying machine learning models. It automates various tasks such as hyperparameter tuning, model selection, and deployment. In this study, AutoML is employed to efficiently search for the most optimal hyperparameters for each of the aforementioned machine learning models. It plays a crucial role in this research by conducting extensive experiments to find the best-performing configurations. Additionally, AutoML offers the ability to document and reproduce experiments, facilitating the

[4] https://www.databricks.com/product/automl

deployment of models at a later stage.

Building on the previous section, where the XGBoost Regression emerged as the best overall model, it is chosen as one of the models in the training framework. Figure 5.4 shows the standardized experiment setup used for training machine learning models. While different models employ the same setup, each uses a distinct training dataset. This consistent experimental framework ensures fairness and comparability across various models, allowing accurate assessment and comparison of their performance.



Configure AutoML experiment

1 Configure — 2 Join Features — 3 Train — 4 Evaluate

Compute Configuration

* Cluster (Databricks Runtime 9.1 LTS ML or above): FD_MachineLearning

Experiment Configuration

* ML problem type: Regression

* Input training dataset: Browse default.fd1_outage

* Prediction target: moisture

* Experiment name: moisture_fd1_outage-2023_08_30-18_52

Advanced Configuration (optional)

* Evaluation metric: RMSE (root mean squared error)

* Training frameworks: lightgbm × sklearn × xgboost ×

Figure 5.4: AutoML Experiment Setup

An AutoML experiment consists of a sequence of iterative model training runs, each systematically exploring a range of hyperparameter combinations. This iterative process continues until it attains a state of convergence, marked by the fulfillment of predefined stopping criteria. By monitoring the performance metrics, the AutoML platform identifies the most successful run based on the designated evaluation criteria, which, in this study, is the RMSE. The source code of the optimal run can then be further examined and replicated, offering a valuable insights into the most optimal hyperparameter configurations. Table 5.1 presents the optimal hyperparameters of the XGBoost Regression for Product 7 in FD 2. The adoption of these optimized hyperparameters brings about a notable enhancement in model performance, as verified by the comprehensive metric comparisons in Table 5.2. Notably, these improvements span across various aspects, such as an increased R^2 score from 0.54 to 0.63, a marked reduction in RMSE from 0.29 to 0.23, and a substantial drop in MAE from 0.21 to 0.17. Further metric comparisons scores of pre- and post-AutoML in the other FD machines can be found in Appendix J.

Table 5.1: Most optimum hyperparameters for FD 2, Product 7

Hyperparameter	Explanation	Value
colsample_bytree	Proportion of features used in each tree	0.7163
learning_rate	Step size at each iteration	0.0137
max_depth	Maximum depth of the decision tree	10
min_child_weight	Minimum sum of instance weight in a child	3
n_estimators	Number of boosting rounds	1604
n_jobs	Number of parallel threads used	100
subsample	Proportion of training data used	0.6176
verbosity	Level of verbosity	0
random_state	Seed for random number generation	566088010

Table 5.2: Metrics Improvement Post-AutoML for FD 2, Product 7

Metrics	Pre-AutoML	Post-AutoML
R^2	0.54	0.63
RMSE	0.29	0.23
MAE	0.21	0.17

While there are no specific studies predicting moisture levels in freeze-dried coffee, similar studies have applied analogous methodologies, as demonstrated in [9, 34, 104]. These studies involve the prediction of previously unexplored dependent features by leveraging domain knowledge and established machine learning algorithms. For instance, the studies have forecasted fishing yield [9], predicted body temperature using smart pillows [34], and developed data prediction models for Wireless Sensor Networks [104]. However, making direct numerical comparisons with these studies is challenging due to disparities in context, feature sets, and various other factors.

Moreover, the feature importance in the XGBoost Regression model is verified through SHAP plots. Figure 5.5 provides an example of a SHAP plot for the XGBoost Regression Model applied to Product 7 in FD 2. The plot reveals that the *weight* and *heat* values of the last six positions located in the last two heating zones have the most significant impact on the model's predictions. Generally, higher *heat* values for these last positions are associated with lower moisture levels, while higher *weight* values correspond to higher moisture levels. Appendix K shows that other products also generated similar SHAP plots, consistently highlighting the importance of the *weight* and *heat* values of the last two heating zones in contributing to the model's predictive accuracy.

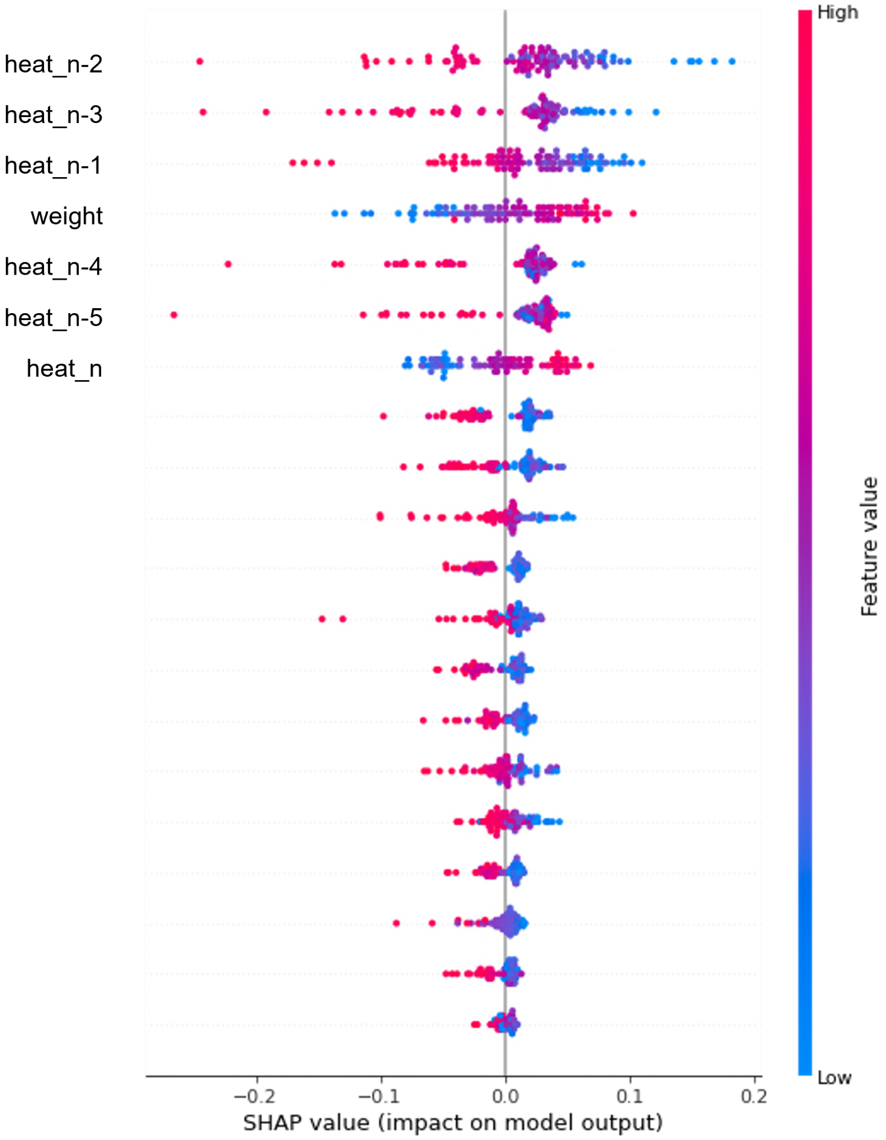


Figure 5.5: SHAP Plot of FD 2, Product 7 XGBoost Regression Model

5.5. CHAPTER SUMMARY

In this chapter, the focus shifts to the exploration of various machine learning algorithms for predicting moisture levels in freeze-dried coffee production. The journey begins by meticulously preparing the data in Section 5.1, which involves feature engineering and scaling techniques. The creation of the composite *heat* feature emerges as a pivotal step, consolidating time and temperature variables to better capture the chemical processes within freeze-drying. The subsequent Section 5.2 delves into the process of model building, where a comprehensive set of regression algorithms is evaluated. This encompasses linear models such as OLS Regression and ElasticNet Regression, as well as non-linear models including SVR, RF Regression, and XGBoost Regression. The observation of non-linear models outperforming linear ones highlights the likely presence of complex, non-linear relationships between variables.

Section 5.4 introduces the concept of AutoML, an automated machine learning solution with the capacity to optimize hyperparameters, select models, and facilitate deployment. AutoML emerges as a pivotal tool in refining model performance and experiment documentation. This section showcases how AutoML synergizes with the XGBoost Regression model, enhancing its predictive capabilities. The optimal hyperparameters derived from AutoML yield substantial performance improvements, as demonstrated in Table 5.1 and Table 5.2. Furthermore, the validation of feature importance through SHAP plots reinforces the critical role of *weight* and *heat* features in driving predictive accuracy.

6

Optimization Methods

This chapter delves into approaches that enable automated temperature adjustments based on the predicted moisture from the previous chapter. It begins with preparing the dataset, detailing feature engineering steps to transform raw data into meaningful inputs. The subsequent section explores Fuzzy Logic and delves into Adaptive Neuro-Fuzzy Inference System (ANFIS), elucidating modeling and optimization steps. Then, the performances of both the optimization method will be evaluated in the next section. Lastly, Sub-RQ6 will be answered, shedding a light in automatic temperature adaptation, revealing how the developed optimization method can be an optimal solution to the problem context of this study.

6.1. Data Preparation

In the pursuit of enhancing moisture prediction accuracy, the initial steps involve data preparation tailored for the optimization methods. An additional column named *prediction* is introduced into the existing FD dataset stored in the SQL server. This column is sourced from the moisture prediction results obtained from the XGBoost model, as elaborated in the previous chapter. Importantly, the introduction of this column does not interfere with the normal operation of the FD machine. This evaluation takes place without imposing any disruptions to the ongoing freeze-drying process. Instead, it acts as a valuable reference, facilitating the comparison of predicted moisture levels with the predefined target moisture levels established for each product recipe. The evaluation of optimization method effectiveness becomes feasible through the assessment of real-world moisture data by leveraging this *prediction* column.

As the accumulation of data progresses, the focus shifts towards three pivotal features that underpin the optimization process. Among these features, two newcomers are introduced to strengthen the optimization methods. The first feature, denoted as *prediction_error*, is computed by subtracting the *prediction* column from the target *moisture* values. The target moisture values are established in accordance with the unique recipe of each product. The *predic-*

tion_error thus offers insights into the disparities between predicted moisture and the predefined target moisture levels. This understanding proves crucial for the temperature adjustment process based on the prediction results. The second feature, *heat_sum*, is a summation of *heat* values up to the last two heating zones, further added by the actual *heat* of the remaining positions. Together with the existing *weight* attribute, these features provide a robust foundation for the optimization methods. Figure 6.1 shows the distributions of these features. Furthermore, to ensure equitable treatment of features with varying scales, StandardScaler is deployed for feature normalization. This is done to normalize all features to a uniform scale, so that the dominance of any single feature is averted.

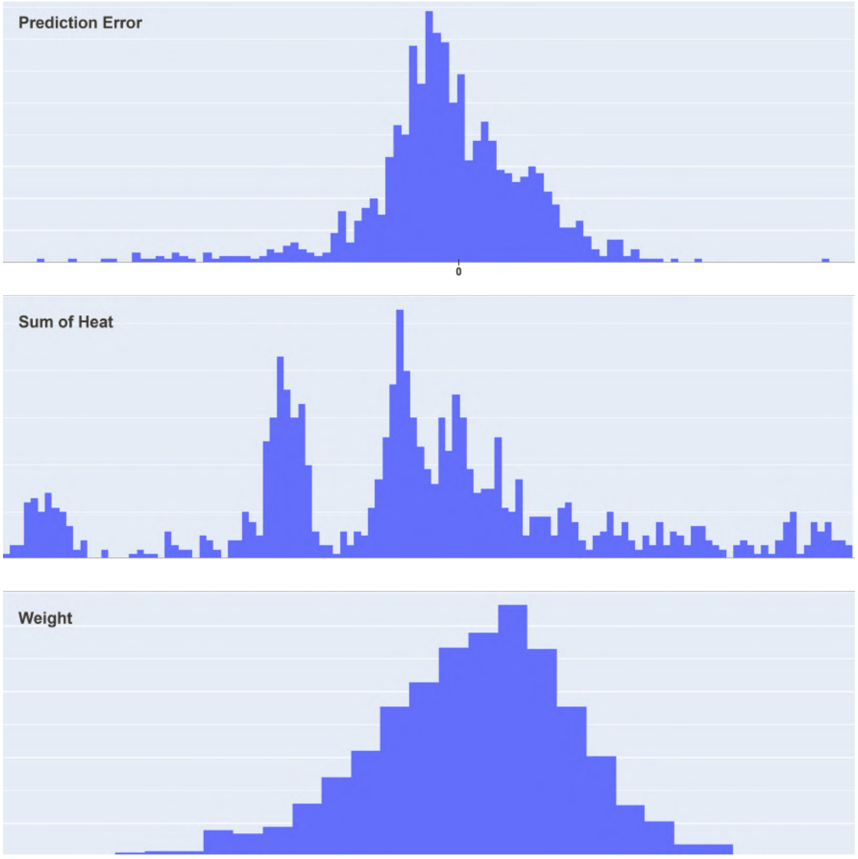


Figure 6.1: Feature Distribution for the Optimization Methods

6.2. FUZZY LOGIC

6.2.1. MODELING

The first step in the application of Fuzzy Logic is to identify the input and output variables. For the Fuzzy Logic model developed in this study, the input variables are the three aforementioned features: *fl_prediction_error*, *fl_heat*, and *fl_weight*. The output variable is the temperature of the second last heating zone, denoted by *fl_temperature_n-1*. Each of these variables spans a universe of information that is divided into a number of fuzzy subsets, with each subset assigned a linguistic variable. Table 6.1 presents the linguistic variables assigned to each feature. These linguistic variables were defined in collaboration with expert engineers from the factory to ensure their relevance and accuracy.

Table 6.1: Linguistic Variable of the Features

Feature	Linguistic Variables
fl_prediction_error	Very Under, Under, Accurate, Over, Very Over
fl_heat	Very Low, Low, Medium, High, Very High
fl_weight	Very Light, Light, Normal, Heavy, Very Heavy
fl_temperature_n	Very Low, Low, Medium, High, Very High

The subsequent step involves the derivation of membership functions for each fuzzy subset. These functions are initially derived from the empirical distribution of each feature within the historical dataset. However, these membership functions do not rely exclusively on data-driven methodologies. Instead, they undergo refinement and calibration in collaboration with the engineers from the factory, whose expertise and profound domain knowledge significantly contribute to the development of the Fuzzy Logic model. This collaboration ensures that the membership functions are aligned with the intricate nuances of the industrial process. In the model, Gaussian membership curves have been selected for all features. This choice is rooted in their capacity to closely approximate the empirical distributions of the features present in the dataset. Figure 6.2 is provided to offer a visual insight into these membership functions. For confidentiality reasons, the real values on the x-axis have been omitted from the figure.

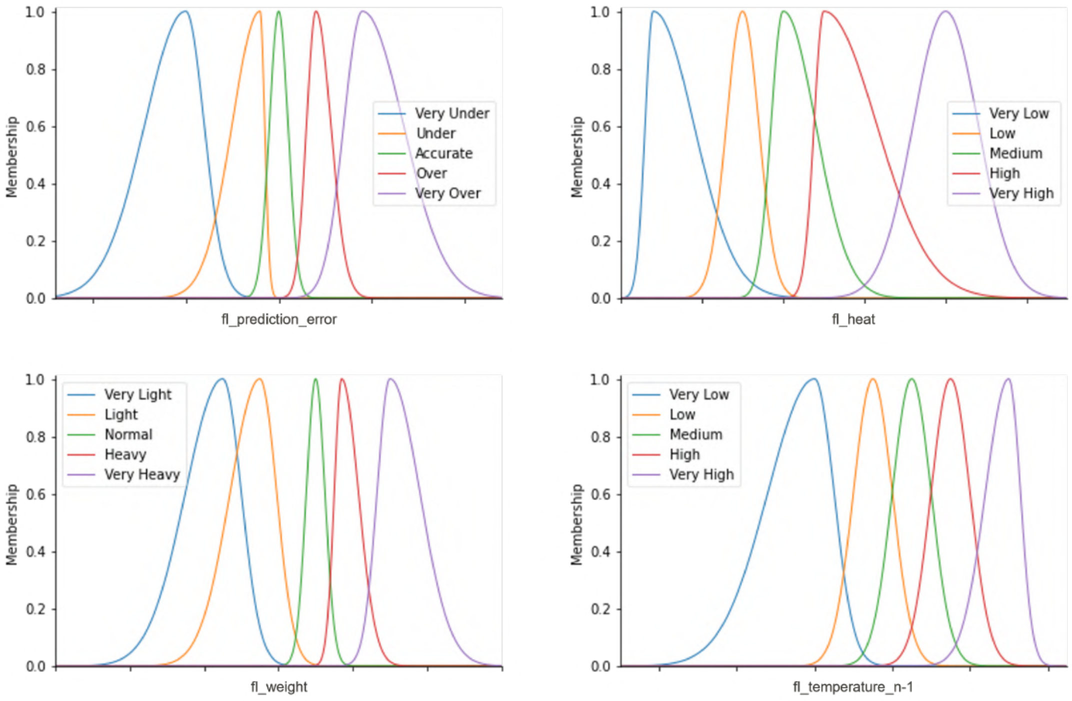


Figure 6.2: Membership Functions of Each Feature

After establishing the rules, the next step is the fuzzification process. This process involves transforming the clear-cut values of the actual inputs into fuzzy values. These fuzzy values are then matched with specific rules to acquire fuzzy output values. This entails identifying the output resulting from each rule through fuzzy approximation reasoning. Subsequently, the fuzzy outputs from all rules are amalgamated. Following this step, defuzzification is performed to convert these fuzzy values back into crisp output values. Mamdani's centroid method is adopted for the implementation of this process. The outcome of these processes is depicted in Figure 6.3, where each point signifies a data point within the Fuzzy Logic system. For ease of representation, one axis, *fl_prediction_error*, has been omitted from this illustration. However, the complete Fuzzy Logic system encompasses all three features.

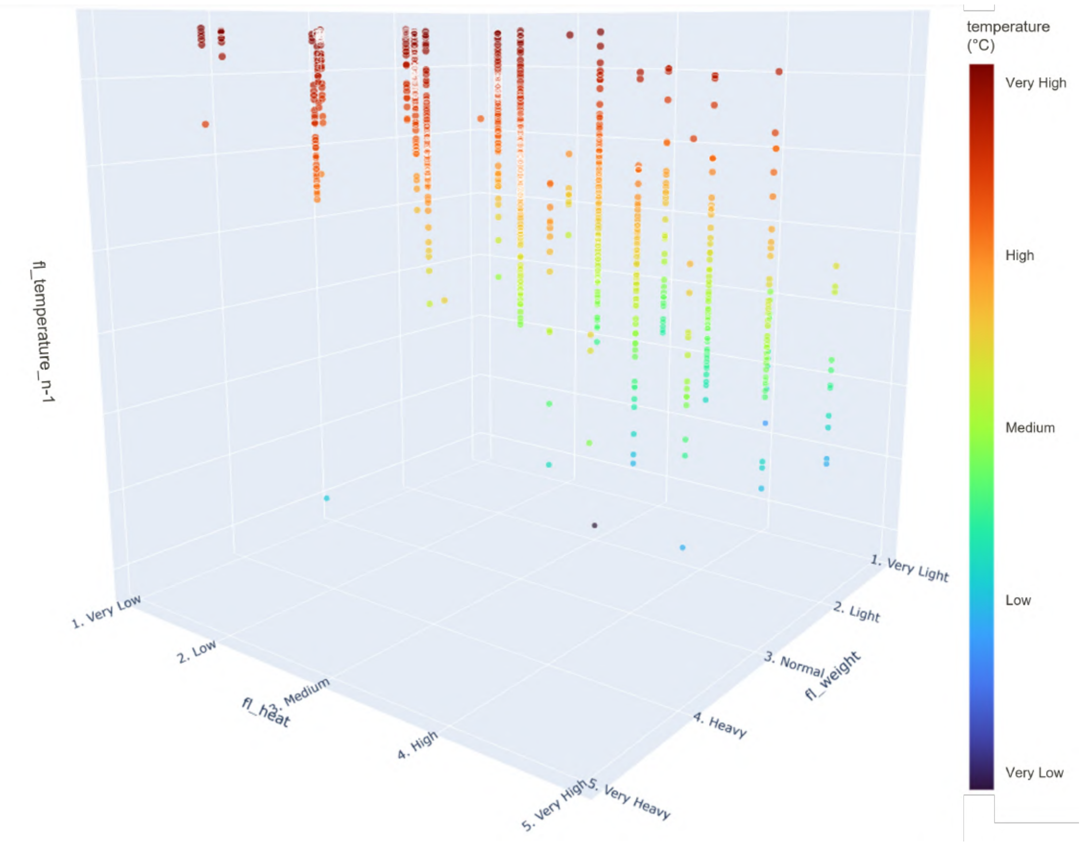


Figure 6.3: Fuzzy Logic Scatter Plot of FD 1, Product 2

Furthermore, the crisp outputs belonging to the same fuzzy rule are grouped and averaged. This averaging results in a rule-surface plane for the product, as shown in Figure 6.4. Similar to the previous figure, one axis has been omitted from this illustration for clarity. This rule-surface plot provides a visual representation of how the Fuzzy Logic model operates and how inputs are transformed into outputs for a specific product.

After the temperature of the second last heating zone has been optimized, a secondary phase of Fuzzy Logic implementation emerges. This second iteration seeks to optimize the temperature settings for the last heating zone of the FD machine. The input variables for this phase mirror the previous model, including *fl_prediction_error* and *fl_weight*. However, a notable distinction arises in the handling of *fl_heat*. In this second Fuzzy Logic model, *fl_heat* incorporates the newly optimized temperature value of the second last heating zone, fine-tuned through the preceding Fuzzy Logic iteration. This introduces a feedback mechanism, where the tempera-

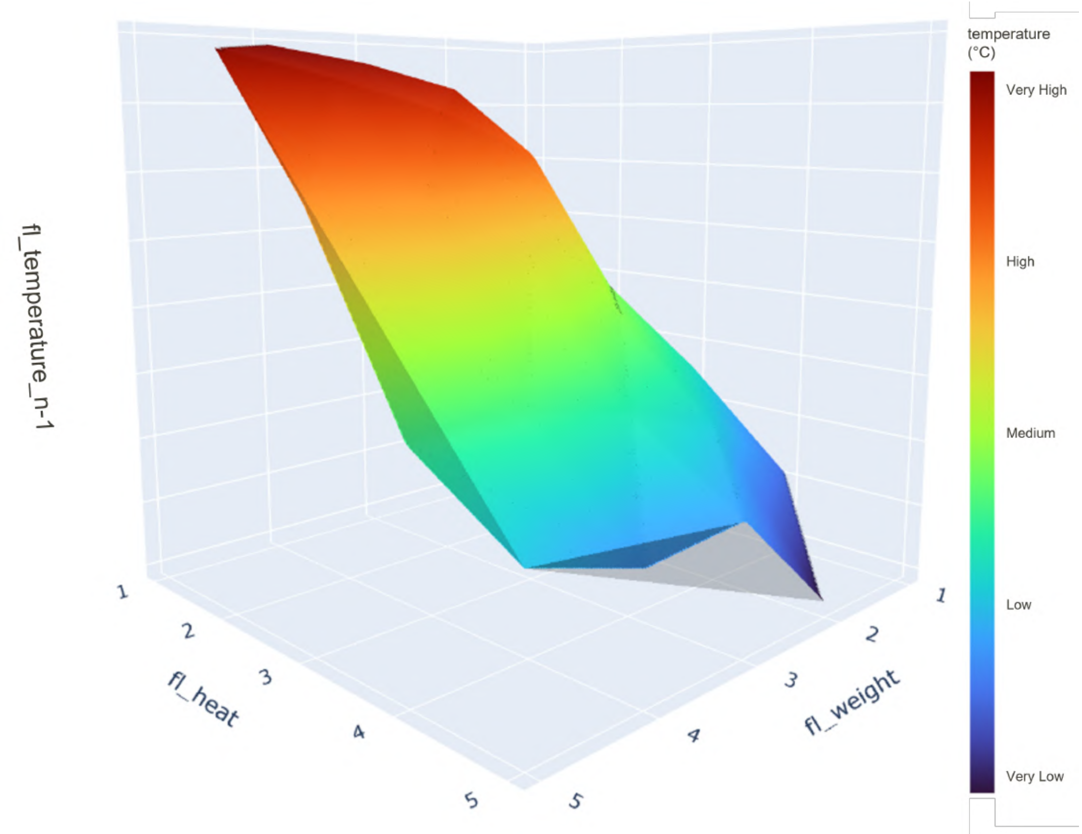


Figure 6.4: Fuzzy Logic Rule-Surface Plot of FD 1, Product 2

ture of the penultimate heating zone actively influences the settings of the last heating zone, fostering continuous optimization.

Similar to the initial Fuzzy Logic model, this second phase follows a structured approach. It commences with the derivation of membership functions tailored to the dataset, forming the basis for rule configuration. Fuzzy rules are then defined, outlining intricate relationships between input variables and the desired output, optimizing the last heating zone's temperature setting. Subsequently, the fuzzification process transforms crisp input values into fuzzy equivalents. These fuzzy values inform the appropriate rules, generating fuzzy output values through a nuanced reasoning process. Finally, defuzzification translates these fuzzy outputs into precise temperature parameters for the last heating zone, thereby optimizing the freeze-drying process to the utmost precision.

6.2.2. EVALUATION

This section critically evaluates the Fuzzy Logic model employed in the freeze-drying process, providing an objective assessment of its merits and limitations in the real operational shifts. One paramount advantage to the Fuzzy Logic model is its inherent customizability. This feature has enabled the fine-tuning of linguistic variables and membership functions through close collaboration with domain experts. This adaptability empowers the system to seamlessly align with the nuanced dynamics of our freeze-drying process, accommodating variations that may arise across different products and production runs. In an industrial context characterized by evolving demands and diverse products, this level of flexibility proves invaluable. Furthermore, the Fuzzy Logic system possesses the unique capability to emulate human deductive reasoning. Its rule-based architecture capitalizes on the collective expertise of the system engineers, effectively encapsulating their domain knowledge and decision-making processes. This quality imbues the system with a level of interpretability that is often elusive in more complex machine learning models. Engineers can readily decipher the logic underpinning the temperature adjustments, fostering an environment of trust towards the optimization method.

However, amid its considerable strengths, the Fuzzy Logic system is not without its limitations. A notable challenge is its sensitivity to outliers within the input data. In instances where outliers in heat or weight persist, the model's ability to provide accurate temperature predictions may diminish. Such instances are particularly conspicuous during significant outages or situations involving empty trays. Over time, there is a gradual and continuous increase in the *fl_prediction_error*. Consequently, the model tends to respond by generating only the very high temperature recommendations, which, if not rectified, can potentially lead to process inefficiencies or compromise product quality. An example of this in the implementation will be further covered in the next chapter. This sensitivity underscores the imperative of continuous vigilance and adaptability. Moreover, addressing this issue necessitates the periodic refinement of Fuzzy Logic rules. As outliers or exceptional scenarios surface in the production process, the engineering team must still intervene to adjust and fine-tune the ruleset. This iterative maintenance process is essential to ensure the system remains responsive to changing conditions and aligned with the evolving intricacies of freeze-drying. Such adaptability demands a high level of human expertise, both in understanding the freeze-drying process itself and in expertly navigating the nuances of Fuzzy Logic modeling.

6.3. Adaptive Neuro-Fuzzy Inference System (ANFIS)

Having explored the development and limitations of Fuzzy Logic in the preceding section, it becomes evident that a more robust optimization method is imperative to automatically fine-tune the temperature in the last two heating zones. As discussed in Section 2.4.4, while Fuzzy Logic stands out for its customizability and interpretability through linguistic rules and membership functions, ANFIS represents an evolutionary step, amalgamating the strengths of Fuzzy Logic with the adaptability inherent in neural networks. It introduces a data-driven approach to modeling, thereby enhancing the capacity to accurately adjust temperature parameters within the freeze-drying process. Notably, ANFIS has found applications in similar studies aimed at forecasting and optimizing comparable dependent features in various domains, including Agriculture [16, 35, 37, 46, 49], Energy [29, 45], and Climate [31, 50]. In this section, the intricacies of ANFIS will be navigated, providing a comprehensive account of its configuration, training process, and the pivotal role it assumes in optimizing heating zone temperatures.

6.3.1. MODELING

In the development of the ANFIS model for this study, a custom-built function was crafted to create the architecture. While there are existing open-source ANFIS packages available, such as anfis[1] and sanfis[2], the decision to create a custom function was driven by the need for enhanced customizability and the ability to conduct rigorous testing throughout the modeling process. The architecture of the customized ANFIS model is determined by some critical parameters:

- Number of Input Features (n): There are three input features, which directly corresponds to the dimensionality of the training dataset. As mentioned before, these input features consist of x_1=*fl_prediction_error*, x_2=*fl_heat*, and x_3=*fl_weight*.
- Number of Fuzzy Subsets (*fs*): This corresponds to the linguistic variables of each fuzzy subset, for instance "Very High", "High", "Medium", "Low", and "Very Low". Aligned with the fuzzy subsets used for the Fuzzy Logic in the previous section, five fuzzy subsets are used for each input feature.
- Number of Fuzzy Rules (m): To effectively capture the underlying data patterns, the number of rules, m, is set to $(fs)^n$. This results in the construction of a total of 125 rules, ensuring a thorough representation of the system's behavior.
- Learning Rate: During the training phase, a learning rate of 0.01 is chosen deliberately. This learning rate strikes a balance between rapid convergence and stability in the training process. A higher learning rate might lead to faster convergence but risks overshooting the optimal parameter values, causing instability and divergence. Conversely, a smaller learning rate ensures stability but may result in slow convergence.
- Number of Epochs: The training process spans 20,000 epochs to ensure convergence.

The ANFIS model is then trained using the `TensorFlow`[3] framework. The training process encompasses the following key stages:

INITIALIZATION

The model parameters, including the means (*mu*) and standard deviations (*sigma*) of Gaussian membership functions, as well as the weights (w) for fuzzy rules, are initialized with random values.

FORWARD PASS (TRAINING)

The training dataset, composed of input features (*X_train*) and target values (*y_train*), is fed into the ANFIS model. Utilizing Gaussian membership functions based on the input features and parameters, the model follows the processes depicted in Figure 3.2 until it derives the predicted output (*Y_train*) in the first epoch.

[1] https://pypi.org/project/anfis/
[2] https://pypi.org/project/sanfis/
[3] https://www.tensorflow.org/

LOSS COMPUTATION (TRAINING)

To gauge the model's performance, the RMSE is employed as the main loss function, quantifying the disparity between the predicted output (Y_train) and the actual target values (y_train).

BACKWARD PASS (BACKPROPAGATION)

The Adam optimizer is employed for minimizing the loss. This optimization phase employs backpropagation to adjust the model's parameters, specifically the membership function parameters (*mu* and *sigma*) and rule weights (w), to gradually enhance predictive accuracy of the model.

TESTING

A separate test dataset (X_test) is used to evaluate the model's generalization capabilities. The ANFIS model computes predictions (Y_test) for this test dataset.

LOSS EVALUATION (TESTING)

The evaluation metrics are recalculated for the test dataset, providing an assessment of the model's ability in generalizing to unseen data.

MODEL SAVING

Upon completion of the training, the trained ANFIS model is saved for future deployment.

The crux of this optimization process centers around minimizing a designated objective function, frequently representing a measure of prediction error. Through iterative refinement, the ANFIS model adapts its fuzzy rules to furnish increasingly precise predictions. The outcome is an adaptive, hybrid system that learns from historical data while retaining the ability to articulate its logic through human-understandable fuzzy rules.

6.4. EVALUATION

The evaluation of the ANFIS model encompasses a comprehensive assessment of its predictive capabilities. To gauge the model's performance, several key metrics are employed, with RMSE serving as the primary indicator of predictive accuracy. Additionally, MAE and R^2 are considered to provide a well-rounded evaluation. Similar to the machine learning evaluation, a robust 5-fold k-fold cross-validation methodology is employed to ensure the reliability of the results. For a visual representation, Figure 6.5 illustrates two scatter plots of predicted against actual temperature values for Product 7 in FD 2, one for the training set and another for the test set. The averaged metrics are also displayed on these plots, providing a consolidated overview of the model's overall performance.

Analyzing the metrics in detail, for the training set, the model achieves an R^2 coefficient of 0.72, indicating its capacity to capture a significant portion of data variance. Additionally, the RMSE value is 7.92, and the MAE value is 5.99, underscoring the model's ability to make predictions with minimal error on the training data. Moving to the test set, the model maintains its predictive strength with an R^2 coefficient of 0.73, demonstrating its robust generalization capabilities.

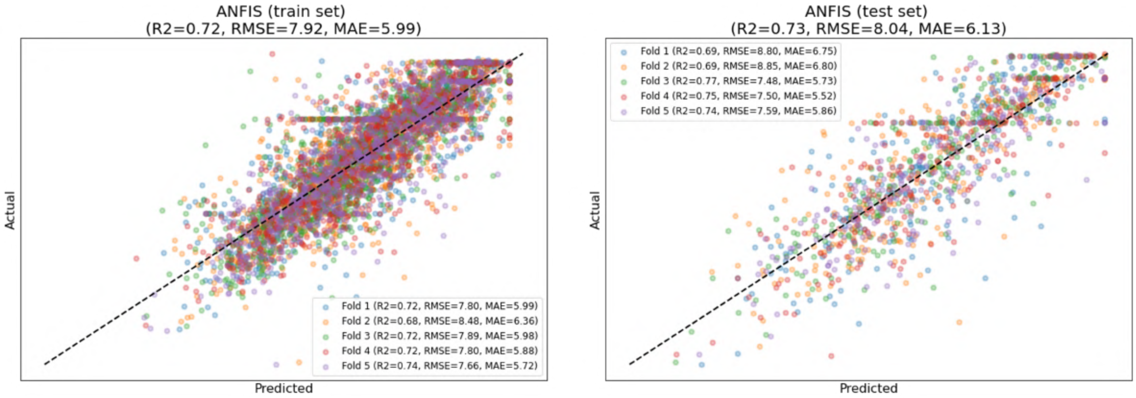


Figure 6.5: ANFIS Scatter Plot of FD 2, Product 7

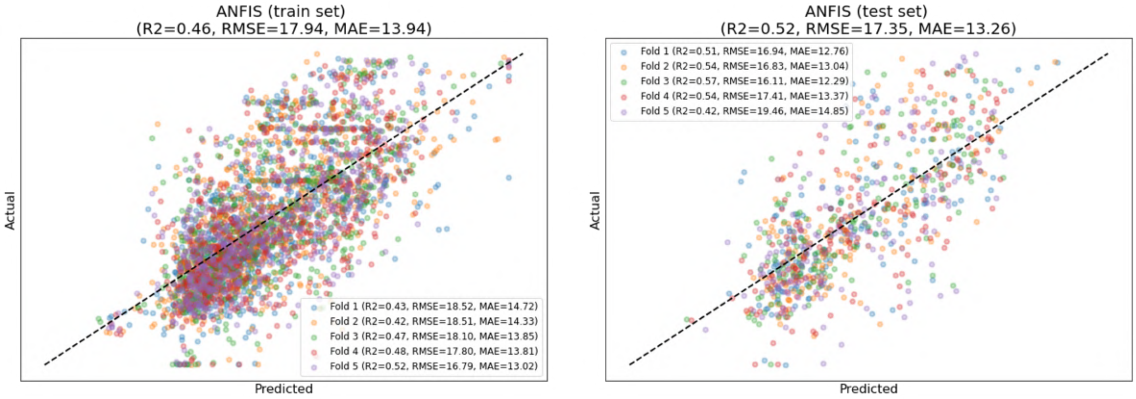


Figure 6.6: ANFIS Scatter Plot of FD 2, Outage

While the RMSE (8.04) and MAE (6.13) in the test set are slightly higher than those in the training set, they remain at reasonable levels, indicating the model's resilience against overfitting. Moreover, in Appendix L, a comprehensive table provides an extensive overview of the ANFIS model's performance across different freeze-drying scenarios. Notably, the test metrics in this table align closely with the train metrics, affirming the model's consistent generalization across various scenarios. However, specific situations, such as outages, present unique challenges, leading to less favorable metrics. Figure 6.6 shows that in outage conditions, the model exhibits a lower R^2 coefficient but higher RMSE and MAE values, signifying the model's reduced predictive accuracy and increased prediction errors. This divergence can be attributed to the complexity and variability inherent in these conditions, with unusual data patterns and limited historical data for training. Nevertheless, it is crucial to highlight that the model's performance decline in outage situations is expected due to the substantial disruptions and anomalies associated with such scenarios. Despite these challenges, the ANFIS model maintains a reasonable level of generalization, underscoring its versatility and robustness in real-world freeze-drying applications.

The ANFIS model exhibits several notable strengths that contribute to its effectiveness in freeze-drying applications, allowing for a comparison with the Fuzzy Logic model. One of its paramount advantages lies in its capacity to capture intricate, nonlinear relationships within the data. This characteristic is particularly advantageous in scenarios where freeze-drying dynamics can be highly complex and nonlinear, outperforming Fuzzy Logic models in handling such intricacies. Moreover, ANFIS is adept at automatic feature selection and extraction, reducing the need for extensive manual engineering of linguistic variables and membership functions, as is often required in Fuzzy Logic models. This automates and streamlines the model-building process, potentially saving time and effort in model development. Furthermore, the model's adaptability and capacity to generalize across different freeze-drying conditions highlight its versatility and real-world applicability, comparable to Fuzzy Logic's customizability.

However, akin to the Fuzzy Logic model, ANFIS is not without its limitations. One notable challenge is the requirement for a substantial amount of high-quality data for training. In the case of freeze-dried coffee production, obtaining such data can be challenging and may necessitate significant effort and resources. This includes a considerable amount of time needed to acquire the data, and potentially the addition of more machines and setups to add features or to scale up the dataset. Additionally, ANFIS performance can be sensitive to the quality of input features and the selection of appropriate hyperparameters, demanding meticulous data preprocessing and tuning, similar to the need for fine-tuning Fuzzy Logic rules. Another aspect to consider is that ANFIS, although interpretable to some extent, may not offer the same level of transparency and interpretability as Fuzzy Logic in certain cases, as Fuzzy Logic inherently emulates human deductive reasoning and encapsulates domain knowledge explicitly. This nuanced comparison illustrates that while ANFIS excels in handling complex data relationships and automating feature engineering, Fuzzy Logic stands out in its interpretability and adaptability to domain-specific expertise, albeit with sensitivity to outliers.

6.5. CHAPTER SUMMARY

This chapter explores optimization methods for automatically adjusting temperature settings in freeze-drying processes based on predicted moisture levels. Two primary approaches are discussed: Fuzzy Logic and ANFIS. The chapter begins with data preparation, introducing a *prediction* column sourced from moisture predictions and highlighting key features. Fuzzy Logic is then detailed, emphasizing its modeling process, customizability, and interpretability. However, this approach is sensitive to outliers and necessitates continuous manual rule adjustments.

The chapter further delves into ANFIS, explaining its modeling process, automatic fuzzy rules engineering, and its unique ability to handle complex, nonlinear relationships while minimizing manual intervention. Both methods are evaluated, with the evaluation of relevant performance metrics, highlighting ANFIS as the preferred choice due to its adaptability and efficiency in handling intricate data dynamics. In summary, ANFIS excels in automating temperature adjustments in freeze-drying processes, offering a robust solution to optimize production efficiency.

7

Implementation and Automation

This chapter provides a practical exploration of how the proposed solution transforms the freeze-dried coffee production process at JDE. Section 7.1 delves into the implementation of machine learning algorithms, carefully tailored to various FD machines and products. Section 7.2 discusses the seamless integration of the optimization method, highlighting its role in automating temperature adjustments within the last heating zones of the FD machine based on moisture predictions. The overview of the entire solution framework is presented in Section 7.3. Finally, in Section 7.4, the tangible results of the implemented solution are presented, offering a comparative analysis against previous manual processes to showcase its real-world effectiveness and efficiency.

7.1. Deployment of Machine Learning Algorithm

The initial phase of implementation involved deploying a machine learning algorithm, a crucial step that laid the foundation for subsequent optimization methods. In commencing this process, the machine learning model was executed in the FD machines 1 and 2. The implementation began with these machines, paving the way for the creation of the new *prediction* column in the dataset and generating initial data. This *prediction* column played a pivotal role in facilitating moisture predictions, as it would be used for calculating the later *prediction_error*. Figure 7.1 visually depicts the interface button within the FD machine system, through which the machine learning algorithm was executed. The successful execution of the model in these initial machines laid the groundwork for its expansion to FD machines 3 and 4, further enhancing the automation of the freeze-drying process.

Figure 7.2 provides a visual representation of the machine learning algorithm's implementation in FD machine 2. Notably, this phase of implementation focused solely on predicting moisture levels, with no temperature adjustments made at this stage. The chart illustrates the model's predictions alongside actual moisture values and target moisture levels. Upon close examina-

Figure 7.1: Machine Learning Enabler Button

tion of the results, it becomes evident that the machine learning model performed well in predicting moisture levels. The predicted values closely align with the actual moisture measurements, with deviations well within acceptable margins. This outcome underscores the model's proficiency in capturing the intricate dynamics of moisture within the freeze-drying process, setting a solid precedent for its role in subsequent optimization phases.

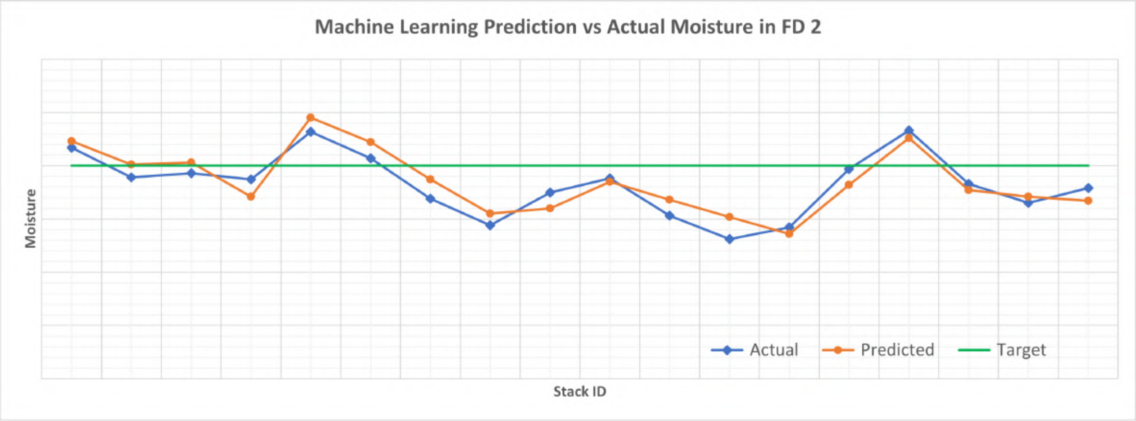
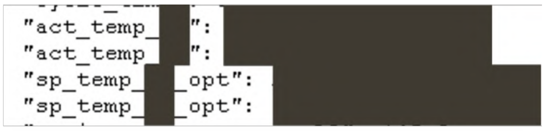


Figure 7.2: Implementation of the Machine Learning Algorithm

7.2. Deployment of Optimization Method

Following the successful implementation and operation of the machine learning algorithm, the next crucial phase in the enhancement of the freeze-dried coffee production process is the deployment of the optimization method. As previously mentioned, the machine learning algorithm precedes the optimization method and plays a pivotal role in generating the requisite dataset. After the *prediction* column accumulates sufficient data through the machine learning algorithm's predictions, this data undergoes the data preparation process, serving as the foundation for the optimization method. The optimization method is strategically designed to harness this enriched dataset, enabling it to make data-driven temperature adjustments in the later stages of the freeze-drying process.

Figure 7.3 offers a visual representation of the machine learning algorithm and optimization method's deployment within FD machine 2. The operation of the machine learning algorithm initiates before the freeze-dried coffee stack enters the last two heating zones, specifically at stack position *n-6*. The machine system interface located in the top right corner of Figure 7.3 displays essential parameters that provide insight into the optimization process. Notably, *act_temp_n* showcases the current actual temperature within the heating zones, while *sp_temp_n_opt*



```
"act_temp_[redacted]": [redacted]
"act_temp_[redacted]":
"sp_temp_[redacted]_opt": [redacted]
"sp_temp_[redacted]_opt":
```

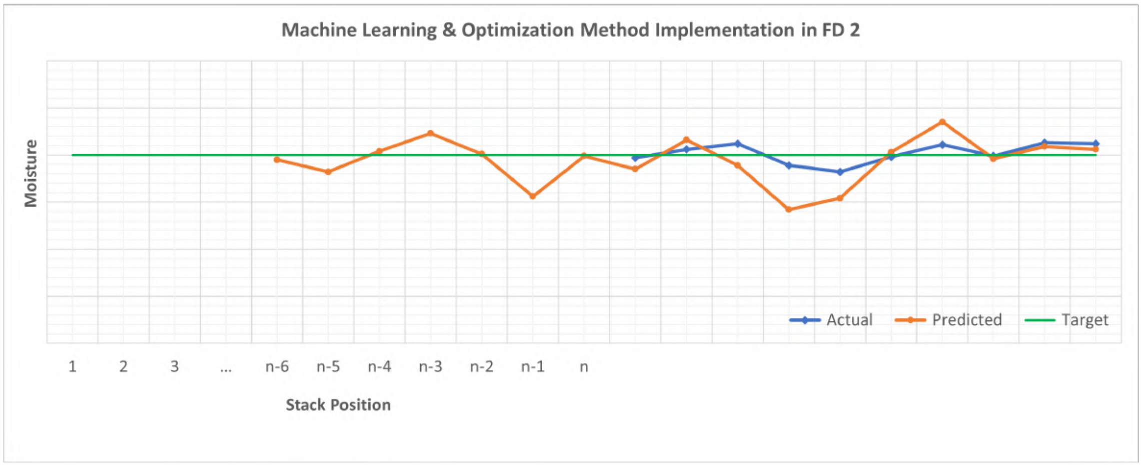


Figure 7.3: Implementation of the Optimization Method

reveals the optimum temperature determined by the optimization method. As the stack progresses through the heating zones, the system adapts, regulating the temperature to align with the calculated optimum temperature, thereby optimizing the moisture content.

The results of this optimization become evident once the stack surpasses position *n*. The accompanying line chart in Figure 7.2 illustrates that the optimization method effectively brings the moisture content closer to the predefined target. When the predicted moisture significantly deviates from the target, the optimization method orchestrates adjustments to the actual moisture, steering it towards proximity with the target. The provided reference numbers corroborate this process, demonstrating how the optimization method diligently fine-tunes the moisture content to achieve consistency, a critical factor in freeze-dried coffee production's quality and efficiency enhancement.

7.3. SOLUTION OVERVIEW

The proposed solution aims to revolutionize the freeze-dried coffee production process by integrating advanced data-driven techniques to enhance efficiency and quality control. As illustrated in Figure 7.4, the framework of the solution encompasses multiple stages, each contributing to the optimization of the production process. The process begins with the acquisition and storage of data from the FD machine. As the stack of trays reaches the few last position in the FD machine's heating zone, a machine learning model is employed to predict the moisture content of that stack. This prediction is based on calculated features derived from the accumulated data, capturing critical nuances that influence the moisture outcome. Notably, distinct machine learning models are tailored for specific FD machines, products, and even instances of outages, acknowledging the aforementioned unique characteristics of each product.

Upon obtaining the moisture prediction, the solution introduces a dynamic adjustment mechanism that holds the potential to reshape the traditional production paradigm. The machine learning model triggers automatic temperature adjustments in the last remaining heating zones of the FD machine by means of ANFIS. This optimization method is meticulously calibrated to ensure that the desired target moisture is achieved consistently across batches. Subsequently, the actual moisture content is measured using a moisture meter and compared against the predictive output of the machine learning model. Any disparities between the prediction and actual outcome are recorded and further assimilated into a feedback loop that continually refines and enhances the predictive capabilities of the machine learning and optimization method model.

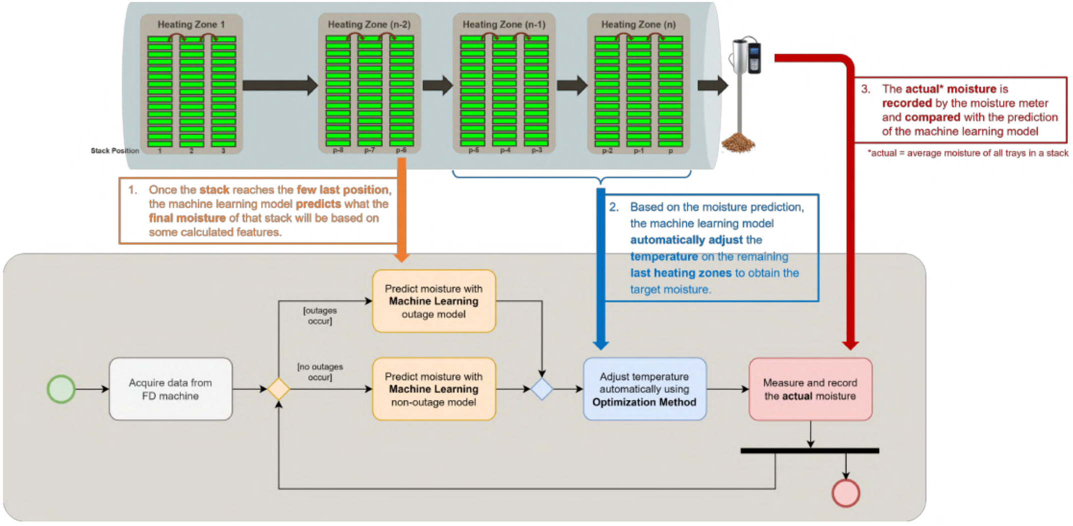

Heating Zone 1
Heating Zone (n-2)
Heating Zone (n-1)
Heating Zone (n)
Stack Position
1
2
3
p-8
p-7
p-6
p-5
p-4
p-3
p-2
p-1
p
3. The **actual* moisture** is **recorded** by the moisture meter and **compared** with the prediction of the machine learning model
*actual = average moisture of all trays in a stack
1. Once the **stack** reaches the **few last position**, the machine learning model **predicts** what the **final moisture** of that stack will be based on some calculated features.
2. Based on the moisture prediction, the machine learning model **automatically adjust** the **temperature** on the remaining **last heating zones** to obtain the target moisture.
Acquire data from FD machine
[outages occur]
[no outages occur]
Predict moisture with **Machine Learning** outage model
Predict moisture with **Machine Learning** non-outage model
Adjust temperature automatically using **Optimization Method**
Measure and record the **actual** moisture

Figure 7.4: Framework of the Proposed Solution

7.4. IMPLEMENTATION RESULT OF THE SOLUTION

Within the scope of this section, a thorough examination of the results achieved during the practical implementation of the proposed solution is presented. Figure 7.5, which illustrates the moisture distribution in each FD machine under non-outage conditions, serves as a valuable reference point for the analysis.

The results indicate a substantial improvement in the performance of the current solution. Notably, the machine learning model's mean moisture content consistently approaches the target value, denoted by the "X" mark on the x-axis, compared to the previous manual approach. This suggests a remarkable level of precision and consistency in the production process. Furthermore, a notable decrease in the standard deviation of moisture levels in the current outcome compared to the previous manual process underscores the superior control and minimized variability attained through the automated solution. Specifically, the automated process yielded moisture levels that were 72% closer to the target, accompanied by a notable 29% reduction in the standard deviation when compared to the manual process. These outcomes are particularly relevant in the context of freeze-dried coffee production, where product quality

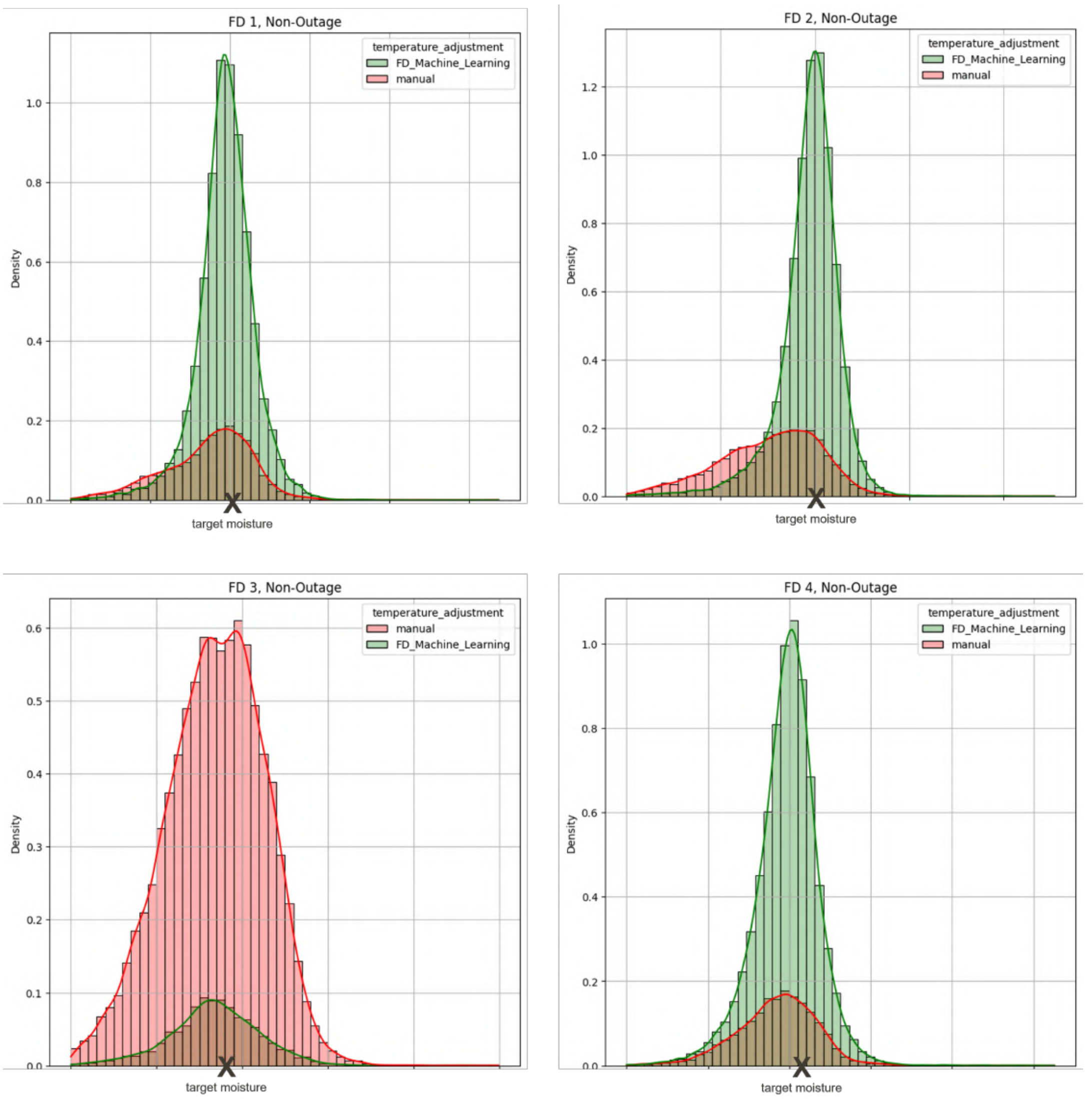


Figure 7.5: Implementation Result in Every FD Machine (Non-Outage)

and consistency are of utmost importance. Furthermore, the analysis extends to the robustness of the solution, as evidenced in Appendix M, which presents results obtained during outage situations. This supplementary data reaffirms the solution's effectiveness, even when facing operational challenges. It is noteworthy that the solution continues to perform admirably during outages, demonstrating its potential to revolutionize freeze-dried coffee production at JDE.

In addition to precision and robustness, the solution showcases its ability to seamlessly execute end-to-end processes within the allocated time limits at each stack along the production line. This means that from data gathering and pre-processing to machine learning-based moisture level predictions and the subsequent optimization method to adjust temperature, every step operates efficiently in synchronization with the production line's movement. Importantly, all of these processes are completed before a stack progresses to the next heating zone. This timeliness is vital for enabling real-time decision-making and process optimization, allowing for swift adjustments in response to dynamic production conditions.

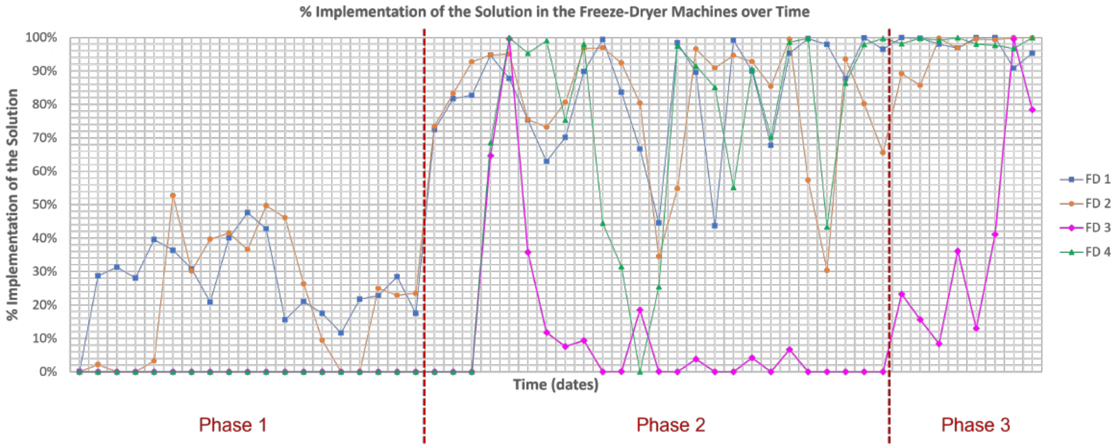


Figure 7.6: Implementation Percentage of the Solution in the FD Machines over Time

The line graph in Figure 7.6 illustrates the utilization percentage of the solution across all FD machines over time. The implementation was divided in three phases, where it was initially deployed in FD 1 and 2 in the first phase. In the second phase, the solution's usage steadily increased, eventually encompassing FD 3 and 4. In the second phase, some questions were also asked to the expert regarding the fluctuation of the implementation in each of the machine. Notably, the periods of reduced usage are typically associated with maintenance activities or ad-hoc tests in the machines. During these events, the end-to-end processes in the machines require manual control and constant monitoring, hence turning off the solution for the period. Additionally, there was a period of improvement in FD 3 that coincided with reduced solution usage, although the specific details of this experiment could not be directly disclosed. The latter portion of the graph indicates that the solution has been consistently implemented in all FD machines during the third phase. Although the graph does not directly convey expert's opinions, it indirectly emphasizes the solution's pivotal role in production, highlighting its alignment with routine activities. Furthermore, the engineer's productivity increment resulting from the FD machines can be quantified by examining the average percentage of its implementation over time. During the first phase, there is a 14% increase in productivity. The second phase demonstrates a remarkable 61% boost, and during the third phase, this figure further escalates to an impressive 83%. These findings underscore the indirect positive impact of the solution on the engineer's productivity, as they can spend their valuable time in doing other improvements.

The insights and methods discussed in this study hold promise for wider use. Although the focus has been on improving freeze-dried coffee production at JDE, but the developed framework and techniques can serve as a valuable starting point for similar projects in other companies and industries. The method, combining machine learning and optimization, can be applied in various scenarios where it's necessary to automatically adjust certain aspects to achieve desired results based on predictions. The solution's resilience, especially its ability to perform well under tough conditions like outages, demonstrates its adaptability to different manufacturing settings where quality control and process improvement are crucial. The core idea of using data-driven techniques for better processes and quality control may be applicable to many situations, such as predicting product characteristics, optimizing production processes, or enhancing quality assurance in different industries. However, it is important to note that each case and industry may require some customization. While the basic principles remain consistent, the way to implement and fine-tune models and optimization methods will vary depending on the specific situation. So, it is essential to tailor the approach to get the best results, even though the method itself has proven effective in this study.

7.5. CHAPTER SUMMARY

This chapter presents the practical implementation of the proposed solution for enhancing freeze-dried coffee production at JDE. The broader solution framework integrates data-driven techniques, from data acquisition, machine learning prediction, to automatic temperature adjustments, enhancing production efficiency and quality control. In addition to presenting the framework, this chapter showcases the tangible results of the implemented solution. Moisture distribution analysis under non-outage conditions demonstrates the solution's precision and consistency in approaching the target moisture content, accompanied by a reduced standard deviation, highlighting enhanced control and minimized variability. Collectively, this section underscores the potential of the automated solution to revolutionize freeze-dried coffee production at JDE, offering improved efficiency and heightened product quality consistency.

8

CONCLUSION

This concluding chapter encapsulates the essence of the research journey, concluding the processes and key findings discovered throughout the study. Section 8.1 provides a comprehensive overview of the study's main conclusions. Subsequently, in Section 8.2, the research addresses sub-research questions 5, 6, and 7, respectively. Furthermore, this chapter delves into the academic and practical contributions of this thesis in Section 8.3, emphasizing the research's relevance and impact on both academic knowledge and real-world applications. Lastly, in Section 8.4, the study acknowledges its limitations and outlines promising avenues for future research in the domain of this study.

8.1. CONCLUSION

This thesis embarked on a mission to develop an optimized and automated approach for the heating process of freeze-dried coffee production at JDE. The primary aim was to diminish the variability in target moisture levels, ultimately leading to a considerable enhancement in product quality through consistent moisture levels, and a substantial boost in efficiency by replacing manual processes with automated schemes. The research journey commenced with an exhaustive examination of the current state of knowledge, spanning predictive techniques, optimization methods, and the implementation of process automation across diverse fields. Through a meticulous analysis of academic articles and research papers, this thesis identified critical gaps and research opportunities in the realm of machine learning, laying the groundwork for the research endeavors.

The methodology was firmly rooted in the CRISP-DM framework, providing a robust structure for the systematic exploration of machine learning algorithms, optimization techniques, and automation approaches. A spectrum of machine learning algorithms was carefully evaluated, drawing inspiration from applications in related fields. The selection of the most adept algorithm, guided by predefined metrics, formed the cornerstone of the approach. Subsequently, optimization methods capable of dynamically adjusting temperature settings based on moisture predictions were seamlessly integrated, ushering in a new era of precision in freeze-dried coffee production.

The culmination of this research journey witnessed the proposal of a groundbreaking solution — a synergy of machine learning and optimization methods. This solution was validated and implemented within the real-world freeze-dried production system at JDE. Through meticulous implementation, the power of data-driven automation was harnessed to achieve the remarkable goal of consistency in moisture levels, significantly elevating product quality. This journey underscores the significant potential of data-driven automation in industrial settings, offering a guiding light towards improved productivity and product quality in freeze-dried coffee production and beyond.

8.2. ANSWERS TO RESEARCH QUESTIONS

8.2.1. SUB-RQ5: MACHINE LEARNING ALGORITHM TO PREDICT THE MOISTURE OF FREEZE-DRIED COFFEE

The ultimate goal in finding the optimal machine learning algorithm to predict moisture levels in freeze-dried coffee production involved the collaboration of advanced algorithms, data complexity, and model performance. The comprehensive evaluation of various regression algorithms highlights the superiority of the XGBoost Regression model in achieving the most accurate and reliable predictions. This conclusion is substantiated by the model's remarkable performance across key metrics such as R^2, RMSE, and MAE. XGBoost Regression consistently achieved the highest counts for the best metrics across different products and FD machines, securing the top position in terms of predictive accuracy. This model's ability to capture complex interactions and non-linear relationships within the data sets it apart from other models.

The prevalence of non-linear models, such as XGBoost Regression, highlights the likely presence of non-linear relationships between predictors and the target variable, indicating the inadequacy of purely linear models in capturing the intricate patterns in the data. The robust performance of XGBoost Regression reinforces the necessity of incorporating non-linear algorithms to accurately predict moisture levels in freeze-dried coffee production. By leveraging ensemble learning and gradient boosting, XGBoost Regression excels in capturing the nuanced relationships and interactions that contribute to the variability in moisture levels. This predictive accuracy can significantly enhance the quality and consistency of freeze-dried coffee products, ultimately benefiting the consistency level in the production process.

8.2.2. SUB-RQ6: OPTIMIZATION METHOD TO AUTOMATICALLY ADJUST THE TEMPERATURE

In identifying the most suitable approach for automatically adjusting temperature settings in heating zones based on predicted moisture levels in freeze-dried coffee production, the Adaptive Neuro-Fuzzy Inference System (ANFIS) method stands out as the favored choice. AN-

FIS offers a unique blend of capabilities that makes it exceptionally well-suited for this task. Freeze-drying processes are often characterized by complex and nonlinear relationships between variables. ANFIS excels in handling these intricacies through its data-driven approach. It has the ability to effectively model the nonlinear dynamics of moisture prediction and subsequent temperature adjustments. This is particularly crucial in a freeze-drying context, where precise control over temperature is paramount for product quality.

One of the significant strengths of ANFIS is its ability to continuously optimize the weights associated with the rules derived from membership functions. Through epoch-based iterative forward and backward passes, ANFIS dynamically adjusts these rule weights to refine its predictive accuracy. This process allows ANFIS to fine-tune its decision-making mechanism, continuously learning from the data to provide increasingly precise temperature recommendations. Additionally, it reduces the need for extensive manual work in crafting linguistic variables and membership functions, streamlining the model-building process. This not only potentially saves time and effort but also reduces the risk of human error in feature engineering. Additionally, ANFIS exhibits remarkable versatility and generalization capabilities across diverse freeze-drying conditions. Real-world production scenarios can vary widely, and ANFIS's adaptability ensures consistent and efficient temperature adjustments. It can effectively handle variations in process dynamics, making it a reliable choice for optimizing temperature settings in the face of changing conditions.

While Fuzzy Logic has its strengths, such as customizability and interpretability, ANFIS's unique ability to handle complex, nonlinear relationships in data and its continuous optimization of rule weights derived from membership functions make it the preferred method for automatically adapting temperature settings in freeze-drying processes. ANFIS not only enhances production efficiency by reducing manual workloads but also ensures precise temperature control amidst intricate process dynamics. Unlike Fuzzy Logic, which often requires manual crafting of linguistic variables and membership functions, ANFIS automates much of the feature engineering process. This automation streamlines model development and reduces the potential for human errors. In essence, ANFIS aligns perfectly with the goal of optimizing temperature settings based on predicted moisture levels while minimizing operational complexities and elevating production efficiency.

8.2.3. Sub-RQ7: Comparison of the Optimized and Automated Heating Process with the Manual Process

The optimized and automated heating process, as implemented in this study, exhibits significant advantages when compared to the current manual processes. Through the deployment of machine learning algorithms and subsequent optimization methods, the automated solution has demonstrated superior performance in controlling moisture levels during the freeze-drying process. The results reveal that the mean moisture content closely aligns with the target value, indicative of a higher degree of precision and consistency in production. Moreover, a substantial reduction in the standard deviation of moisture levels underscores the enhanced control and reduced variability achieved by the automated solution.

Additionally, the analysis extends to outage situations, where the automated solution continues to perform admirably, reaffirming its robustness and reliability even when facing operational challenges. These findings underscore the potential of the automated solution to revolutionize freeze-dried coffee production at JDE, offering the promise of improved efficiency in production processes and heightened consistency in product quality. This comprehensive performance evaluation provides concrete evidence of the significant benefits and advancements brought about by the implementation of the proposed solution, marking a pivotal step towards the future of freeze-dried coffee production.

8.3. STUDY CONTRIBUTIONS

8.3.1. ACADEMIC CONTRIBUTION

This study makes several notable academic contributions. Firstly, it conducts an extensive SLR to comprehensively analyze the existing body of knowledge related to predictive techniques, optimization methods, and process automation across various fields. In doing so, it identifies the research gaps, particularly within the Manufacturing domain, which appears relatively underexplored. Significantly, within the broader engineering domain closely associated with manufacturing, only a few studies have integrated both machine learning algorithms and optimization methods. Moreover, none of these existing studies have combined all three crucial components of the proposed solutions in this thesis: machine learning algorithms, optimization methods, and implementation & process automation, in the domain of freeze-dried coffee production.

This research addresses this gap by exploring, evaluating, and selecting the most suitable machine learning algorithms for predicting moisture levels in freeze-dried coffee production. The rigorous assessment of these algorithms according to predefined metrics adds to the existing knowledge base in predictive modeling. Additionally, the study investigates optimization methods tailored for temperature adjustments based on moisture predictions, thereby enhancing the understanding of optimization techniques in industrial applications. These contributions aim to advance the academic discourse surrounding data-driven automation in manufacturing, potentially leading to the development of decision support systems that optimize manufacturing operations and enhance decision-making processes.

8.3.2. PRACTICAL CONTRIBUTION

In practical terms, this research offers substantial contributions to the field of freeze-dried coffee production at JDE. The developed solution, encompassing machine learning algorithms and optimization methods, has been implemented and validated within the real production system. This practical application showcases the feasibility and effectiveness of the proposed approach in a real-world industrial setting, providing actionable insights for process improvement. Moreover, the implementation of data-driven automation contributes to enhanced efficiency and consistency in freeze-dried coffee production. This transition from manual processes to automated schemes is expected to yield tangible benefits for JDE, including reduced variability in moisture levels, improved product quality, and it supports increased productivity of the engineers. In addition to validating the feasibility of the proposed approach, these practical contributions also hold the potential to drive positive transformations within the manufacturing industry, particularly in coffee production.

8.4. STUDY LIMITATIONS AND DIRECTION FOR FUTURE RESEARCH

While this research has made significant strides in advancing data-driven automation within freeze-dried coffee production, certain limitations should be acknowledged. First, this research primarily operates within the domain of freeze-dried coffee production at JDE. Although the findings are promising, the generalizability of the proposed solution to other manufacturing processes or industries remains untested. It is essential to recognize that different production systems may exhibit unique complexities that necessitate tailored solutions. The optimization methods employed in this study can be further fine-tuned and customized to suit other specific production scenarios, for instance predicting defects or quality issues in automotive parts production, thus optimizing the manufacturing process in engineering field, or forecasting variations in drug formulations and optimizing the pharmaceutical production to account for these variations. Comparative studies in various industries could elucidate the transferability and adaptability of the approach. Future research could also delve into advanced optimization techniques or explore the potential of reinforcement learning for dynamic process control, as several papers has proposed new techniques such as Deep Q-Networks, Proximal Policy Optimization, and Actor-Critic Methods [105–107].

Second, the focus of this study primarily centered on moisture prediction and optimization of heating processes. Other critical aspects of the production pipeline, such as quality control at subsequent stages or environmental sustainability, were not explored. One avenue for further investigation is the integration of additional data sources, such as environmental conditions or equipment health parameters, into the predictive and control frameworks. This expansion of data inputs could enhance the accuracy and robustness of the automation system. Additionally, exploring advanced quality control mechanisms within the production process, possibly utilizing computer vision or sensor technologies, could be a valuable area of research.

Moreover, time constraints have played a role in the validation process. The implementation and validation of the proposed solution occurred over a relatively limited time frame. While the results are promising, an extended validation period would provide a more robust assessment of the solution's performance, especially in capturing long-term variations and production dynamics. Therefore, future research should consider longer-term validation studies to further enhance the solution's reliability and effectiveness in real-world industrial scenarios.

References

[1] F. Martínez-Plumed, L. Contreras-Ochando, C. Ferri, J. Hernández-Orallo, M. Kull, N. Lachiche, M. J. Ramírez-Quintana, P. Flach. Crisp-dm twenty years later: From data mining processes to data science trajectories. IEEE Transactions on Knowledge and Data Engineering 33 (2021) 3048–3061. doi:10.1109/TKDE.2019.2962680.

[2] J.-S. Jang. ANFIS: adaptive-network-based fuzzy inference system. IEEE Transactions on Systems, Man, and Cybernetics 23 (1993) 665–685. URL: https://doi.org/10.1109/21.256541. doi:10.1109/21.256541.

[3] Buencafé, Freeze-dried coffee by buencafé, 2015. URL: https://www.youtube.com/watch?v=tvKgz_1oU1Y.

[4] International Coffee Organization, World Coffee Consumption, http://www.ico.org/prices/new-consumption-table.pdf, 2021. [Online; accessed 16 June 2023].

[5] Jacobs Douwe Egberts, Jacobs Douwe Egberts 2022 Annual Report, https://www.jdepeets.com/siteassets/documents/jde-peets-annual-report-2022.pdf, 2022. [Online; accessed 23 June 2023].

[6] O. Gonzalez-Rios, M. Suarez-Quiroz, B. Renaud, M. Barel, B. Guyot, J.-P. Guiraud, S. Schorr-Galindo. Impact of "ecological" post-harvest processing on coffee aroma: Ii. roasted coffee. Journal of Food Composition and Analysis - J FOOD COMPOS ANAL 20 (2007) 297–307. doi:10.1016/j.jfca.2006.12.004.

[7] B. Kitchenham, S. Charters. Guidelines for performing systematic literature reviews in software engineering 2 (2007).

[8] S. Kumar, A. Kar, V. Ilavarasan. Applications of text mining in services management: A systematic literature review. International Journal of Information Management Data Insights 1 (2021) 100008. doi:10.1016/j.jjimei.2021.100008.

[9] H. Yuan, M. Tan, Y. Chen. A model for fishery forecast based on cluster analysis and nonlinear regression (2014) 174–181. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85037370373&doi=10.1142%2f9789814619998_0031&partnerID=40&md5=388b927164412e4ec926e1f6e80d1192https://www.worldscientific.com/doi/abs/10.1142/9789814619998_0031. doi:10.1142/9789814619998_0031, export Date: 8 June 2023.

[10] F. Sabrina, S. Sohail, F. Farid, S. Jahan, F. Ahamed, S. Gordon. An interpretable artificial intelligence based smart agriculture system. Computers, Materials and Continua 72 (2022) 3777–3797. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85127341562&doi=10.32604%2fcmc.2022.026363&partnerID=40&md5=e108fa83bf09e2ec1ea42205cb1001d8https://www.techscience.com/cmc/v72n2/47236. doi:10.32604/cmc.2022.026363, export Date: 8 June 2023.

[11] Harsawardana, B. Samodro, B. Mahesworo, T. Suparyanto, S. Atmaja, B. Pardamean. Maintaining the quality and aroma of coffee with fuzzy logic coffee roasting machine. IOP Conference Series: Earth and Environmental Science 426 (2020) 012148. doi:10.1088/1755-1315/426/1/012148.

[12] A. M. Imammuddien, S. Wirayoga, M. D. Muliono, in: 2022 International Conference on Electrical and Information Technology (IEIT), pp. 420–424. doi:10.1109/IEIT56384.2022.9967854.

[13] Y. E. Isikdemir, G. Erturk, H. Ates, M. O. Tas. Fuzzy inference and machine learning based hvac control system for smart buildings (2022) 116–119. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85146488248&doi=10.1109%2fGEC55014.2022.9987083&partnerID=40&md5=04b4beb3d1b823d8014c06ab070d04a3https://ieeexplore.ieee.org/stampPDF/getPDF.jsp?tp=&arnumber=9987083&ref=. doi:10.1109/GEC55014.2022.9987083, export Date: 8 June 2023.

[14] A. G. Mohapatra, S. K. Lenka. Hybrid decision support system using plsr-fuzzy model for gsm-based site-specific irrigation notification and control in precision agriculture. International Journal of Intelligent Systems Technologies and Applications 15 (2016) 4–18. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84978388612&doi=10.1504%2fIJISTA.2016.076101&partnerID=40&md5=698fc222addc96f95127eabd40237373. doi:10.1504/IJISTA.2016.076101, export Date: 8 June 2023.

[15] P. Patel, Y. Patel, U. Patel, V. Patel, N. Patel, P. Oza, U. Patel. Towards automating irrigation: a fuzzy logic-based water irrigation system using iot and deep learning. Modeling Earth Systems and Environment 8 (2022) 5235–5250. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85133628272&doi=10.1007%2fs40808-022-01452-0&partnerID=40&md5=06c53a28ada2adec2c8d9d90760f0900https://link.springer.com/article/10.1007/s40808-022-01452-0. doi:10.1007/s40808-022-01452-0, export Date: 8 June 2023.

[16] M. Abbaspour-Gilandeh, Y. Abbaspour-Gilandeh. Modelling soil compaction of agricultural soils using fuzzy logic approach and adaptive neuro-fuzzy inference system (anfis) approaches. Modeling Earth Systems and Environment 5 (2019) 13–20. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.

0-85070988796&doi=10.1007%2fs40808-018-0514-1&partnerID=40&md5=30ab8509c30df4522a11eefa8d2e8e8chttps://link.springer.com/article/10.1007/s40808-018-0514-1. doi:10.1007/s40808-018-0514-1, export Date: 8 June 2023.

[17] S. A. Abdul-Wahab, A. S. M. Omer, K. Yetilmezsoy, M. Bahramian. Modelling the clogging of gas turbine filter houses in heavy-duty power generation systems. Mathematical and Computer Modelling of Dynamical Systems 26 (2020) 119–143. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85078860310&doi=10.1080%2f13873954.2020.1713821&partnerID=40&md5=42d6fde31974aae9c45a76ff186f2459https://www.tandfonline.com/doi/full/10.1080/13873954.2020.1713821. doi:10.1080/13873954.2020.1713821, export Date: 8 June 2023 CODEN: MCMSF.

[18] F. Al-Shanableh, M. Bilin, A. Evcil, M. A. Savaş, in: 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), pp. 1–5. URL: https://ieeexplore.ieee.org/stampPDF/getPDF.jsp?tp=&arnumber=9255214&ref=. doi:10.1109/ISMSIT50672.2020.9255214.

[19] K. N. Amrutha, Y. K. Bharath, J. Jayanthi, in: 2019 4th International Conference on Recent Trends on Electronics, Information, Communication Technology (RTEICT), pp. 39–44. URL: https://ieeexplore.ieee.org/stampPDF/getPDF.jsp?tp=&arnumber=9016703&ref=. doi:10.1109/RTEICT46194.2019.9016703.

[20] K. Boma, S. Palizdar. Short time price forecasting for electricity market based on hybrid fuzzy wavelet transform and bacteria foraging algorithm. Journal of Information Systems and Telecommunication 4 (2016) 210–214. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85020479622&doi=10.7508%2fjist.2016.04.001&partnerID=40&md5=bc94ba46773184c8721bd64bc77c5043. doi:10.7508/jist.2016.04.001, export Date: 8 June 2023.

[21] K. T. T. Bui, D. Tien Bui, J. Zou, C. Van Doan, I. Revhaug. A novel hybrid artificial intelligent approach based on neural fuzzy inference model and particle swarm optimization for horizontal displacement modeling of hydropower dam. Neural Computing and Applications 29 (2018) 1495–1506. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84995482286&doi=10.1007%2fs00521-016-2666-0&partnerID=40&md5=81bab482d7100ab6d780ed1baaab3ee4https://link.springer.com/article/10.1007/s00521-016-2666-0. doi:10.1007/s00521-016-2666-0, export Date: 8 June 2023.

[22] A. Choudhary, D. Pandey, S. Bhardwaj. Overview of solar radiation estimation techniques with development of solar radiation model using artificial neural network. Advances in Science, Technology and Engineering Systems 5 (2020) 589–593. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85092896168&doi=10.

25046%2fAJ050469&partnerID=40&md5=f429fb12064748517a6836d39b646f8b. doi:10.25046/AJ050469, export Date: 8 June 2023.

[23] M. El Midaoui, M. Qbadou, K. Mansouri. A fuzzy-based prediction approach for blood delivery using machine learning and genetic algorithm. International Journal of Electrical and Computer Engineering 12 (2022) 1056–1068. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85118996855&doi=10.11591%2fijece.v12i1.pp1056-1068&partnerID=40&md5=c38af2dde75d7d2221515c9974ebb66ahttps://ijece.iaescore.com/index.php/IJECE/article/download/24993/15415. doi:10.11591/ijece.v12i1.pp1056-1068, export Date: 8 June 2023.

[24] G. Ellina, G. Papaschinopoulos, B. K. Papadopoulos. Research of fuzzy implications via fuzzy linear regression in data analysis for a fuzzy model. Journal of Computational Methods in Sciences and Engineering 20 (2020) 879–888. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85092793069&doi=10.3233%2fJCM-194015&partnerID=40&md5=f4b5429fa1c07f4c0785c8d765c953e4https://content.iospress.com/articles/journal-of-computational-methods-in-sciences-and-engineering/jcm194015. doi:10.3233/JCM-194015, export Date: 8 June 2023.

[25] M. Fauziyah, S. Adhisuwignjo, M. Rifai, D. Dewatama, in: IOP Conference Series: Materials Science and Engineering 434 012202, IOP Publishing, 2018. doi:10.1088/1757-899X/434/1/012202.

[26] C. G. Gay, B. O. Bastien. Global temperature fuzzy model as a function of carbon emissions a fuzzy 'regression' from historical data (2014). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85113747537&partnerID=40&md5=e3ebd2fca13222b69ed9a0a8659ad51c, export Date: 8 June 2023.

[27] M. K. Goyal, B. Bharti, J. Quilty, J. Adamowski, A. Pandey. Modeling of daily pan evaporation in sub tropical climates using ann, ls-svr, fuzzy logic, and anfis. Expert Systems with Applications 41 (2014) 5267–5276. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84898464831&doi=10.1016%2fj.eswa.2014.02.047&partnerID=40&md5=632c5d74bb6e90e4f844a2e76f028352https://www.sciencedirect.com/science/article/abs/pii/S0957417414001237?via%3Dihub. doi:10.1016/j.eswa.2014.02.047, export Date: 8 June 2023 CODEN: ESAPE.

[28] M. Gustin, R. S. McLeod, K. J. Lomas. Forecasting indoor temperatures during heatwaves: Do more complex models provide better predictions? 6 (2019) 4243–4250. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85100423766&partnerID=40&md5=f128c2ee042a25b921517ab7d5cdcf77, export Date: 8 June 2023.

[29] A. Khosravi, R. N. N. Koury, L. Machado, J. J. G. Pabon. Prediction of hourly solar radiation in abu musa island using machine learning algorithms. Journal of Cleaner Production 176 (2018) 63–75. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85040707921&doi=10.1016%2fj.jclepro.2017.12.065&partnerID=40&md5=8fc2986f6fc9aa491b1c4895cb009cc9https://www.sciencedirect.com/science/article/abs/pii/S0959652617330007?via%3Dihub. doi:10.1016/j.jclepro.2017.12.065, export Date: 8 June 2023 CODEN: JCROE.

[30] J. Y. Kim. Coffee beans quality prediction using machine learning (2022). URL: https://dx.doi.org/10.2139/ssrn.4024785.

[31] C. E. Lachouri, K. Mansouri, M. M. Lafifi. Greenhouse climate modeling using fuzzy neural network machine learning technique. Revue d'Intelligence Artificielle 36 (2022) 925–930. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85148281985&doi=10.18280%2fria.360614&partnerID=40&md5=56bbc6b68f623fc9d567f55203e4b5dd. doi:10.18280/ria.360614, export Date: 8 June 2023.

[32] T. L. Lam. Low-cost non-contact pcbs temperature monitoring and control in a hot air reflow process based on multiple thermocouples data fusion. IEEE Access 9 (2021) 123566–123574. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85097923350&doi=10.1109%2fACCESS.2020.3036527&partnerID=40&md5=dea9a1c0645903509567f2c2e4e62a9ehttps://ieeexplore.ieee.org/ielx7/6287639/9312710/09250461.pdf?tp=&arnumber=9250461&isnumber=9312710&ref=. doi:10.1109/ACCESS.2020.3036527, export Date: 8 June 2023.

[33] C. K. Leung, J. D. Elias, S. M. Minuk, A. R. R. d. Jesus, A. Cuzzocrea, in: 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–8. URL: https://ieeexplore.ieee.org/stampPDF/getPDF.jsp?tp=&arnumber=9177823&ref=. doi:10.1109/FUZZ48607.2020.9177823.

[34] S. Li, in: 2019 IEEE International Conference on Mechatronics and Automation (ICMA), pp. 421–426. URL: https://ieeexplore.ieee.org/stampPDF/getPDF.jsp?tp=&arnumber=8816226&ref=. doi:10.1109/ICMA.2019.8816226.

[35] J. Liang, X. Liu, K. Liao. Soil moisture retrieval using uwb echoes via fuzzy logic and machine learning. IEEE Internet of Things Journal 5 (2018) 3344–3352. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85031780207&doi=10.1109%2fJIOT.2017.2760338&partnerID=40&md5=f99362b30797e9d6c372b46376f0c44dhttps://ieeexplore.ieee.org/stampPDF/getPDF.jsp?tp=&arnumber=8062797&ref=. doi:10.1109/JIOT.2017.2760338, export Date: 8 June 2023.

[36] D. M. Minhas, R. R. Khalid, G. Frey. Short term load forecasting using hybrid adaptive fuzzy neural system: The performance evaluation (2017) 468–473. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85028639108&doi=10.1109%2fPowerAfrica.2017.7991270&partnerID=40&md5=142d7cf106f315e4692f12e7f09dee38https://ieeexplore.ieee.org/stampPDF/getPDF.jsp?tp=&arnumber=7991270&ref=. doi:10.1109/PowerAfrica.2017.7991270, export Date: 8 June 2023.

[37] F. Mirzaei, M. Delavar, I. Alzoubi, B. Nadjar Arrabi. Modeling and predict environmental indicators for land leveling using adaptive neuro-fuzzy inference system (anfis), and regression. International Journal of Energy Sector Management 12 (2018) 484–506. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85050918393&doi=10.1108%2fIJESM-02-2017-0003&partnerID=40&md5=2e585f3a0094719381d28d56552aa33e. doi:10.1108/IJESM-02-2017-0003, export Date: 8 June 2023.

[38] S. K. Mousavi Mashhadi, H. Yadollahi, A. Marvian Mashhad. Design and manufacture of tds measurement and control system for water purification in reverse osmosis by pid fuzzy logic controller with the ability to compensate effects of temperature on measurement. Turkish Journal of Electrical Engineering and Computer Sciences 24 (2016) 2589–2608. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84974728170&doi=10.3906%2felk-1402-65&partnerID=40&md5=797a02f85360292e772f24ce438916e3https://journals.tubitak.gov.tr/cgi/viewcontent.cgi?article=2627&context=elektrik. doi:10.3906/elk-1402-65, export Date: 8 June 2023.

[39] H. Neog, P. E. Dutta, N. Medhi. Health condition prediction and covid risk detection using healthcare 4.0 techniques. Smart Health 26 (2022). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85139275493&doi=10.1016%2fj.smhl.2022.100322&partnerID=40&md5=49ed79cf1aa3c1179aadaacf003358dchttps://www.sciencedirect.com/science/article/abs/pii/S2352648322000563?via%3Dihub. doi:10.1016/j.smhl.2022.100322, export Date: 8 June 2023.

[40] A. H. Orta, I. Kayabasi, M. Tunc, in: 2018 IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS), pp. 1–6. URL: https://ieeexplore.ieee.org/stampPDF/getPDF.jsp?tp=&arnumber=8440350&ref=. doi:10.1109/PMAPS.2018.8440350.

[41] V. K. Patil, V. R. Pawar, in: 2022 International Conference on Signal and Information Processing (IConSIP), pp. 1–6. URL: https://ieeexplore.ieee.org/stampPDF/getPDF.jsp?tp=&arnumber=10007500&ref=. doi:10.1109/ICoNSIP49665.2022.10007500.

[42] B. Petković, D. Petković, B. Kuzman, M. Milovančević, K. Wakil, L. S. Ho, K. Jermsittiparsert. Neuro-fuzzy estimation of reference crop evapotranspiration by neuro

fuzzy logic based on weather conditions. Computers and Electronics in Agriculture 173 (2020). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85083201464&doi=10.1016%2fj.compag.2020.105358&partnerID=40&md5=405a563e6ce84807380784f296211174https://www.sciencedirect.com/science/article/abs/pii/S0168169920302192?via%3Dihub. doi:10.1016/j.compag.2020.105358, export Date: 8 June 2023 CODEN: CEAGE.

[43] M. R. C. Qazani, V. Pourmostaghimi, M. Moayyedian, S. Pedrammehr. Estimation of tool–chip contact length using optimized machine learning in orthogonal cutting. Engineering Applications of Artificial Intelligence 114 (2022). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85133362848&doi=10.1016%2fj.engappai.2022.105118&partnerID=40&md5=802de691e65c361698ca158fdc8970bchttps://www.sciencedirect.com/science/article/abs/pii/S0952197622002524?via%3Dihub. doi:10.1016/j.engappai.2022.105118, export Date: 8 June 2023 CODEN: EAAIE.

[44] J. Refonaa, M. Lakshmi. Remote sensing based rain fall prediction using big data assisted integrated routing framework. Journal of Ambient Intelligence and Humanized Computing (2021). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85098778828&doi=10.1007%2fs12652-020-02726-0&partnerID=40&md5=7500309a1a9ef1ebac1e7dacc54a3abdhttps://link.springer.com/article/10.1007/s12652-020-02726-0. doi:10.1007/s12652-020-02726-0, export Date: 8 June 2023.

[45] S. Sharma, R. K. Agrawal, M. M. Tripathi, in: 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), pp. 165–169. URL: https://ieeexplore.ieee.org/stampPDF/getPDF.jsp?tp=&arnumber=9076429&ref=. doi:10.1109/ICCMC48092.2020.ICCMC-00033.

[46] A. Shastry, H. A. Sanjay, M. Hegde. A parameter based anfis model for crop yield prediction (2015) 253–257. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84941985246&doi=10.1109%2fIADCC.2015.7154708&partnerID=40&md5=3ac70cda2061fba1d0028b0ce9203f06https://ieeexplore.ieee.org/stampPDF/getPDF.jsp?tp=&arnumber=7154708&ref=. doi:10.1109/IADCC.2015.7154708, export Date: 8 June 2023.

[47] J. M. Siqueira, T. A. Paço, J. C. Silvestre, F. L. Santos, A. O. Falcão, L. S. Pereira. Generating fuzzy rules by learning from olive tree transpiration measurement - an algorithm to automatize granier sap flow data analysis. Computers and Electronics in Agriculture 101 (2014) 1–10. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84890715465&doi=10.1016%2fj.compag.2013.11.013&partnerID=40&md5=aa9d07fb12180ce85cdad5349275fbf3https://www.sciencedirect.com/science/article/abs/pii/S0168169913002834?

via%3Dihub. doi:10.1016/j.compag.2013.11.013, export Date: 8 June 2023 CODEN: CEAGE.

[48] Q. T. T. Tran, K. Davies, L. Roose, in: 2021 IEEE/IAS 57th Industrial and Commercial Power Systems Technical Conference (ICPS), pp. 1–5. URL: https://ieeexplore.ieee.org/stampPDF/getPDF.jsp?tp=&arnumber=9416618&ref=. doi:10.1109/ICPS51807.2021.9416618.

[49] V. Vivekanandhan, S. Sakthivel, M. Manikandan. Adaptive neuro fuzzy inference system to enhance the classification performance in smart irrigation system. Computational Intelligence 38 (2022) 308–322. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85120558157&doi=10.1111%2fcoin.12492&partnerID=40&md5=59cb39084c74e1227176c80017a1f3f8https://onlinelibrary.wiley.com/doi/pdfdirect/10.1111/coin.12492?download=true. doi:10.1111/coin.12492, export Date: 8 June 2023 CODEN: COMIE.

[50] G. Zhang, S. S. Band, S. Ardabili, K. W. Chau, A. Mosavi. Integration of neural network and fuzzy logic decision making compared with bilayered neural network in the simulation of daily dew point temperature. Engineering Applications of Computational Fluid Mechanics 16 (2022) 713–723. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85125952053&doi=10.1080%2f19942060.2022.2043187&partnerID=40&md5=ed7b8641743fed70f237e77b70fa699fhttps://www.tandfonline.com/doi/pdf/10.1080/19942060.2022.2043187. doi:10.1080/19942060.2022.2043187, export Date: 8 June 2023.

[51] M. Bohanec, M. Robnik-Šikonja, M. K. Borštnar. Organizational learning supported by machine learning models coupled with general explanation methods: A case of b2b sales forecasting. Organizacija 50 (2017) 217–233. URL: https://doi.org/10.1515/orga-2017-0020. doi:10.1515/orga-2017-0020.

[52] C. Schröer, F. Kruse, J. M. Gómez. A systematic literature review on applying CRISP-DM process model. Procedia Computer Science 181 (2021) 526–534. URL: https://doi.org/10.1016/j.procs.2021.01.199. doi:10.1016/j.procs.2021.01.199.

[53] P. Tziachris, M. Nikou, V. Aschonitis, A. Kallioras, K. Sachsamanoglou, M. D. Fidelibus, E. Tziritis. Spatial or random cross-validation? the effect of resampling methods in predicting groundwater salinity with machine learning in mediterranean region. Water 15 (2023) 2278. URL: https://doi.org/10.3390/w15122278. doi:10.3390/w15122278.

[54] D. Kamelesun, R. Saranya, P. Kathiravan, A benchmark study by using various machine learning models for predicting covid-19 trends, 2023. URL: https://arxiv.org/abs/2301.11257. doi:10.48550/ARXIV.2301.11257.

[55] D. Bertsimas, G. Lukin, L. Mingardi, O. Nohadani, A. Orfanoudaki, B. Stellato, H. Wiberg, S. Gonzalez-Garcia, C. L. Parra-Calderón, K. Robinson, M. Schneider, B. Stein, A. Es-

tirado, L. a Beccara, R. Canino, M. D. Bello, F. Pezzetti, A. P. and. COVID-19 mortality risk assessment: An international multi-center study. PLOS ONE 15 (2020) e0243262. URL: https://doi.org/10.1371/journal.pone.0243262. doi:10.1371/journal.pone.0243262.

[56] R. I. Hamilton, P. N. Papadopoulos. Using shap values and machine learning to understand trends in the transient stability limit. IEEE Transactions on Power Systems (2023) 1–12. doi:10.1109/TPWRS.2023.3248941.

[57] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, 2017. URL: https://arxiv.org/abs/1705.07874. doi:10.48550/ARXIV.1705.07874.

[58] D. Chicco, M. J. Warrens, G. Jurman. The coefficient of determination r-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science 7 (2021) e623. URL: https://doi.org/10.7717/peerj-cs.623. doi:10.7717/peerj-cs.623.

[59] J. Kaliappan, K. Srinivasan, S. M. Qaisar, K. Sundararajan, C.-Y. Chang, S. C. Performance evaluation of regression models for the prediction of the COVID-19 reproduction rate. Frontiers in Public Health 9 (2021). URL: https://doi.org/10.3389/fpubh.2021.729795. doi:10.3389/fpubh.2021.729795.

[60] D. Maulud, A. Mohsin Abdulazeez. A review on linear regression comprehensive in machine learning. Journal of Applied Science and Technology Trends 1 (2020) 140–147. doi:10.38094/jastt1457.

[61] C. Yu, W. Yao. Robust linear regression: A review and comparison. Communications in Statistics - Simulation and Computation 46 (2016) 6261–6282. URL: https://doi.org/10.1080/03610918.2016.1202271. doi:10.1080/03610918.2016.1202271.

[62] E. N. Pratama, E. Suwarni, M. A. Handayani. Effect of job satisfaction and organizational commitment on turnover intention with person organization fit as moderator variable. APTISI Transactions on Management (ATM) 6 (2022) 74–82. URL: https://doi.org/10.33050/atm.v6i1.1722. doi:10.33050/atm.v6i1.1722.

[63] S. Etemadi, M. Khashei. Etemadi multiple linear regression. Measurement 186 (2021) 110080. URL: https://doi.org/10.1016/j.measurement.2021.110080. doi:10.1016/j.measurement.2021.110080.

[64] H. Han, K. J. Dawson. Applying elastic-net regression to identify the best models predicting changes in civic purpose during the emerging adulthood. Journal of Adolescence 93 (2021) 20–27. URL: https://doi.org/10.1016/j.adolescence.2021.09.011. doi:10.1016/j.adolescence.2021.09.011.

[65] J. M. K. Aheto, H. O. Duah, P. Agbadi, E. K. Nakua. A predictive model, and predictors of under-five child malaria prevalence in ghana: How do LASSO, ridge and

elastic net regression approaches compare? Preventive Medicine Reports 23 (2021) 101475. URL: https://doi.org/10.1016/j.pmedr.2021.101475. doi:10.1016/j.pmedr.2021.101475.

[66] A. Lakshmanarao, M. Kumar, K. Ratnakar, Y. Satwika, in: 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), pp. 423–426. doi:10.1109/ICAAIC56838.2023.10141462.

[67] M. Maabreh, G. Almasabha. Machine learning regression algorithms for shear strength prediction of SFRC-DBs: Performance evaluation and comparisons. Arabian Journal for Science and Engineering (2023). URL: https://doi.org/10.1007/s13369-023-08176-y. doi:10.1007/s13369-023-08176-y.

[68] M. R. Keyvanpour, M. B. Shirzad, in: Application of Machine Learning in Agriculture, Elsevier, 2022, pp. 283–305. URL: https://doi.org/10.1016/b978-0-323-90550-3.00011-4. doi:10.1016/b978-0-323-90550-3.00011-4.

[69] J. Wei, X. He. Support vector regression model with variant tolerance. Measurement and Control (2023). URL: https://doi.org/10.1177/00202940231180620. doi:10.1177/00202940231180620.

[70] T. C. Au. Random forests, decision trees, and categorical predictors: The "absent levels" problem (2017). URL: https://arxiv.org/abs/1706.03492. doi:10.48550/ARXIV.1706.03492.

[71] E. Carrizosa, C. Molero-Río, D. R. Morales. Mathematical optimization in classification and regression trees. TOP 29 (2021) 5–33. URL: https://doi.org/10.1007/s11750-021-00594-1. doi:10.1007/s11750-021-00594-1.

[72] R. Abedi, R. Costache, H. Shafizadeh-Moghadam, Q. B. Pham. Flash-flood susceptibility mapping based on XGBoost, random forest and boosted regression trees. Geocarto International 37 (2021) 5479–5496. URL: https://doi.org/10.1080/10106049.2021.1920636. doi:10.1080/10106049.2021.1920636.

[73] E. Lasso, D. C. Corrales, J. Avelino, E. de Melo Virginio Filho, J. C. Corrales. Discovering weather periods and crop properties favorable for coffee rust incidence from feature selection approaches. Computers and Electronics in Agriculture 176 (2020) 105640. URL: https://doi.org/10.1016/j.compag.2020.105640. doi:10.1016/j.compag.2020.105640.

[74] G. Biau, B. Cadre, Optimization by gradient boosting, 2017. URL: https://arxiv.org/abs/1707.05023. doi:10.48550/ARXIV.1707.05023.

[75] T. Chen, C. Guestrin, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016. URL: https://doi.org/10.1145%2F2939672.2939785. doi:10.1145/2939672.2939785.

[76] X. Yao, X. Fu, C. Zong. Short-term load forecasting method based on feature preference strategy and LightGBM-XGboost. IEEE Access 10 (2022) 75257–75268. URL: https://doi.org/10.1109%2Faccess.2022.3192011. doi:10.1109/access.2022.3192011.

[77] A. Shehadeh, O. Alshboul, R. E. A. Mamlook, O. Hamedat. Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression. Automation in Construction 129 (2021) 103827. URL: https://doi.org/10.1016%2Fj.autcon.2021.103827. doi:10.1016/j.autcon.2021.103827.

[78] C. Mao, W. Xu, Y. Huang, X. Zhang, N. Zheng, X. Zhang. Investigation of passengers' perceived transfer distance in urban rail transit stations using XGBoost and SHAP. Sustainability 15 (2023) 7744. URL: https://doi.org/10.3390%2Fsu15107744. doi:10.3390/su15107744.

[79] L. Teng, Gradient boosting-based numerical methods for high-dimensional backward stochastic differential equations, 2021. URL: https://arxiv.org/abs/2107.06673. doi:10.48550/ARXIV.2107.06673.

[80] L. Zadeh. Fuzzy sets. Information and Control 8 (1965) 338–353. URL: https://www.sciencedirect.com/science/article/pii/S001999586590241X. doi:https://doi.org/10.1016/S0019-9958(65)90241-X.

[81] C. Fatih. Forecasting models based on fuzzy logic: An application on international coffee prices. Econometrics 26 (2022) 1–16. URL: https://doi.org/10.15611/eada.2022.4.01. doi:10.15611/eada.2022.4.01.

[82] A. Jain, A. Sharma. Membership function formulation methods for fuzzy logic systems: A comprehensive review. Journal of Critical Reviews 7 (2020) 8717–8733.

[83] E. Hadianto, D. Amanda, D. Hindarto, A. Makmur, H. Santoso. Design and development of coffee machine control system using fuzzy logic. Sinkron 8 (2023) 130–138. URL: https://doi.org/10.33395/sinkron.v8i1.11917. doi:10.33395/sinkron.v8i1.11917.

[84] E. Mamdani, S. Assilian. An experiment in linguistic synthesis with a fuzzy logic controller. International Journal of Man-Machine Studies 7 (1975) 1–13. URL: https://www.sciencedirect.com/science/article/pii/S0020737375800022. doi:https://doi.org/10.1016/S0020-7373(75)80002-2.

[85] T. Takagi, M. Sugeno. Fuzzy identification of systems and its applications to modeling and control. IEEE Transactions on Systems, Man, and Cybernetics SMC-15 (1985) 116–132. doi:10.1109/TSMC.1985.6313399.

[86] Y. Tsukamoto, in: D. Dubois, H. Prade, R. R. Yager (Eds.), Readings in Fuzzy Sets for Intelligent Systems, Morgan Kaufmann, 1993, pp. 523–529. URL: https:

//www.sciencedirect.com/science/article/pii/B9781483214504500559. doi:https://doi.org/10.1016/B978-1-4832-1450-4.50055-9.

[87] C. Klaidaeng, S. Chudjuarjeen, C. Pomsen, P. Charoenwiangnuea. Prediction of roasted coffee bean level from a coffee house-ware using fuzzy logic. Materials Today: Proceedings (2023). URL: https://doi.org/10.1016/j.matpr.2023.04.524. doi:10.1016/j.matpr.2023.04.524.

[88] S. Sengupta, S. Basak, P. Saikia, S. Paul, V. Tsalavoutis, F. Atiah, V. Ravi, A. Peters. A review of deep learning with special emphasis on architectures, applications and recent trends. Knowledge-Based Systems 194 (2020) 105596. URL: https://doi.org/10.1016/j.knosys.2020.105596. doi:10.1016/j.knosys.2020.105596.

[89] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006. URL: /bib/bishop/Bishop2006/Pattern-Recognition-and-Machine-Learning-Christophe-M-Bishop.pdf,/bib/bishop/Bishop2006/978-0-387-31073-2_sm.pdf,https://www.microsoft.com/en-us/research/people/cmbishop/#!prml-book.

[90] Y. LeCun, Y. Bengio, G. Hinton. Deep learning. Nature 521 (2015) 436–444. URL: https://doi.org/10.1038/nature14539. doi:10.1038/nature14539.

[91] E. M. T. Caballero, A. M. R. Duke, in: 2020 5th International Conference on Control and Robotics Engineering (ICCRE), pp. 246–251. doi:10.1109/ICCRE49379.2020.9096435.

[92] O. A. M. López, A. M. López, J. Crossa, in: Multivariate Statistical Machine Learning Methods for Genomic Prediction, Springer International Publishing, 2022, pp. 379–425. URL: https://doi.org/10.1007/978-3-030-89010-0_10. doi:10.1007/978-3-030-89010-0_10.

[93] T. Chen, C. Shang, P. Su, E. Keravnou-Papailiou, Y. Zhao, G. Antoniou, Q. Shen. A decision tree-initialised neuro-fuzzy approach for clinical decision support. Artificial Intelligence in Medicine 111 (2021) 101986. URL: https://doi.org/10.1016/j.artmed.2020.101986. doi:10.1016/j.artmed.2020.101986.

[94] A. F. Schmidt, C. Finan. Linear regression and the normality assumption. Journal of Clinical Epidemiology 98 (2018) 146–151. URL: https://www.sciencedirect.com/science/article/pii/S0895435617304857. doi:https://doi.org/10.1016/j.jclinepi.2017.12.006.

[95] N. M. Razali, Y. B. Wah. URL: https://api.semanticscholar.org/CorpusID:18639594.

[96] E. González-Estrada, W. Cosmes. Shapiro–wilk test for skew normal distributions based on data transformations. Journal of Statistical Computation and Simulation 89 (2019) 3258–3272. URL: https://doi.org/10.1080/00949655.2019.1658763. doi:10.1080/00949655.2019.1658763.

[97] C. J. Leggetter, P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. Computer speech & language 9 (1995) 171–185.

[98] S. Chowdhury, Y. Lin, B. Liaw, L. Kerby, in: 2022 International Conference on Intelligent Data Science Technologies and Applications (IDSTA), pp. 17–25. doi:10.1109/IDSTA55301.2022.9923169.

[99] I. A. Talin, M. H. Abid, M. A.-M. Khan, S.-H. Kee, A.-A. Nahid. Finding the influential clinical traits that impact on the diagnosis of heart disease using statistical and machine-learning techniques. Scientific Reports 12 (2022). URL: https://doi.org/10.1038/s41598-022-24633-4. doi:10.1038/s41598-022-24633-4.

[100] X. Liu, Z. Yan, F. Leng, Y. Bao, Y. Huang. Machine learning predictive model for electronic slurries for smart grids. Frontiers in Energy Research 10 (2023). URL: https://doi.org/10.3389/fenrg.2022.1031118. doi:10.3389/fenrg.2022.1031118.

[101] S. Senthilnathan. Usefulness of correlation analysis. SSRN Electronic Journal (2019). URL: https://doi.org/10.2139/ssrn.3416918. doi:10.2139/ssrn.3416918.

[102] N. Shrestha. Detecting multicollinearity in regression analysis. American Journal of Applied Mathematics and Statistics 8 (2020) 39–42. URL: https://doi.org/10.12691/ajams-8-2-1. doi:10.12691/ajams-8-2-1.

[103] E. Uman, M. Colonna-Dashwood, L. Colonna-Dashwood, M. Perger, C. Klatt, S. Leighton, B. Miller, K. T. Butler, B. C. Melot, R. W. Speirs, C. H. Hendon. The effect of bean origin and temperature on grinding roasted coffee. Scientific Reports 6 (2016). URL: https://doi.org/10.1038/srep24483. doi:10.1038/srep24483.

[104] S. Ramalingam, K. Baskaran. An efficient data prediction model using hybrid harris hawk optimization with random forest algorithm in wireless sensor network. Journal of Intelligent &amp Fuzzy Systems 40 (2021) 5171–5195. URL: https://doi.org/10.3233/jifs-201921. doi:10.3233/jifs-201921.

[105] S. Spielberg, A. Tulsyan, N. P. Lawrence, P. D. Loewen, R. B. Gopaluni. Deep reinforcement learning for process control: A primer for beginners (2020). URL: https://arxiv.org/abs/2004.05490. doi:10.48550/ARXIV.2004.05490.

[106] R. de Rezende Faria, B. D. O. Capron, A. R. Secchi, M. B. de Souza. Where reinforcement learning meets process control: Review and guidelines. Processes 10 (2022) 2311. URL: https://doi.org/10.3390/pr10112311. doi:10.3390/pr10112311.

[107] H. Yoo, H. E. Byun, D. Han, J. H. Lee. Reinforcement learning for batch process control: Review and perspectives. Annual Reviews in Control 52 (2021) 108–119. URL: https://www.sciencedirect.com/science/article/pii/S136757882100081X. doi:https://doi.org/10.1016/j.arcontrol.2021.10.006.

A

APPENDIX A: FREEZE-DRIED COFFEE PRODUCTION PROCESS

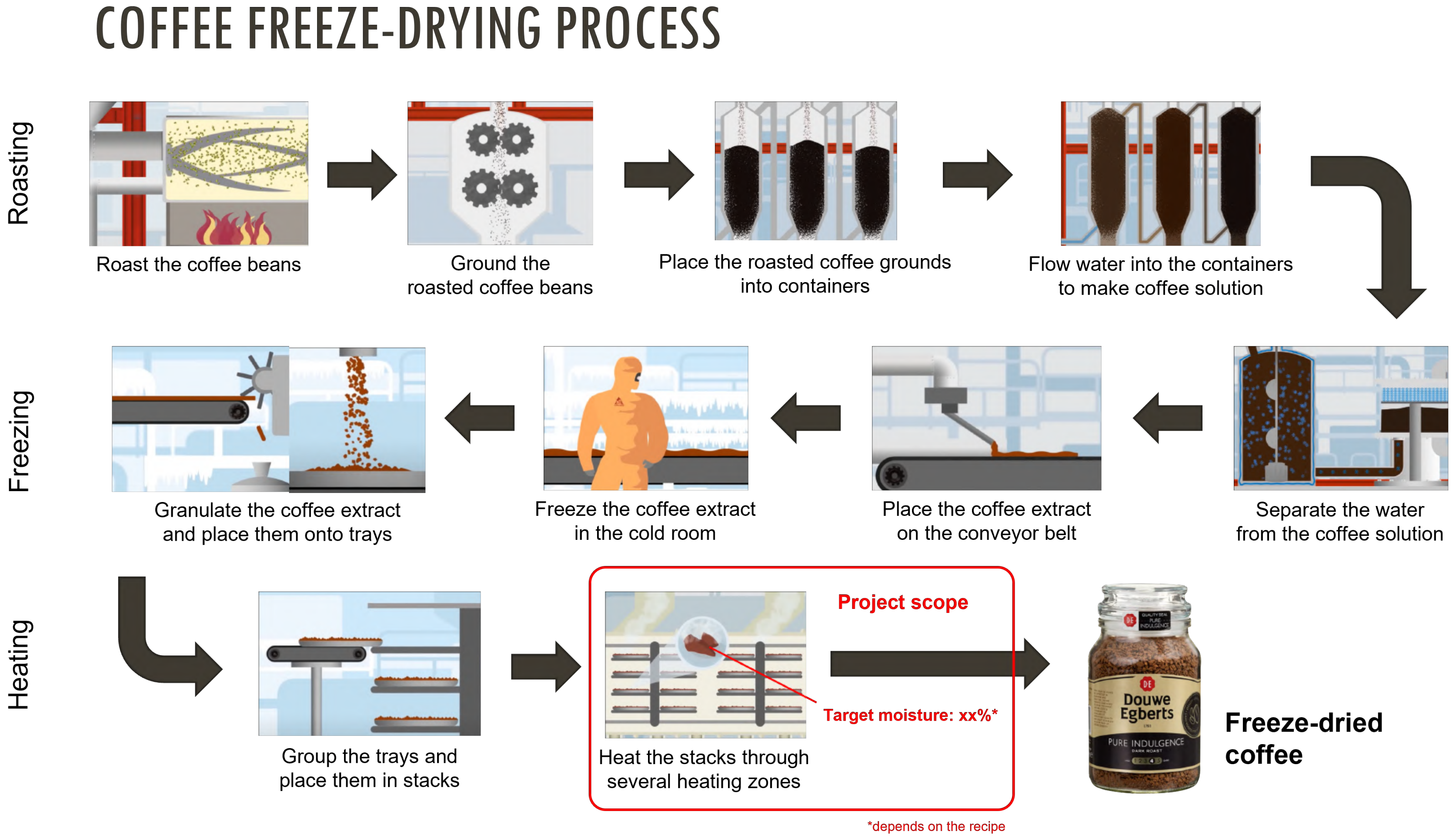

COFFEE FREEZE-DRYING PROCESS
Roasting
Roast the coffee beans
Ground the roasted coffee beans
Place the roasted coffee grounds into containers
Flow water into the containers to make coffee solution
Freezing
Separate the water from the coffee solution
Place the coffee extract on the conveyor belt
Freeze the coffee extract in the cold room
Granulate the coffee extract and place them onto trays
Heating
Group the trays and place them in stacks
Project scope
Heat the stacks through several heating zones
Target moisture: xx%*
Freeze-dried coffee
*depends on the recipe

Figure A.1: Freeze-Dried Coffee Production Process (Illustration by [3])

B

Appendix B: Qualitative Analysis of the Systematic Literature Review

Table B.1: Qualitative Analysis of the Literature

Literature	Main Purpose	Main Findings/Results	Other Remarks
M. Abbaspour-Gilandeh and Y. Abbaspour-Gilandeh (2019) [16]	"The aim of this study was predicting arable soils cone index values by effective parameters on the soil cone index, including bulk density, soil moisture content and soil electrical conductivity by using Fuzzy and neuro-fuzzy systems."	In comparison with regression models, ANFIS model has high accuracy and can be used to estimate the soil cone index in agricultural land.	–
S. A. Abdul-Wahab, A. S. M. Omer, K. Yetilmezsoy and M. Bahramian (2020) [17]	"A prognostic approach based on a MISO (multiple inputs and single output) fuzzy logic model was introduced to estimate the pressure difference across a gas turbine (GT) filter house in a heavy-duty power generation system."	The results revealed that the proposed fuzzy logic model produced very small deviations and showed a superior predictive performance than the conventional multiple regression methodology.	Considering the usefulness of an artificial intelligence–based modelling scheme, a MIMO (multiple inputs and single output) fuzzy logic-based model (introduction of additional model components and specification of new membership functions with different levels) will be useful to improve the proposed strategy on the GT filter houses. It is also needed to provide additional experimental data from the literature for the validity of the implemented deep learning strategy.
F. Al-Shanableh, M. Bilin, A. Evcil and M. A. Savaş (2020) [18]	"a fuzzy logic (FL) model and a multilinear regression model (MLR) for the prediction of jojoba oil yield was developed for the cases where the optimum conditions could not be attained easily in practice."	It was noted that the proposed FL model can provide high accuracy and reliability for predicting the oil yield as compared to the linear mathematical MLR models.	–
K. N. Amrutha, Y. K. Bharath and J. Jayanthi (2019) [19]	"In this paper, a model has been proposed to obtain the performance deterioration of Turboprop engine."	Both ANN and Fuzzy systems are good prediction models and by comparing the error performance values, it can be seen that ANN model gives slightly better performance.	,–
K. Boma and S. Palizdar (2016) [20]	"In this study, linear prediction methods and neural networks and fuzzy logic have been studied and emulated. An optimized fuzzy-wavelet prediction method is proposed to predict the price of electricity."	The use of fuzzy logic-wavelet forecasting method resulted in an improved performance, compared with that of fuzzy logic forecasting method. Also choosing two different types of filters; low-pass and high-pass, in the wavelet transform, increased the efficiency of the predictor in the fuzzy prediction method.	Fuzzy-wavelet method has a higher computational volume due to the use of wavelet transform as well as double use of fuzzy prediction.
K. T. T. Bui, D. Tien Bui, J. Zou, C. Van Doan and I. Revhaug (2018) [21]	"This paper proposes a novel hybrid artificial intelligent approach, namely swarm optimized neural fuzzy inference system (SONFIS), for modeling and forecasting of the horizontal displacement of hydropower dams."	High performance of the SONFIS model, both on the training and validation datasets, implies that the SONFIS has successfully modeled a typical complex nonlinear problem of hydropower dam displacement. Overall, the SONFIS model outperforms the five benchmark methods.	The high performance of the proposed method indicates that the selection, processing, and coding of the input variables must be carried out accordingly.

A. Choudhary, D. Pandey and S. Bhardwaj (2020) [22]	"This paper is framed to briefly provide the idea behind different solar radiation estimation models with the methodology used. Soft computing-based models are mainly analyzed here."	The developed ANN-based Global Solar Irradiance Estimation Model has better results for training, testing, validation, and all compared to the current solar radiation estimation models.	This model may be used for the estimation of Global Solar Irradiance for other stations also.
M. El Midaoui, M. Qbadou and K. Mansouri (2022) [23]	"The paper presents routing and scheduling system based on artificial intelligence to deliver blood from the blood-banks to hospitals based on single blood bank and multiple blood banks with respect of the vehicle capacity used to deliver the blood and creating the shortest path. The next section consists on solution for predicting the blood needs for each hospital based on transfusion history using machine learning and fuzzy logic."	The adoption of fuzzification data to get a clear idea about demand with linguistic values, train a model by using transfer learning to predict a future need in blood components by using historical data.	The combination of fuzzy logic and machine learning besides genetic algorithms is rarely used in the resolution of similar matters. Also, even if this solution appears to be specific to the health-care and blood delivery sector, this work can concern further fields like agricultural and food-processing. It should be noticed that the model can be improved by considering applying parallelization for GA and taking into consideration reverse logistics to recover the unused/expired blood.
G. Ellina, G. Papaschinopoulos and B. K. Papadopoulos (2020) [24]	"Our purpose is to investigate some of the factors responsible for eutrophication (water temperature, nitrates, total phosphorus, Secchi depth, chlorophyll-a) using fuzzy logic. In this paper, we propose a method of evaluating fuzzy implications constructing triangular fuzzy numbers for all of the studied factors coming from statistical data."	Fuzzy logic can be used as a powerful tool in categorizing environmental status and describing multifaceted changes. The main advantage of this tool is the ability to unite many kinds of perceptions by offering stability between social, economic and biological impacts.	–
M. Fauziyah, S. Adhisuwignjo, M. Rifai and D. Dewatama (2018) [25]	"The aim of this research is to control the temperature of roasting process to be stable. This paper presents the development of the coffee roaster machine where the temperature is controlled using fuzzy logic control algorithm."	The Error Steady State gained is 0.52% of the results and 1.85% of the results using fuzzy logic controller does not experience an error that is too large so that the temperature can be properly maintained.	PLC (Programmable Logic Controller) and HMI (Human Machine Interaction) were implemented.
C. G. Gay and B. O. Bastien (2014) [26]	"In this study we approach the problem of correlating global mean temperature with Carbon emissions using statistical analysis using fuzzy logic analysis and inference systems, which is a pioneer method in climate modelling."	The fuzzy model created can relate the change in mean global temperature with the carbon emissions. Thanks to it fuzziness allow us to involve variables with high uncertainty, such as measurements of annually emitted Carbon or Atmospheric CO2 concentration.	This fuzzy model will be very useful to project future temperatures based on possible values of emissions, due to the uncertainty nature of the problem.

M. K. Goyal, B. Bharti, J. Quilty, J. Adamowski and A. Pandey (2014) [27]	"This paper investigates the abilities of Artificial Neural Networks (ANN), Least Squares – Support Vector Regression (LS-SVR), Fuzzy Logic, and Adaptive Neuro-Fuzzy Inference System (ANFIS) techniques to improve the accuracy of daily pan evaporation estimation in subtropical climates."	The Fuzzy Logic and LS-SVR approaches can be employed successfully in modeling the daily evaporation process from the available climatic data. In addition, results showed that the machine learning models outperform the traditional HGS and SS empirical methods.	Using only the minimum and maximum temperatures as inputs gives poor estimates for all machine learning models. Future research will explore multiple watersheds, include additional new state-of-the-art machine learning methods, employ wavelets as a pre-processing method, create ensembles of models, and will also look to develop various forms of uncertainty assessment for model predictions.
M. Gustin, R. S. McLeod and K. J. Lomas (2019) [28]	"A novel application of semi-parametric Generalized Additive Models (GAMs) was developed to forecast elevated indoor temperatures."	More complex non-linear models do not necessarily produce better forecasts and that particular attention should be given to the use of GAMs when predicting out-of-range.	There will always be limited data at the lower and upper ranges of the independent variables, which engenders increasing uncertainty when forecasting beyond the ranges for which the models were originally trained, with errors that are likely to amplify at longer forecasting horizons.
Harsawardana, B. Samodro, B. Mahesworo, T. Suparyanto, S. Atmaja and B. Pardamean (2020) [11]	"In this study, we proposed a temperature control system based on fuzzy logic for coffee roaster machine. The system controls the temperature in accordance with the demanded roasting level."	A prototype was developed which has thermal camera was mounted inside the roasting chamber to monitor the coffee beans temperature uniformity. Inside the chamber, there is a stirring mechanism with Pulse Width Modulation (PWM) motor to stir the coffee beans.	There are no discussions on the real results from the prototype.
A. M. Imammuddien, S. Wirayoga and M. D. Muliono (2022) [12]	"This study aims to make a detection tool for roasting maturity levels in coffee beans by utilizing the frequency value obtained from the RC oscillator. To increase the accuracy of the classification, a color sensor is added to determine the color intensity of the coffee beans. The capacitance value and the value of the color sensor from each roasting classification are entered into fuzzy logic and are expected to be able to classify the type of roasting of coffee beans accurately."	The results of testing the system with Soegeno's fuzzy logic showed the appropriate results between the input values from the color sensor and the oscillator circuit to the results of the classification of the maturity level of dampit and kawi coffee beans.	In the defuzzification test, it produced fuzzy output values from several frequency input values and color sensors based on the rules on the fuzzy inference system. Mapping was done in the defuzzification process by grouping fuzzy sets into firm sets.
Y. E. Isikdemir, G. Erturk, H. Ates and M. O. Tas (2022) [13]	"In this study, a fuzzy inference and machine learning based HVAC control system is proposed that is aware of the condition change and automatically adjusts the optimal conditions for the building occupants."	Among the multiple machine learning algorithms, Random Forest yields better performance to estimate air quality index. A four-input, one-output with 81 rules FIS is utilized for temperature control. According to the simulation result, there is no overshoot is observed while the temperature reaches a steady state.	Future work involves build web-based user interface and backend server to communicate the sensors via MQTT protocols, where the proposed algorithm can be tracked and configured remotely.

A. Khosravi, R. N. N. Koury, L. Machado and J. J. G. Pabon (2018) [29]	"The main contribution of this study is to developed machine learning algorithms in order to predict the hourly solar radiation in two separate networks. For the first time, in the current study, three types of ANFIS model are developed to predict the time-series hourly solar radiation."	For this model, MFs were determined as gaussmf for the inputs and output and the Mamdani model as FIS structure. The number of MFs are obtained through a trial and error process. The best performance was achieved by considering the four MFs for each input and target. The SVR model and MLFFNN have the maximum efficiency for forecasting the solar radiation.	–
J. Y. Kim (2022) [30]	"This study introduces the machine learning model, random forest, to predict coffee quality."	The random forest model achieved 86.6% and 84.1% accuracy on the train set and the test set, respectively. It also showed an F1 score of 61.7% on the test set. Category two defects, growing altitude, and bag weight were found to be important variables in predicting the quality of coffee.	–
C. E. Lachouri, K. Mansouri and M. M. Lafifi (2022) [31]	"This paper proposes an adaptive system based on artificial neural networks technique embedded with fuzzy logic technique calls Adaptive Neuro Fuzzy Inference System (ANFIS) to predict air humidity, air temperature, internal radiation, and CO2 concentration while the seeds grow, in order to produce favorable greenhouse climate conditions."	ANFIS is an accurate and efficient prediction method for greenhouse climates. This system can yield accuracy as high as 98% in all of the four components when trained with the least square algorithm and back propagation.	Chances of error in predicting internal climate values in combination with gaussian and sigmoidal membership function is just 2%. Accuracy with the adoption of the triangular membership function would be 92% with an average error chance of 8%.
T. L. Lam (2021) [32]	"In order to provide a low-cost and accurate temperature control solution for reflow systems, a cost-effective non-contact temperature approximation and control system is proposed in this article. The proposed temperature approximation is achieved based on a machine learning method with multiple-input single-output strategies to get a relationship between the temperatures near the PCBs and the onboard temperature."	Compared with the result of the PID controller, the FL controller's overshoot is improved from 3◦C to 1◦C, and its steady-state error is improved from ±1 ◦C to ±0.5◦C. Regression method is combined with RNN and FL to supervise the onboard temperatures in real-time, ensuring a good quality of productions.	The shortcoming of using the proposed method is that it may need to conduct data collection and regression for every new product to keep the high-temperature approximation accuracy.
C. K. Leung, J. D. Elias, S. M. Minuk, A. R. R. d. Jesus and A. Cuzzocrea (2020) [33]	"We design and develop an innovative fuzzy logic-based machine learning algorithm for supporting predictive analytics on big transportation data to help detect and predict the expected delay of streetcars (aka trolley cars) in the Canadian city of Toronto."	Using 5-fold cross validation (in which each fold used a 75%-25% split for training/test data), the ensemble of 40 decision trees in the RF regression led to a 70% reduction in the prediction error when compared with the mean rule algorithm.	The algorithm augments transit data with weather information, pre-processes them with fuzzy-logic based categorization, visualizes and analyzes the augmented data and their characteristics, mines frequent patterns and interesting association rules, and predicts delays with a RF.

S. Li (2019) [34]	"This work seeks to obtain an optimum way from regression of Supervised Learning to predict body temperature data from a smart pillow."	From the comparison of metrics (MSE, MAE, R-squared) for models, KNN and RF are better than other models.	For future research, more input features, more training data, and Deep Learning are deserved further attention. Only two features in this paper, it is not good enough for RF that needs more features to optimize results.
J. Liang, X. Liu and K. Liao (2018) [35]	"In this paper, we compare type-1 FL System and ANFIS to extract fuzzy parameters of soil. Moreover, two machine learning algorithms: RF and ANN with principal component analysis algorithms are applied in the SM classifications."	ANFIS with RF provides the best VWC correct recognition rate compared to other algorithms. the lower RMSEs demonstrate ANFIS's higher precisions than that of FL System in feature extractions. In recognitions, though ANFIS's features are corrupted by noise, it still shows a perfect correct recognition rate when combining with RF classification.	Fig. 5 shows the structure of the special five-layer network interpretation as a general expressions of ANFIS.
D. M. Minhas, R. R. Khalid and G. Frey (2017) [36]	"In this paper, an evaluation theory of hybrid model for short-term electricity load forecasting is presented using simple soft-technique of predicting data. A model that integrates fuzzy system with neural network database is demonstrated and eventually compared with a traditional statistical method of linear regression."	Power load forecasting errors especially for weekends, which is much higher than that of weekdays, is reduced using the probabilistic and stochastic natured Hybrid Adaptive Fuzzy Neural System (HAFNS) method. HAFNS showed much better forecasting result and errors compared to the simple linear regression.	In this paper, only the day-ahead profile is executed, whose data depends on a week-before load and temperature profile.
F. Mirzaei, M. Delavar, I. Alzoubi and B. Nadjar Arrabi (2018) [37]	"The purpose of this paper is to develop three methods including artificial bee colony algorithm (ABC-ANN), regression and adaptive neural fuzzy inference system (ANFIS) to predict the environmental indicators for land leveling and to analysis the sensitivity of these parameters."	Only three parameters of sand per cent, slope and soil, cut/fill volume had significant effects on energy consumption. All developed models had satisfactory performance in predicting aforementioned parameters in various field conditions. The ANFIS has the most capability in prediction according to least RMSE and the highest R^2 value.	ANFIS with a hybrid method of the gradient descent and the least-squares method was applied to find the optimal learning parameters using various membership functions (MFs). The implementations divulged that Gaussian membership function (gaussmf) and Trapezoidal membership function (tramf) configurations were found to denote MSE of 0.0166 and R2 of 0.98 for traction coefficient
A. G. Mohapatra and S. K. Lenka (2016) [14]	"In this paper, a partial least square regression (PLSR) and FL based smart decision support system (DSS) for crop-specific irrigation notification and control in precision agriculture is proposed, and this can be implemented in farm land, green-house and poly-house."	A comparative analysis on PLSR is done for soil moisture content prediction. It is observed from the errors that the prediction model produces exact soil moisture content as per the target value. Various SMS notifications are also generated using fuzzy-based model for sending the SMS notification to the farmer's handset.	This paper is trying out several membership function (MFs) for input and output variables.
S. K. Mousavi Mashhadi, H. Yadollahi and A. Marvian Mashhad (2016) [38]	"The system designed and manufactured in this paper measures, displays, and controls the total dissolved solids (TDS) of water by installation on domestic and industrial purification systems."	As evident from Figure 24, performance of the fuzzy controller is very acceptable in comparison with the classical PID controller, which was manually configured.	In this strategy, triangular membership functions with uniform distribution are used to make the input values fuzzy.

H. Neog, P. E. Dutta and N. Medhi (2022) [39]	"In this paper, we propose a remote Health Monitoring System for the prediction of health status of a person as well as detection of the risk of getting Covid for a particular patient using IoT and ML technologies."	The paper highlights the main steps, which can be elaborated as follows: Data preprocessing, feature engineering, model training (LSTM and KNN achieved the best metrics), SARIMA + LSTM-Markov to detect the risk of getting Covid, and finally FL optimization.	–
A. H. Orta, I. Kayabasi and M. Tunc (2018) [40]	"The purpose of this study is to develop a methodology to forecast short term wind power from the minimum number of input and to determine which turbines should be selected for Numerical Weather Prediction (NWP) models in the Bandirma power plant."	Using 6 or 4 turbine wind speeds, temperatures and wind directions with NARX model are the most powerful methods for Bandirma power plant. ANN models are more successful when the number of input is small. Furthermore, the most successful models for small number of inputs are regression models.	–
P. Patel, Y. Patel, U. Patel, V. Patel, N. Patel, P. Oza, et al. (2022) [15]	"We present a model for an automated watering system that attempts to reduce both human interaction and water usage."	The system takes soil moisture, weather parameters, and crop quality as input. Precipitation and its probability extracted from a weather API and crop quality was considered to optimize the system. DenseNet201 provided the most accurate classification, and all these parameters are then passed into our fuzzy system controller, deciding the exact time to irrigate the crop.	This automated system can assist in reducing water wastage and the need for manual monitoring. The concept may be improved by including a GSM module that allows farmers to receive notifications.
V. K. Patil and V. R. Pawar (2022) [41]	"We are proposing a system that can automatically identify human emotions with the help of sensors to be used at entrance gates."	The presented technique integrates the K-means clustering along with linear regression. Further, ANN and Fuzzy Classification paradigms was done to achieve emotion recognition through the Data collected by the IoT sensors and the dataset. The extensive experimentation for the recognition of the errors in the methodology through the use of MSE and RMSE reveals that the methodology achieves significant improvements in accuracy and reliability in emotion recognition.	FL and ANN methods are used with machine learning algorithms, which results in fewer errors in prediction. The measures of errors are given by MSE and RMSE.
B. Petković, D. Petković, B. Kuzman, M. Milovančević, K. Wakil, L. S. Ho, et al. (2020) [42]	"to establish regression models of the reference evapotranspiration in regard to several input weather parameters. The main aim is to achieve predictive capable models for the reference evapotranspiration."	Global radiation has the strongest influence on the reference evapotranspiration. Moreover, the combination of daily average temperature and global radiation is the optimal combination for the ET_0 estimation.	The main goal is minimization of the root mean square errors (RMSE). Additionally, coefficient of determination (R^2) and Pearson coefficient (r) were used for a more comprehensive analysis of the ANFIS prediction accuracy.

M. R. C. Qazani, V. Pourmostaghimi, M. Moayyedian and S. Pedrammehr (2022) [43]	"The main objective of this study is to calculate the tool–chip contact length using a highly advanced machine learning method without any time-consuming and expensive experiments. In this study, we proposed the ANFIS to predict the tool–chip contact length for the first time in orthogonal cutting using depth of cut, feed-rate, and cutting speed as inputs of the proposed model."	The GWO-ANFIS can decrease the mean square error between the actual and predicted tool–chip contact length of 15.60%, 3.67%, 89.75%, and 92.17% in comparison with those of GA-ANFIS, PSO-ANFIS, B-ANFIS, and GP, respectively. In addition, the fuzzy logic rule surface of the GWO-ANFIS shows 57.20%, 30.95%, and 11.85% dependency of tool–chip contact length to cutting speed, feed-rate, and depth of cut as the inputs of the orthogonal cutting process, respectively.	–
J. Refonaa and M. Lakshmi (2021) [44]	"An appropriate and error-reducing evaluation for rainfall prediction is performed in combination with the proposed BIRSM model and ANN, which can be very helpful for agriculture and food management."	The forecasting effect of FL and BIRSM is ideal. It can not only understand the extent of rainfall but also can accurately handle the condition of no rainfall.	The Mean Square Error (MSE) was chosen for performance analysis function.
F. Sabrina, S. Sohail, F. Farid, S. Jahan, F. Ahamed and S. Gordon (2022) [10]	"In this paper, we propose a novel artificial intelligence-based agriculture system that uses IoT data to monitor the environment and alerts farmers to take the required actions for maintaining ideal conditions for crop production."	The experimental results show that the proposed system is interpretable, can detect anomalous data, and triggers actions accurately based on crop requirements.	The strength of the proposed system is in its interpretability which makes it easy for farmers to understand, trust and use it. The use of fuzzy logic makes the system customisable in terms of types/number of sensors, type of crop, and adaptable for any soil types and weather conditions.
S. Sharma, R. K. Agrawal and M. M. Tripathi (2020) [45]	"A novel method of short-term load forecasting based on the combination of FL with RNN model has been proposed in this paper. The proposed approach combines the advantages of fuzzy logic and neural networks to predict the next day's load."	The computed results conclude that the synergetic use of FL with RNN model is successful in achieving higher accuracy by efficiently mapping the effect of weather parameters with a change in load demand. Fuzzy-RNN has performed best with the highest accuracy in load forecasting amongst the six models considered.	–
A. Shastry, H. A. Sanjay and M. Hegde (2015) [46]	"This work analyses how yield of a particular crop is determined by few attributes. In this paper, several ML models are used for predicting the yield of wheat by considering biomass, extractable soil water (ESW), radiation and rain as input parameters."	ANFIS model performs better than Multiple LinReg and FL models with a lower RMSE value.	All the models are compared based on the RMSE values.

J. M. Siqueira, T. A. Paço, J. C. Silvestre, F. L. Santos, A. O. Falcão and L. S. Pereira (2014) [47]	"The present study aims at developing an intelligent system of automating data analysis and prediction embedded in a FL algorithm to capture the relationship between environmental variables and sap flow measurements (Granier method)."	The FL algorithm shows to be an effective approach for system optimization, allowing a less time consuming process, yet not discarding the human decision capacity, since it mimetizes the process. In addition, it provides the opportunity for earlier reaction to data because it allows for a real-time treatment possibility.	–
Q. T. T. Tran, K. Davies and L. Roose (2021) [48]	"This paper proposed a low-cost method to build the machine learning training dataset for assessing service transformer health by using FL method."	The FL workflow validated by SVM demonstrates that the built-in training data set is efficient, applicable to assess the transformer condition. Additionally, the data generation proposed in this paper has high feature continuity and good scalability that can be used as a training data for machine learning, deep learning models.	Figure 2 shows the range of each parameter corresponding to various conditions. These map through expert rules to overall transformer health conditions and recommended actions.
V. Vivekanandhan, S. Sakthivel and M. Manikandan (2022) [49]	"This proposed work introduces an ANFIS technique for analyzing agricultural plant growth based on soil, water level, temperature, and moisture conditions."	ANFIS results in an overall better result compared to the other algorithms; higher precision and recall, higher accuracy, and lower error rate.	–
H. Yuan, M. Tan and Y. Chen (2014) [9]	"This paper proposes a non-linear regression model (NRM) for fishing forecast. It employs cluster analysis and nonlinear regression to help forecast fishing yield based on marine environmental data."	The comparison result reveals this new NRM model increases both the accuracy in fishery forecast and the reliability in guiding fishery production and related activities, which is proved by higher coefficiency of determination compared to other methods. It can also help explore and discover the distribution of fishing grounds.	The model combines fishery domain expert knowledge, marine environmental factor data such as water temperature, chlorophyll concentration and sea surface level as base data and applies cluster analysis that incorporates function fitting and nonlinear regression for data analysis and processing.
G. Zhang, S. S. Band, S. Ardabili, K. W. Chau and A. Mosavi (2022) [50]	"In this study, data-driven simulation is used to model dew point temperature (DPT). The forecasting method based on ANFIS is used to estimate this factor at Tabriz, one of the earliest Iranian meteorological stations."	The results reveal that the ANFIS method is capable of identifying data patterns with a high degree of accuracy in general. ANFIS model is very stable for almost all numbers of membership functions. Moreover, it is efficient regarding the computations, it can also be applied when different and complex parameters are engaged in the processes.	Sugeno FIS was used. For a short number of iterations, both training and testing processes had high RSME, but as the number of iterations increases, RSME decreased. The number of numerical iterations in the testing procedure for both MMSE and RSME rose between 50 and 100. These parameters, however, diminished after around 100 iterations.

C

Appendix C: Open Source Libraries Used in the Project

Table C.1: Open Source Libraries

Library	Description	Link
anfis	The Adaptive Neuro-Fuzzy Inference System (ANFIS) library for combining fuzzy logic and neural networks in machine learning.	https://pypi.org/project/anfis/
matplotlib	A comprehensive data visualization library for creating static, animated, and interactive plots and charts in Python.	https://pypi.org/project/matplotlib/
numpy	A fundamental library for numerical computing in Python, supporting large, multi-dimensional arrays and mathematical functions.	https://pypi.org/project/numpy/
pandas	A popular data manipulation library with data structures for efficiently handling structured data.	https://pypi.org/project/pandas/
plotly	A library for creating interactive and visually appealing data visualizations, including various chart types.	https://pypi.org/project/plotly/
pyspark	The Python API for Apache Spark, enabling distributed data processing and big data analysis.	https://pypi.org/project/pyspark/
sanfis	An extension of ANFIS incorporating self-adaptation mechanisms for improved model performance.	https://pypi.org/project/sanfis/
scikeras	A library for integrating Keras with scikit-learn, simplifying deep learning model building within scikit-learn.	https://pypi.org/project/scikeras/
scipy	An open-source library for mathematics, science, and engineering, extending NumPy with additional modules.	https://pypi.org/project/scipy/
seaborn	A data visualization library based on Matplotlib, designed for creating informative and attractive statistical graphics.	https://pypi.org/project/seaborn/
skfuzzy	A fuzzy logic toolkit for scikit-learn, used for modeling uncertainty and decision-making.	https://pypi.org/project/scikit-fuzzy/
sklearn	Scikit-learn, a machine learning library offering a wide range of tools and algorithms for various tasks.	https://pypi.org/project/scikit-learn/
statsmodels	A Python library for estimating and interpreting statistical models, covering linear and non-linear models, time series analysis, and more.	https://pypi.org/project/statsmodels/
statistics	A library providing mathematical and statistical functions for basic statistical calculations.	https://pypi.org/project/statistics/
sweetviz	An automated exploratory data analysis (EDA) library, generating visualizations and summaries of dataset characteristics.	https://pypi.org/project/sweetviz/
tensorflow	A deep learning framework developed by Google, providing tools and libraries for neural network building and training.	https://pypi.org/project/tensorflow/
xgboost	A gradient boosting library widely used for supervised machine learning tasks, known for its high performance and accuracy.	https://pypi.org/project/xgboost/

D

Appendix D: Feature Distribution of the Products in Freeze Dryer Machines

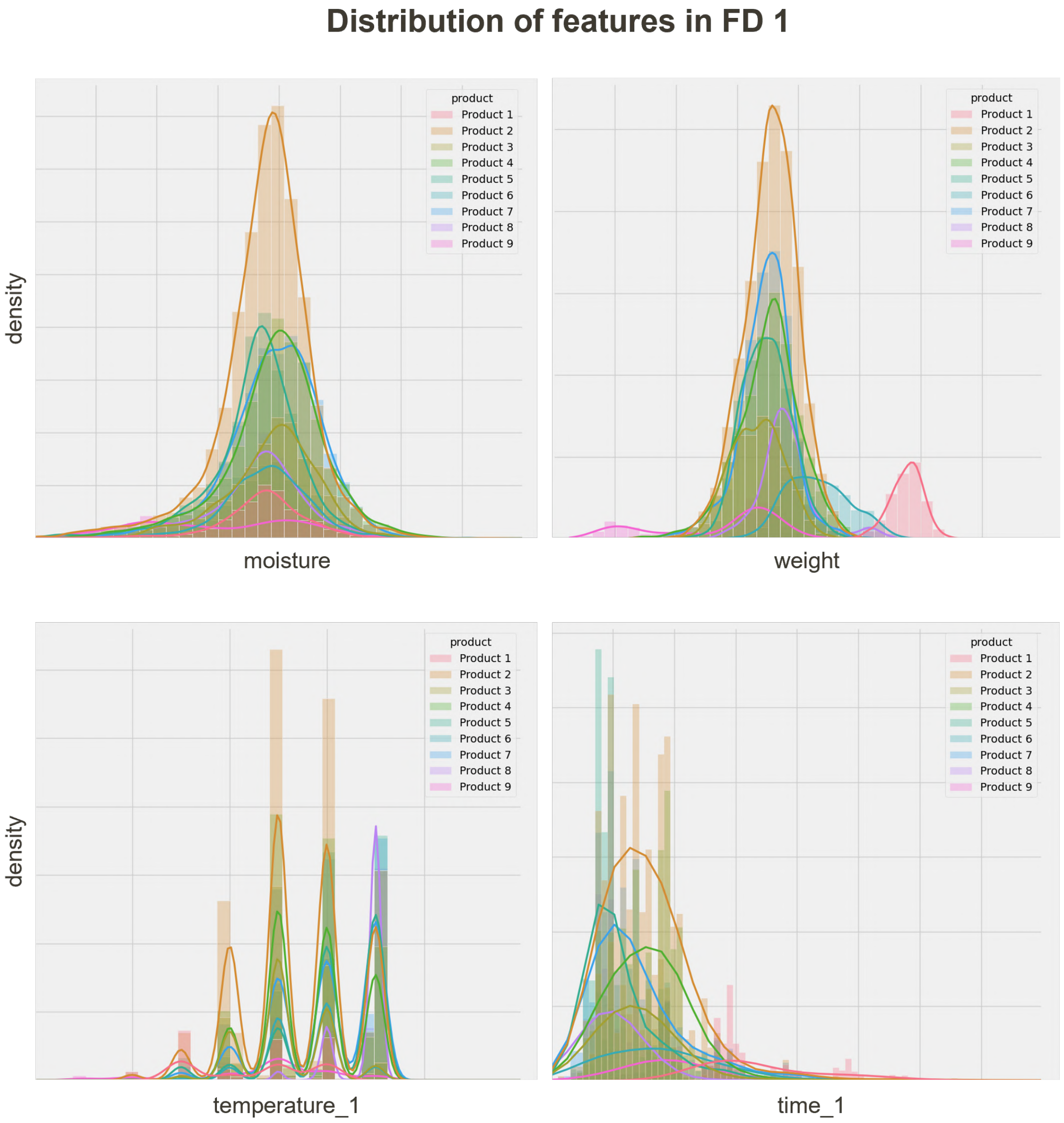


Figure D.1: Distribution of Features in FD 1

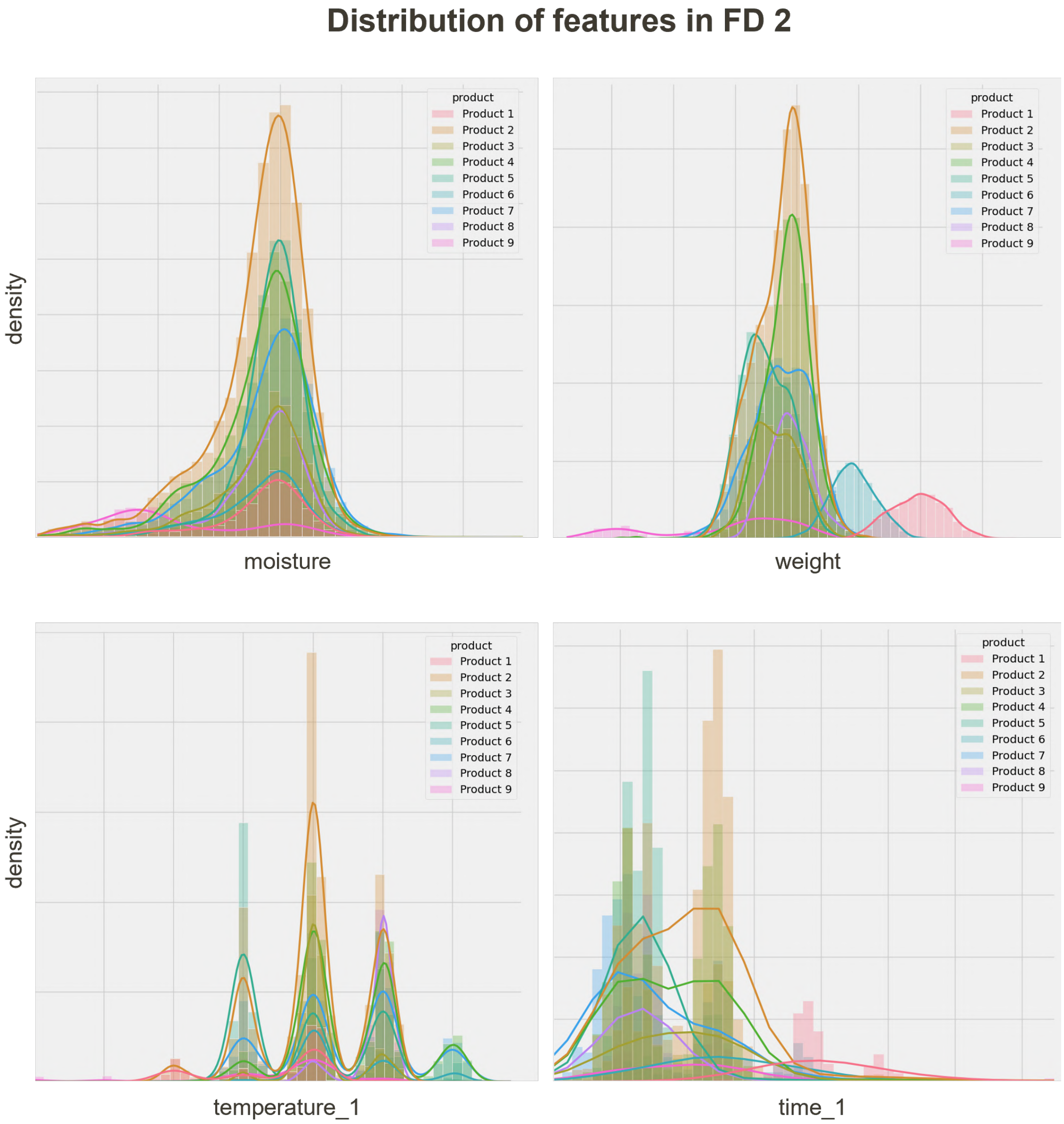


Figure D.2: Distribution of Features in FD 2

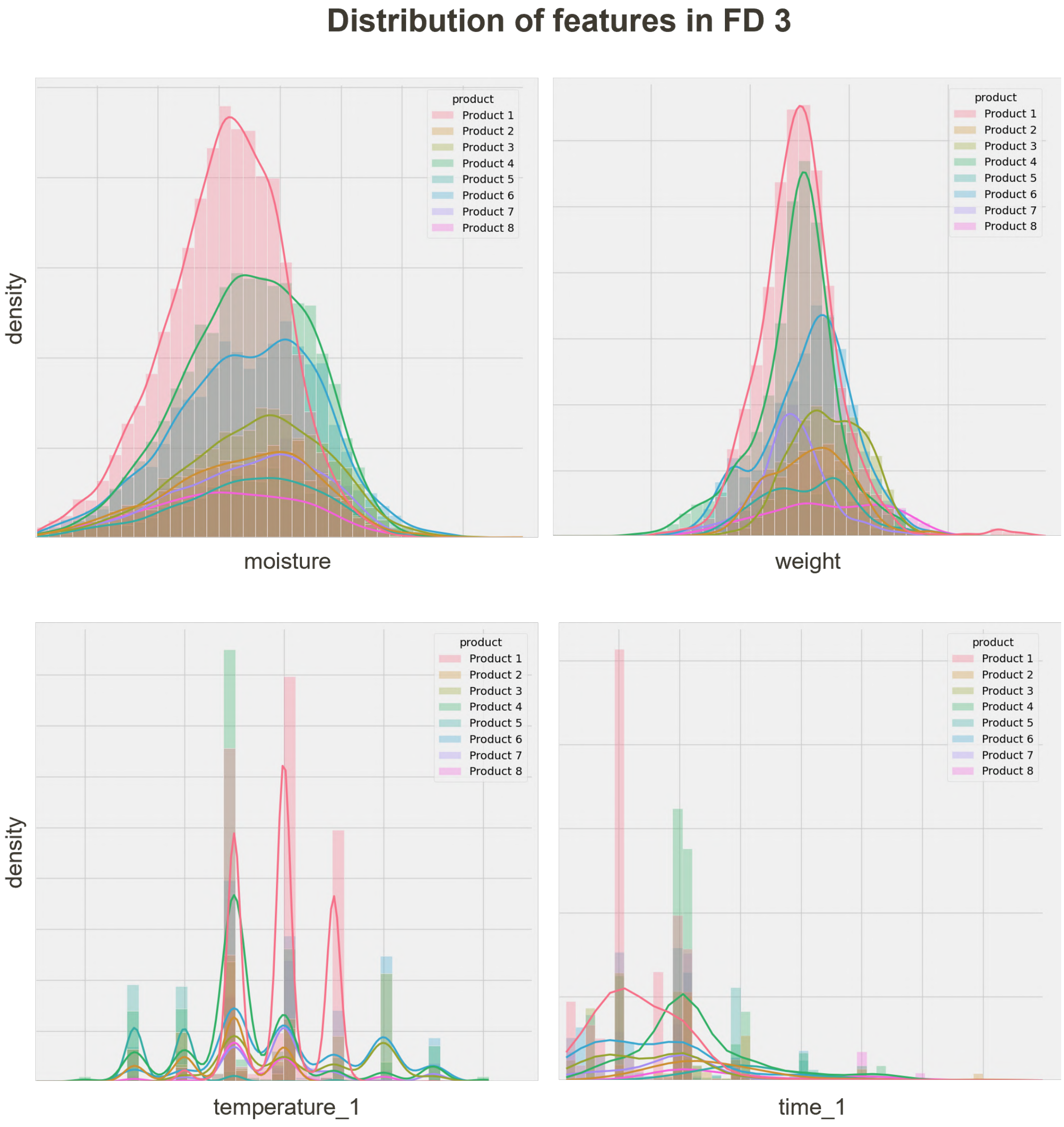


Figure D.3: Distribution of Features in FD 3

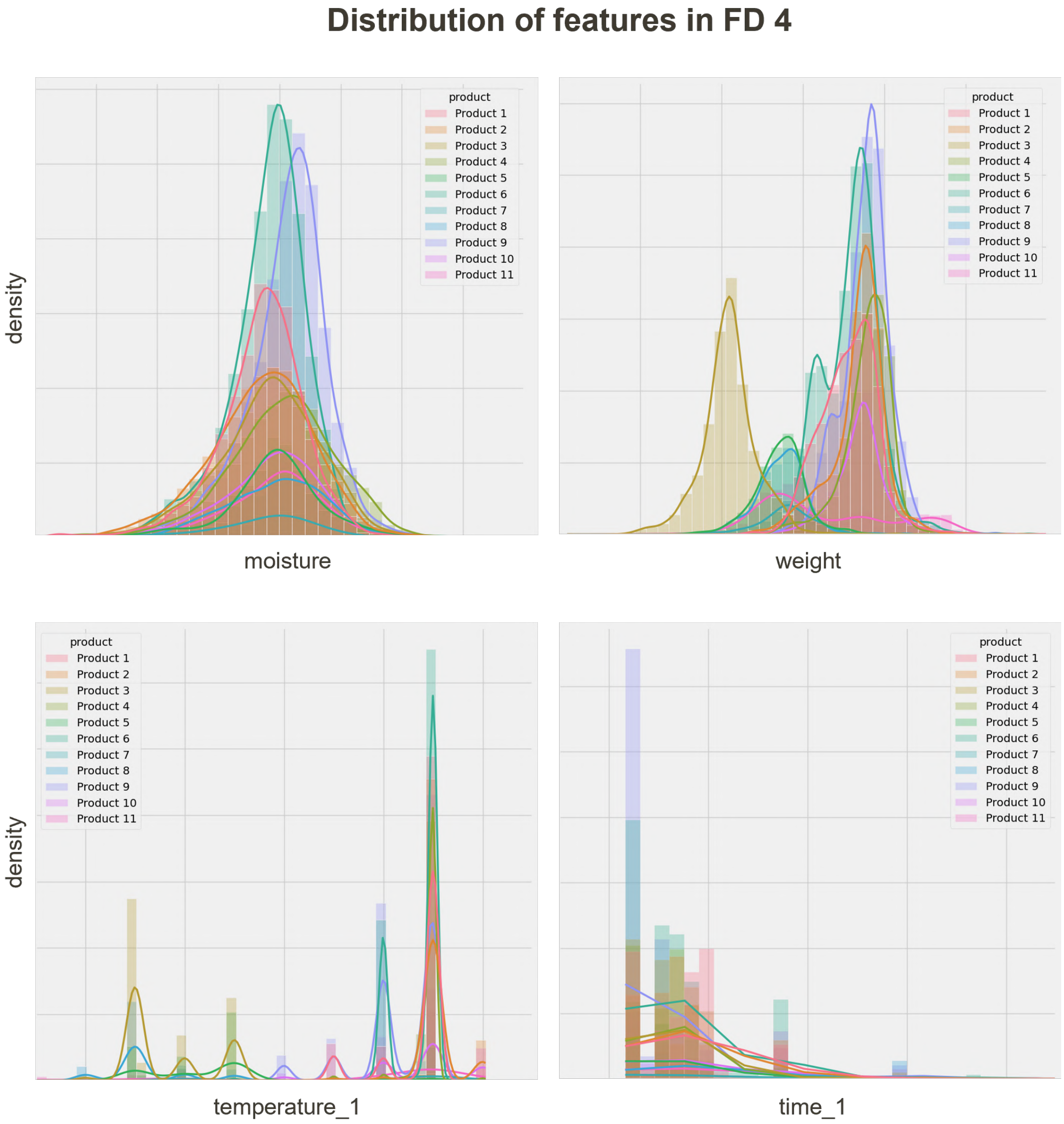


Figure D.4: Distribution of Features in FD 4

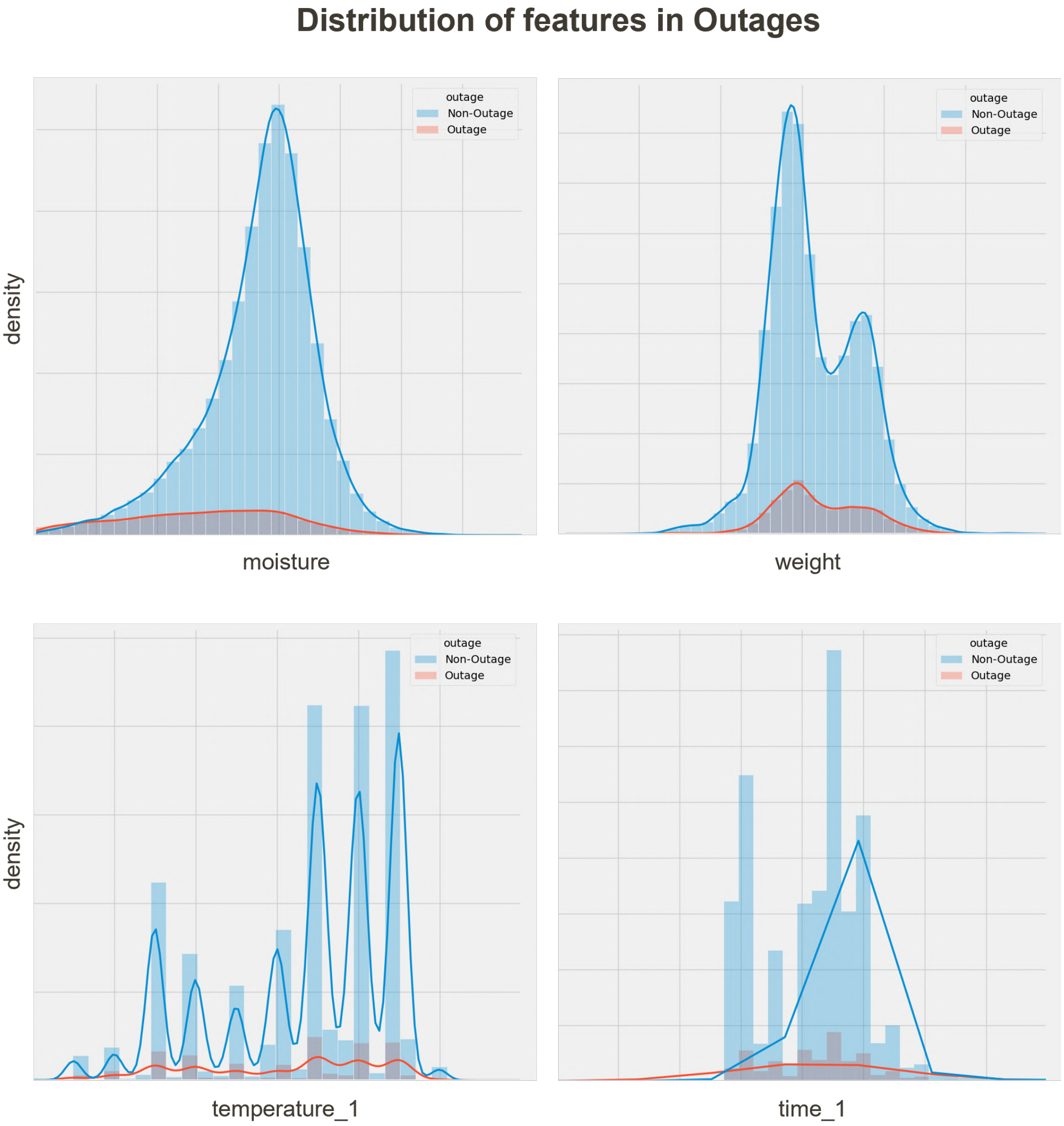


Figure D.5: Feature Distribution in Outages

E

Appendix E: Shapiro-Wilk Test Results

Table E.1: Shapiro-Wilk Test Results for Features in Different FD Machines

Feature	FD	Product	Outage	p-value	Description
moisture	1	All	All	Close to 0	Does not follow a normal distribution
moisture	2	All	All	Close to 0	Does not follow a normal distribution
moisture	3	All	All	7.35e-35	Does not follow a normal distribution
moisture	4	All	All	Close to 0	Does not follow a normal distribution
weight	1	All	All	Close to 0	Does not follow a normal distribution
weight	2	All	All	Close to 0	Does not follow a normal distribution
weight	3	All	All	8.97e-39	Does not follow a normal distribution
weight	4	All	All	Close to 0	Does not follow a normal distribution
temperature_1	1	All	All	Close to 0	Does not follow a normal distribution
temperature_1	2	All	All	Close to 0	Does not follow a normal distribution
temperature_1	3	All	All	Close to 0	Does not follow a normal distribution
temperature_1	4	All	All	Close to 0	Does not follow a normal distribution
time_1	1	All	All	Close to 0	Does not follow a normal distribution
time_1	2	All	All	Close to 0	Does not follow a normal distribution
time_1	3	All	All	Close to 0	Does not follow a normal distribution
time_1	4	All	All	Close to 0	Does not follow a normal distribution
total_heat	1	All	All	Close to 0	Does not follow a normal distribution
total_heat	2	All	All	Close to 0	Does not follow a normal distribution
total_heat	3	All	All	Close to 0	Does not follow a normal distribution
total_heat	4	All	All	Close to 0	Does not follow a normal distribution

Table E.2: Shapiro-Wilk Test Results for Features in FD 1

Feature	FD	Product	Outage	p-value	Description
moisture	1	1	All	1.04e-22	Does not follow a normal distribution
moisture	1	2	All	Close to 0	Does not follow a normal distribution
moisture	1	3	All	1.16e-25	Does not follow a normal distribution
moisture	1	4	All	6.72e-35	Does not follow a normal distribution
moisture	1	5	All	1.37e-41	Does not follow a normal distribution
moisture	1	6	All	1.46e-18	Does not follow a normal distribution
moisture	1	7	All	3.23e-31	Does not follow a normal distribution
moisture	1	8	All	2.16e-28	Does not follow a normal distribution
moisture	1	9	All	6.07e-14	Does not follow a normal distribution
weight	1	1	All	7.66e-28	Does not follow a normal distribution
weight	1	2	All	2.92e-24	Does not follow a normal distribution
weight	1	3	All	9.54e-08	Does not follow a normal distribution
weight	1	4	All	2.21e-23	Does not follow a normal distribution
weight	1	5	All	5.08e-12	Does not follow a normal distribution
weight	1	6	All	3.26e-15	Does not follow a normal distribution
weight	1	7	All	6.69e-33	Does not follow a normal distribution
weight	1	8	All	2.62e-33	Does not follow a normal distribution
weight	1	9	All	1.61e-29	Does not follow a normal distribution
temperature_1	1	1	All	1.15e-32	Does not follow a normal distribution
temperature_1	1	2	All	Close to 0	Does not follow a normal distribution
temperature_1	1	3	All	1.70e-39	Does not follow a normal distribution
temperature_1	1	4	All	Close to 0	Does not follow a normal distribution
temperature_1	1	5	All	Close to 0	Does not follow a normal distribution
temperature_1	1	6	All	2.78e-35	Does not follow a normal distribution
temperature_1	1	7	All	Close to 0	Does not follow a normal distribution
temperature_1	1	8	All	Close to 0	Does not follow a normal distribution
temperature_1	1	9	All	6.89e-29	Does not follow a normal distribution
time_1	1	1	All	Close to 0	Does not follow a normal distribution
time_1	1	2	All	Close to 0	Does not follow a normal distribution
time_1	1	3	All	Close to 0	Does not follow a normal distribution
time_1	1	4	All	Close to 0	Does not follow a normal distribution
time_1	1	5	All	Close to 0	Does not follow a normal distribution
time_1	1	6	All	Close to 0	Does not follow a normal distribution

time_1	1	7	All	Close to 0	Does not follow a normal distribution
time_1	1	8	All	Close to 0	Does not follow a normal distribution
time_1	1	9	All	Close to 0	Does not follow a normal distribution
total_heat	1	1	All	4.07e-08	Does not follow a normal distribution
total_heat	1	2	All	6.68e-43	Does not follow a normal distribution
total_heat	1	3	All	5.03e-19	Does not follow a normal distribution
total_heat	1	4	All	4.17e-27	Does not follow a normal distribution
total_heat	1	5	All	6.25e-13	Does not follow a normal distribution
total_heat	1	6	All	4.94e-19	Does not follow a normal distribution
total_heat	1	7	All	4.11e-15	Does not follow a normal distribution
total_heat	1	8	All	2.12e-11	Does not follow a normal distribution
total_heat	1	9	All	1.95e-17	Does not follow a normal distribution

Table E.3: Shapiro-Wilk Test Results for Features in FD 2

Feature	FD	Product	Outage	p-value	Description
moisture	2	1	All	5.85e-26	Does not follow a normal distribution
moisture	2	2	All	Close to 0	Does not follow a normal distribution
moisture	2	3	All	3.86e-37	Does not follow a normal distribution
moisture	2	4	All	Close to 0	Does not follow a normal distribution
moisture	2	5	All	Close to 0	Does not follow a normal distribution
moisture	2	6	All	4.58e-28	Does not follow a normal distribution
moisture	2	7	All	6.14e-41	Does not follow a normal distribution
moisture	2	8	All	7.13e-40	Does not follow a normal distribution
moisture	2	9	All	1.43e-21	Does not follow a normal distribution
weight	2	1	All	Close to 0	Does not follow a normal distribution
weight	2	2	All	5.76e-31	Does not follow a normal distribution
weight	2	3	All	4.42e-11	Does not follow a normal distribution
weight	2	4	All	2.80e-45	Does not follow a normal distribution
weight	2	5	All	2.53e-13	Does not follow a normal distribution
weight	2	6	All	0.08e-10	Does not follow a normal distribution
weight	2	7	All	9.94e-17	Does not follow a normal distribution
weight	2	8	All	6.52e-13	Does not follow a normal distribution
weight	2	9	All	5.39e-26	Does not follow a normal distribution
temperature_1	2	1	All	3.01e-36	Does not follow a normal distribution
temperature_1	2	2	All	Close to 0	Does not follow a normal distribution
temperature_1	2	3	All	Close to 0	Does not follow a normal distribution
temperature_1	2	4	All	1.54e-44	Does not follow a normal distribution
temperature_1	2	5	All	Close to 0	Does not follow a normal distribution
temperature_1	2	6	All	5.26e-37	Does not follow a normal distribution
temperature_1	2	7	All	1.54e-39	Does not follow a normal distribution
temperature_1	2	8	All	Close to 0	Does not follow a normal distribution
temperature_1	2	9	All	2.50e-36	Does not follow a normal distribution
time_1	2	1	All	Close to 0	Does not follow a normal distribution
time_1	2	2	All	Close to 0	Does not follow a normal distribution
time_1	2	3	All	Close to 0	Does not follow a normal distribution
time_1	2	4	All	Close to 0	Does not follow a normal distribution
time_1	2	5	All	Close to 0	Does not follow a normal distribution
time_1	2	6	All	Close to 0	Does not follow a normal distribution

time_1	2	7	All	Close to 0	Does not follow a normal distribution
time_1	2	8	All	Close to 0	Does not follow a normal distribution
time_1	2	9	All	Close to 0	Does not follow a normal distribution
total_heat	2	1	All	1.13e-05	Does not follow a normal distribution
total_heat	2	2	All	7.47e-10	Does not follow a normal distribution
total_heat	2	3	All	1.15e-06	Does not follow a normal distribution
total_heat	2	4	All	6.97e-32	Does not follow a normal distribution
total_heat	2	5	All	8.87e-20	Does not follow a normal distribution
total_heat	2	6	All	9.83e-14	Does not follow a normal distribution
total_heat	2	7	All	6.98e-29	Does not follow a normal distribution
total_heat	2	8	All	4.49e-23	Does not follow a normal distribution
total_heat	2	9	All	1.56e-19	Does not follow a normal distribution

Table E.4: Shapiro-Wilk Test Results for Features in FD 3

Feature	FD	Product	Outage	p-value	Description
moisture	3	1	All	7.88e-20	Does not follow a normal distribution
moisture	3	2	All	7.74e-10	Does not follow a normal distribution
moisture	3	3	All	2.95e-09	Does not follow a normal distribution
moisture	3	4	All	6.34e-17	Does not follow a normal distribution
moisture	3	5	All	7.82e-09	Does not follow a normal distribution
moisture	3	6	All	1.10e-13	Does not follow a normal distribution
moisture	3	7	All	1.12e-13	Does not follow a normal distribution
moisture	3	8	All	1.45e-06	Does not follow a normal distribution
weight	3	1	All	Close to 0	Does not follow a normal distribution
weight	3	2	All	4.18e-12	Does not follow a normal distribution
weight	3	3	All	1.14e-08	Does not follow a normal distribution
weight	3	4	All	1.30e-34	Does not follow a normal distribution
weight	3	5	All	2.21e-13	Does not follow a normal distribution
weight	3	6	All	4.75e-19	Does not follow a normal distribution
weight	3	7	All	2.36e-15	Does not follow a normal distribution
weight	3	8	All	3.31e-09	Does not follow a normal distribution
temperature_1	3	1	All	Close to 0	Does not follow a normal distribution
temperature_1	3	2	All	3.35e-36	Does not follow a normal distribution
temperature_1	3	3	All	2.45e-40	Does not follow a normal distribution
temperature_1	3	4	All	Close to 0	Does not follow a normal distribution
temperature_1	3	5	All	1.09e-42	Does not follow a normal distribution
temperature_1	3	6	All	4.78e-41	Does not follow a normal distribution
temperature_1	3	7	All	3.57e-41	Does not follow a normal distribution
temperature_1	3	8	All	1.01e-34	Does not follow a normal distribution
time_1	3	1	All	Close to 0	Does not follow a normal distribution
time_1	3	2	All	Close to 0	Does not follow a normal distribution
time_1	3	3	All	Close to 0	Does not follow a normal distribution
time_1	3	4	All	Close to 0	Does not follow a normal distribution
time_1	3	5	All	Close to 0	Does not follow a normal distribution
time_1	3	6	All	Close to 0	Does not follow a normal distribution
time_1	3	7	All	Close to 0	Does not follow a normal distribution
time_1	3	8	All	4.86e-43	Does not follow a normal distribution
total_heat	3	1	All	2.11e-07	Does not follow a normal distribution

total_heat	3	2	All	4.69e-06	Does not follow a normal distribution
total_heat	3	3	All	1.58e-23	Does not follow a normal distribution
total_heat	3	4	All	3.56e-23	Does not follow a normal distribution
total_heat	3	5	All	1.62e-26	Does not follow a normal distribution
total_heat	3	6	All	3.23e-10	Does not follow a normal distribution
total_heat	3	7	All	6.68e-20	Does not follow a normal distribution
total_heat	3	8	All	7.43e-16	Does not follow a normal distribution

Table E.5: Shapiro-Wilk Test Results for Features in FD 4

Feature	FD	Product	Outage	p-value	Description
moisture	4	1	All	3.17e-23	Does not follow a normal distribution
moisture	4	2	All	9.78e-20	Does not follow a normal distribution
moisture	4	3	All	2.75e-21	Does not follow a normal distribution
moisture	4	4	All	1.14e-24	Does not follow a normal distribution
moisture	4	5	All	9.48e-26	Does not follow a normal distribution
moisture	4	6	All	3.19e-41	Does not follow a normal distribution
moisture	4	7	All	1.56e-15	Does not follow a normal distribution
moisture	4	8	All	2.58e-09	Does not follow a normal distribution
moisture	4	9	All	1.04e-43	Does not follow a normal distribution
moisture	4	10	All	7.60e-19	Does not follow a normal distribution
moisture	4	11	All	5.64e-21	Does not follow a normal distribution
weight	4	1	All	3.16e-31	Does not follow a normal distribution
weight	4	2	All	6.02e-44	Does not follow a normal distribution
weight	4	3	All	1.69e-37	Does not follow a normal distribution
weight	4	4	All	2.80e-45	Does not follow a normal distribution
weight	4	5	All	2.22e-32	Does not follow a normal distribution
weight	4	6	All	2.38e-44	Does not follow a normal distribution
weight	4	7	All	7.39e-14	Does not follow a normal distribution
weight	4	8	All	5.75e-27	Does not follow a normal distribution
weight	4	9	All	4.90e-44	Does not follow a normal distribution
weight	4	10	All	2.47e-37	Does not follow a normal distribution
weight	4	11	All	1.72e-23	Does not follow a normal distribution
temperature_1	4	1	All	Close to 0	Does not follow a normal distribution
temperature_1	4	2	All	Close to 0	Does not follow a normal distribution
temperature_1	4	3	All	Close to 0	Does not follow a normal distribution
temperature_1	4	4	All	Close to 0	Does not follow a normal distribution
temperature_1	4	5	All	1.16e-42	Does not follow a normal distribution
temperature_1	4	6	All	Close to 0	Does not follow a normal distribution
temperature_1	4	7	All	1.50e-33	Does not follow a normal distribution
temperature_1	4	8	All	3.78e-44	Does not follow a normal distribution
temperature_1	4	9	All	Close to 0	Does not follow a normal distribution
temperature_1	4	10	All	1.34e-41	Does not follow a normal distribution
temperature_1	4	11	All	2.80e-45	Does not follow a normal distribution

time_1	4	1	All	Close to 0	Does not follow a normal distribution
time_1	4	2	All	Close to 0	Does not follow a normal distribution
time_1	4	3	All	Close to 0	Does not follow a normal distribution
time_1	4	4	All	Close to 0	Does not follow a normal distribution
time_1	4	5	All	Close to 0	Does not follow a normal distribution
time_1	4	6	All	Close to 0	Does not follow a normal distribution
time_1	4	7	All	Close to 0	Does not follow a normal distribution
time_1	4	8	All	Close to 0	Does not follow a normal distribution
time_1	4	9	All	Close to 0	Does not follow a normal distribution
time_1	4	10	All	Close to 0	Does not follow a normal distribution
time_1	4	11	All	Close to 0	Does not follow a normal distribution
total_heat	4	1	All	1.57e-18	Does not follow a normal distribution
total_heat	4	2	All	1.29e-31	Does not follow a normal distribution
total_heat	4	3	All	8.18e-10	Does not follow a normal distribution
total_heat	4	4	All	3.49e-29	Does not follow a normal distribution
total_heat	4	5	All	7.22e-12	Does not follow a normal distribution
total_heat	4	6	All	8.77e-23	Does not follow a normal distribution
total_heat	4	7	All	1.32e-19	Does not follow a normal distribution
total_heat	4	8	All	6.66e-19	Does not follow a normal distribution
total_heat	4	9	All	4.98e-16	Does not follow a normal distribution
total_heat	4	10	All	7.66e-09	Does not follow a normal distribution
total_heat	4	11	All	4.09e-28	Does not follow a normal distribution

F

Appendix F: Number of Rows of the Freeze Dryer Datasets

Table F.1: Number of Rows of the Categorized FD Dataset

Freeze Dryer	**Product ID**	**Number of Rows**
FD 1	Product 1	1,164
	Product 2	10,275
	Product 3	3,203
	Product 4	5,946
	Product 5	4,949
	Product 6	2,048
	Product 7	6,053
	Product 8	2,431
	Product 9	1,080
	Outage	4,760
FD 2	Product 1	1,344
	Product 2	10,235
	Product 3	3,133
	Product 4	6,729
	Product 5	5,410
	Product 6	1,716
	Product 7	5,750
	Product 8	2,754
	Product 9	1,128
	Outage	4,217
FD 3	Product 1	7,654
	Product 2	2,080
	Product 3	2,827
	Product 4	5,941
	Product 5	1,432
	Product 6	4,811
	Product 7	1,972
	Product 8	1,243
	Outage	6,085
FD 4	Product 1	4,346
	Product 2	4,013
	Product 3	3,683
	Product 4	3,518
	Product 5	1,626
	Product 6	6,521
	Product 7	534
	Product 8	1,432
	Product 9	6,112
	Product 10	1,931
	Product 11	1,462
	Outage	2,875

G

Appendix G: Initial Set of Hyperparameters for the Machine Learning Models

Table G.1: Initial Set of Hyperparameters for the Machine Learning Models

ML Model	Hyperparameter	Description	Initial Value
ElasticNet Regression	alpha	Regularization strength	0-1 in steps of 0.01
	l1_ratio	L1/L2 regularization ratio	0-1 in steps of 0.1
Support Vector Regression	kernel	Type of kernel function	*rbf, poly, sigmoid*
	C	Regularization parameter	0.01, 0.1, 1, 10, 100
	epsilon	Margin of error tolerance value	0.01, 0.1, 1, 10, 100
Random Forest Regression	n_estimators	Number of trees in the forest	100-1000 in steps of 100
	max_depth	Maximum depth of each tree	3-10 in steps of 1
	min_samples_split	Minimum number of samples required to split an internal node	3-10 in steps of 1
	max_features	Maximum number of features to consider for splitting	*sqrt, log2*
XGBoost Regression	n_estimators	Number of boosting rounds	100-1000 in steps of 100
	max_depth	Maximum depth of each tree	3-10 in steps of 1
	learning_rate	Step size shrinkage to prevent overfitting	0.001, 0.01, 0.1, 0.5, 0.9, 1.0
	subsample	Fraction of samples used for fitting the trees	0.5-1 in steps of 0.05
	colsample_bytree	Fraction of features used for fitting the trees	0.5-1 in steps of 0.05

H

Appendix H: Scatter Plots of the Machine Learning Models

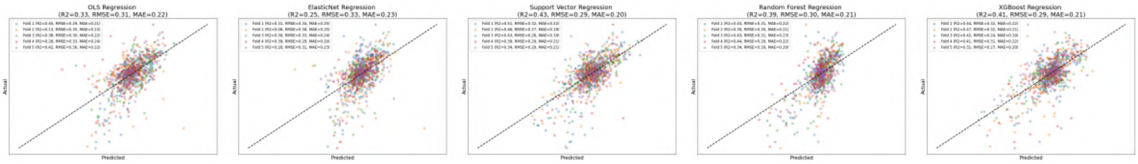


Figure H.1: FD 1, Product 1

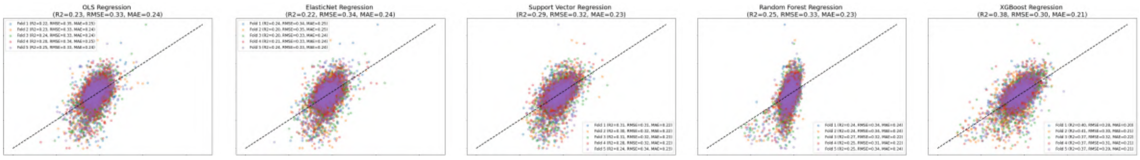


Figure H.2: FD 1, Product 2

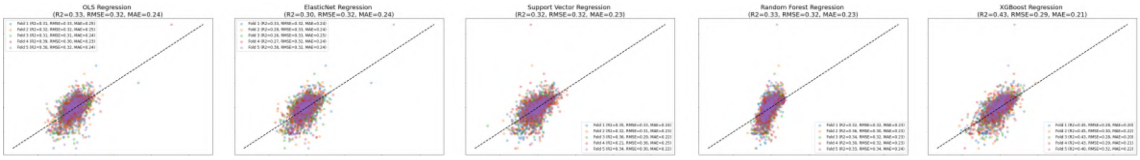


Figure H.3: FD 1, Product 3

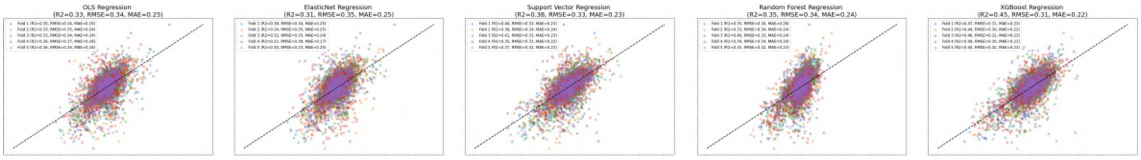


Figure H.4: FD 1, Product 4

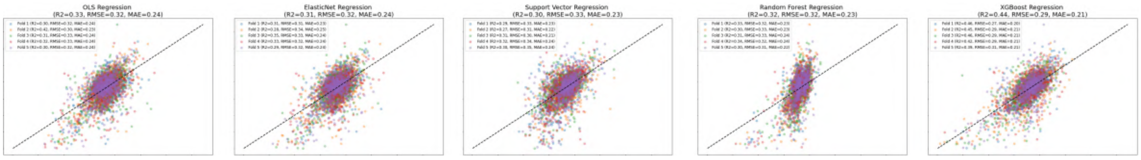


Figure H.5: FD 1, Product 5

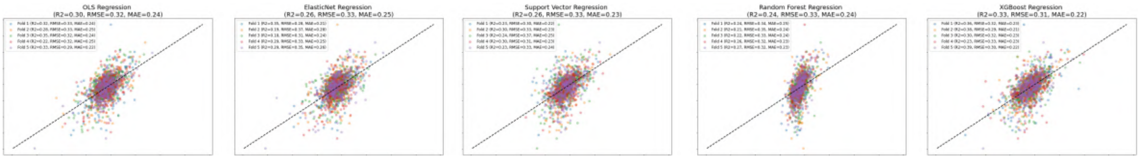


Figure H.6: FD 1, Product 6

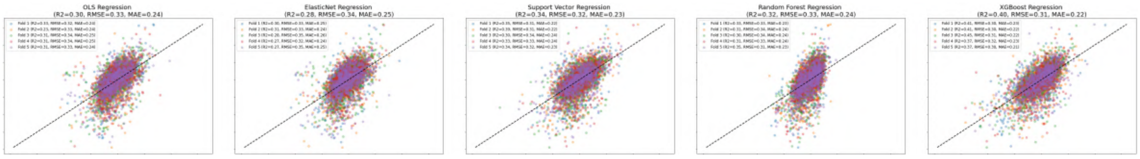


Figure H.7: FD 1, Product 7

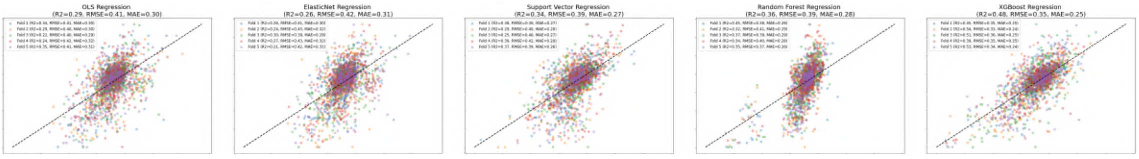


Figure H.8: FD 1, Product 8

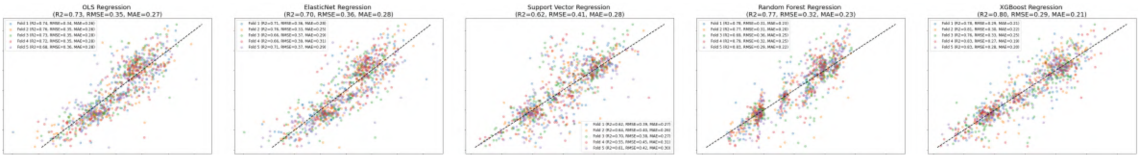


Figure H.9: FD 1, Product 9

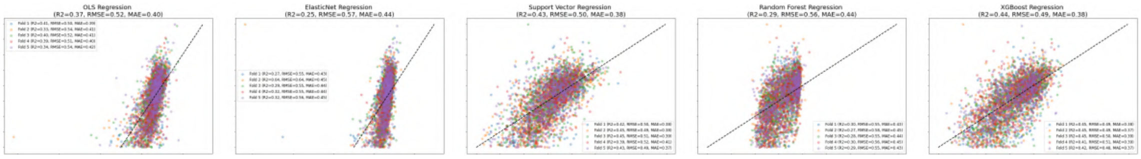


Figure H.10: FD 1, Outage

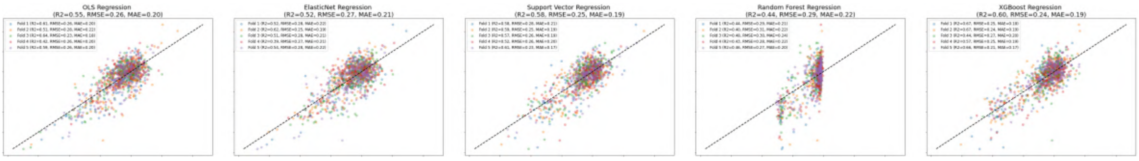


Figure H.11: FD 2, Product 1

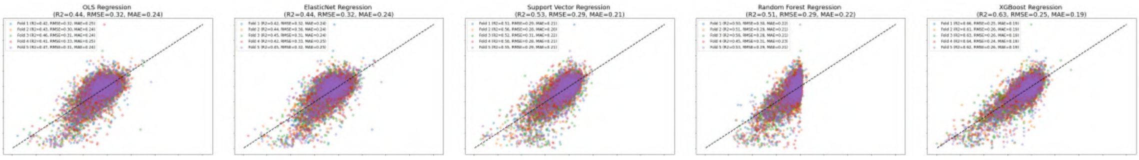


Figure H.12: FD 2, Product 2

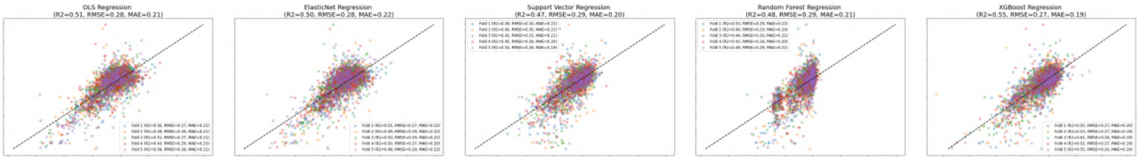


Figure H.13: FD 2, Product 3

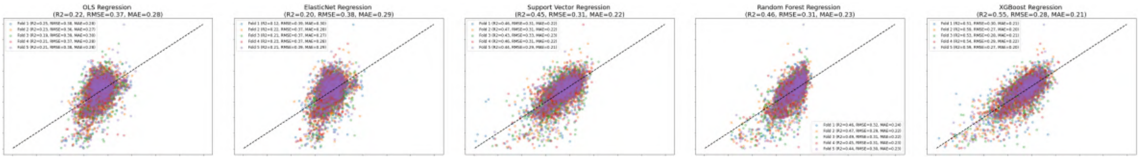


Figure H.14: FD 2, Product 4

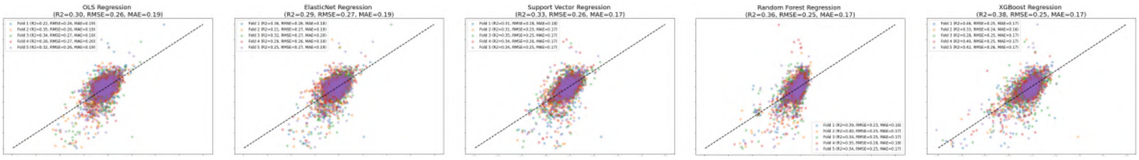


Figure H.15: FD 2, Product 5

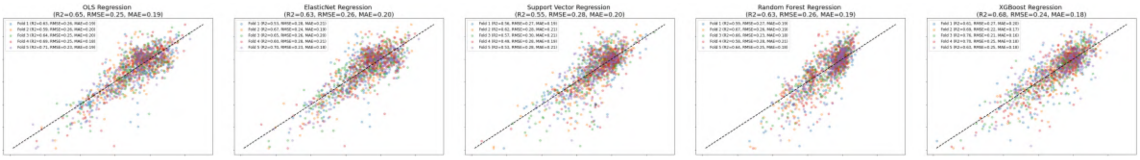


Figure H.16: FD 2, Product 6

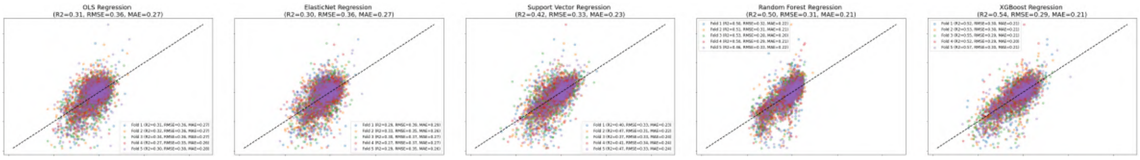


Figure H.17: FD 2, Product 7

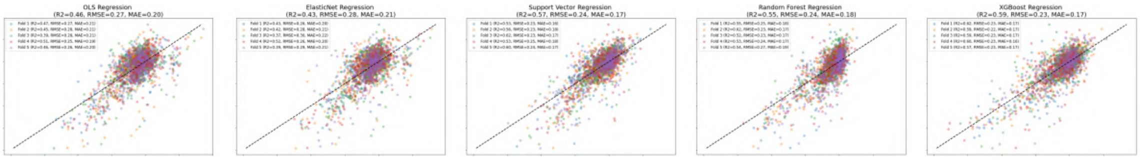


Figure H.18: FD 2, Product 8

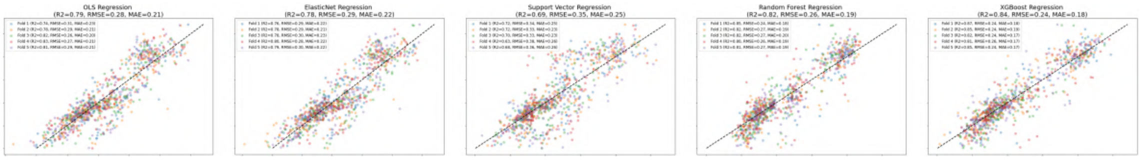


Figure H.19: FD 2, Product 9

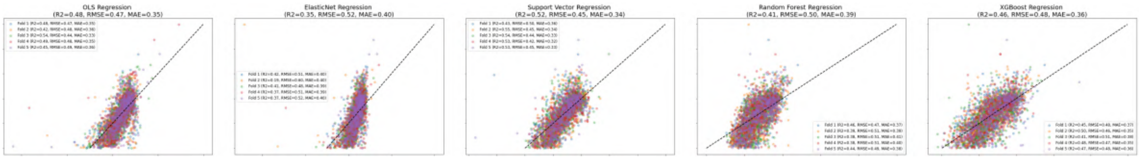


Figure H.20: FD 2, Outage

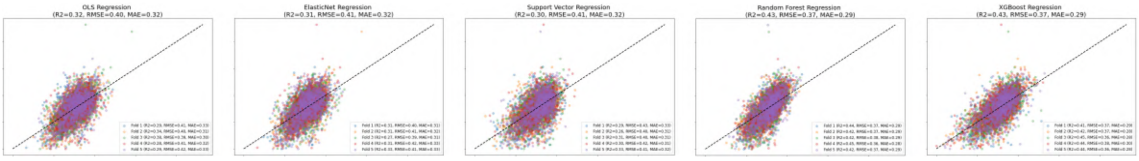


Figure H.21: FD 3, Product 1

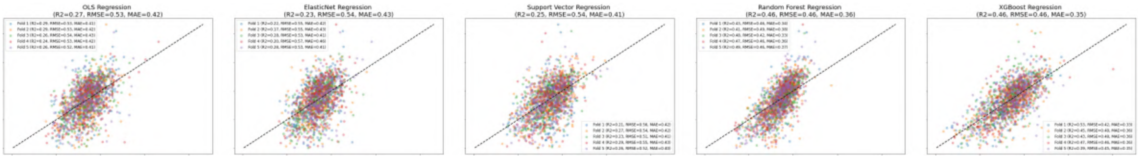


Figure H.22: FD 3, Product 2

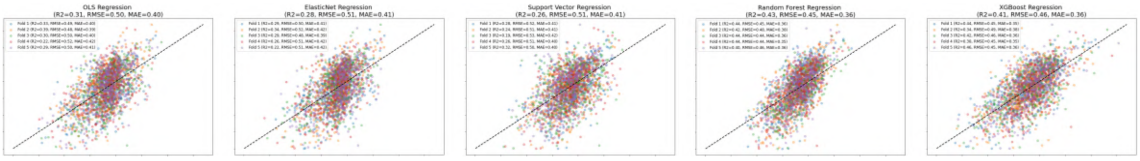


Figure H.23: FD 3, Product 3

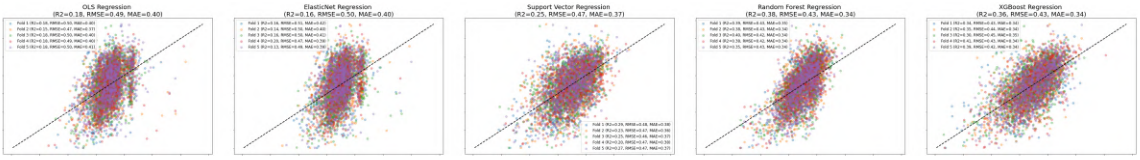


Figure H.24: FD 3, Product 4

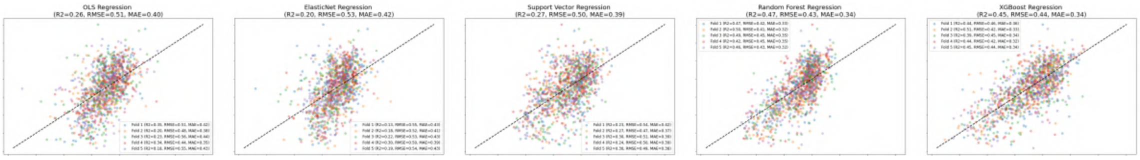


Figure H.25: FD 3, Product 5

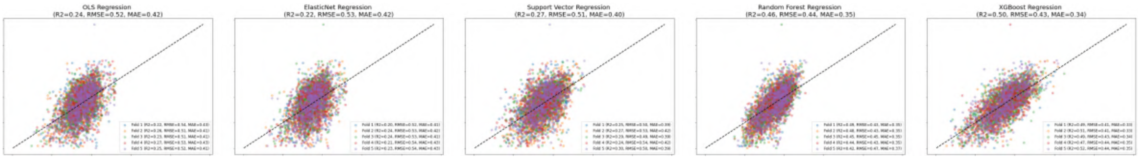


Figure H.26: FD 3, Product 6

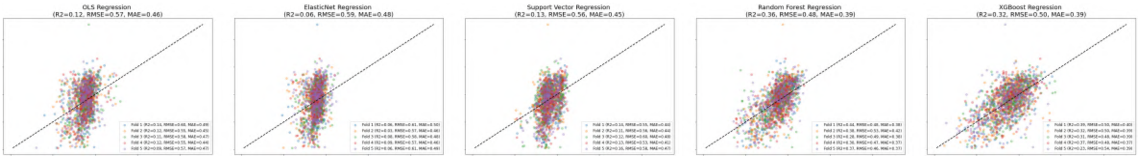


Figure H.27: FD 3, Product 7

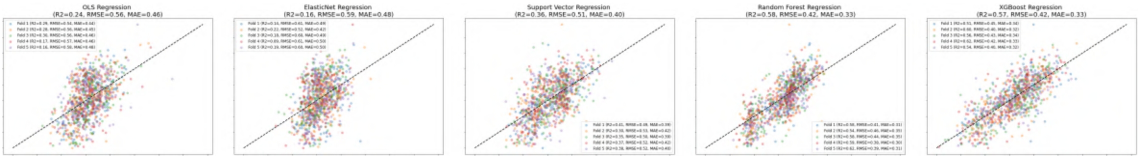


Figure H.28: FD 3, Product 8

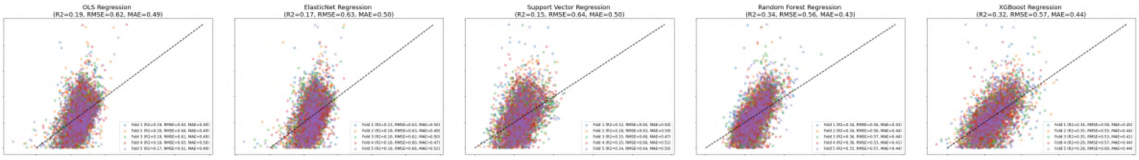


Figure H.29: FD 3, Outage

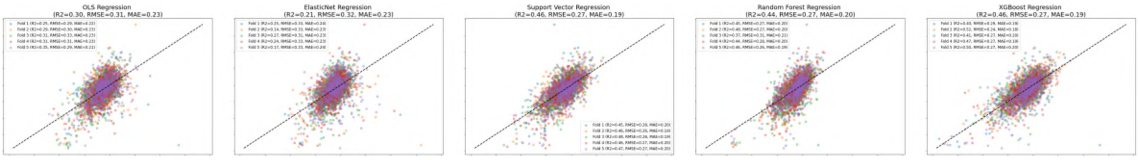


Figure H.30: FD 4, Product 1

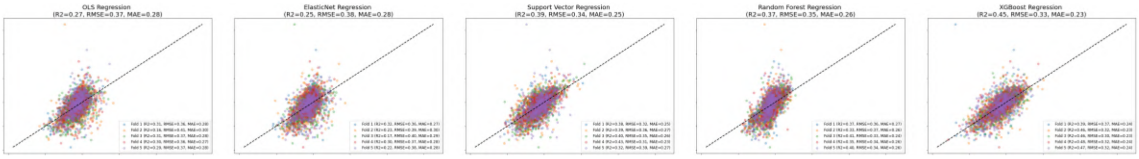


Figure H.31: FD 4, Product 2

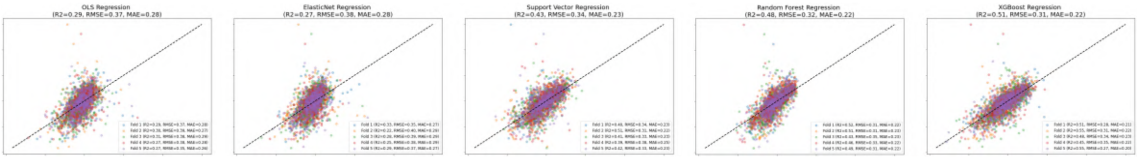


Figure H.32: FD 4, Product 3

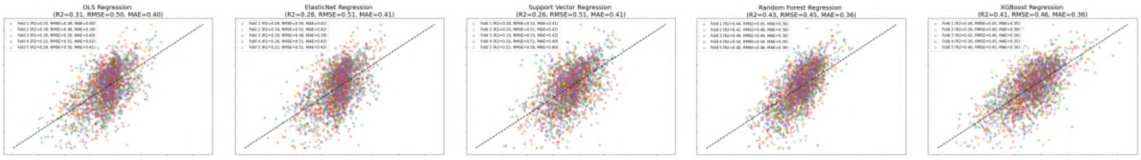


Figure H.33: FD 4, Product 4

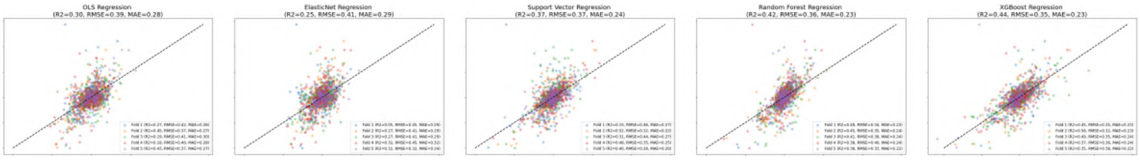


Figure H.34: FD 4, Product 5

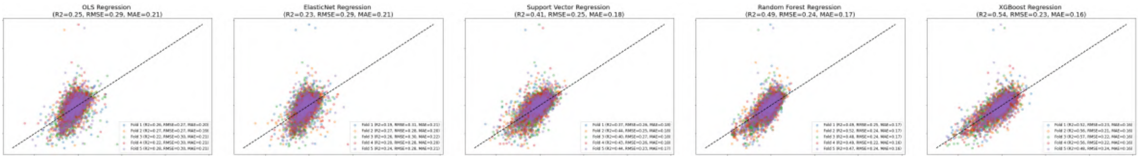


Figure H.35: FD 4, Product 6

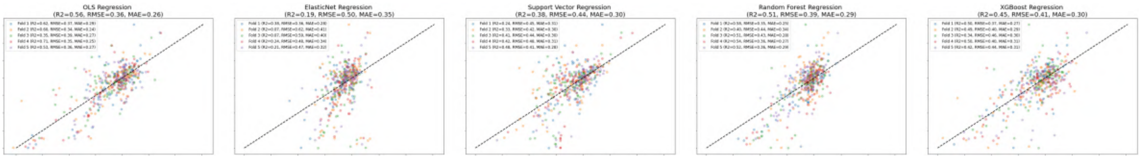


Figure H.36: FD 4, Product 7

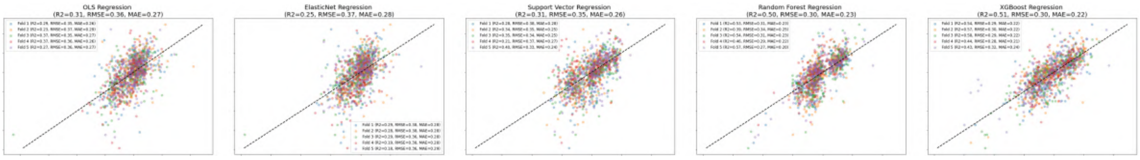


Figure H.37: FD 4, Product 8

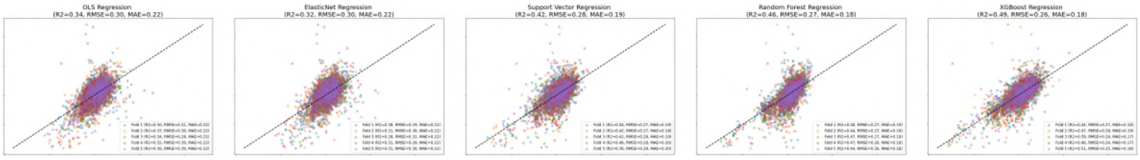


Figure H.38: FD 4, Product 9

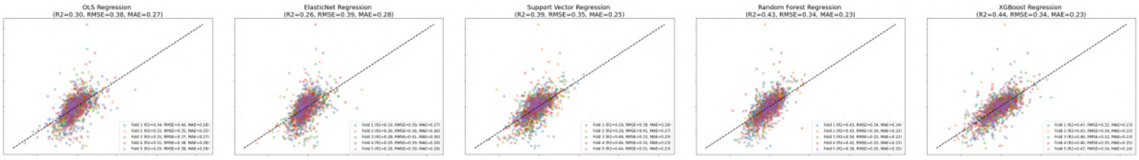


Figure H.39: FD 4, Product 10

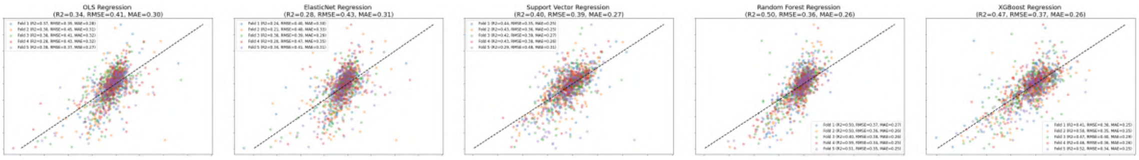


Figure H.40: FD 4, Product 11

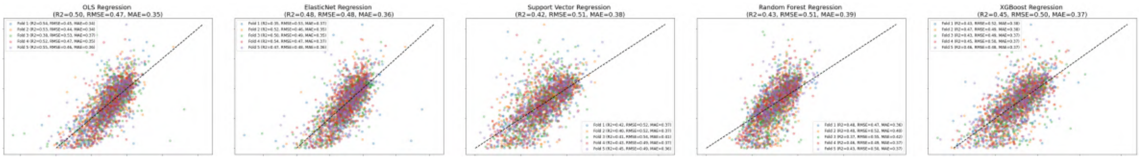
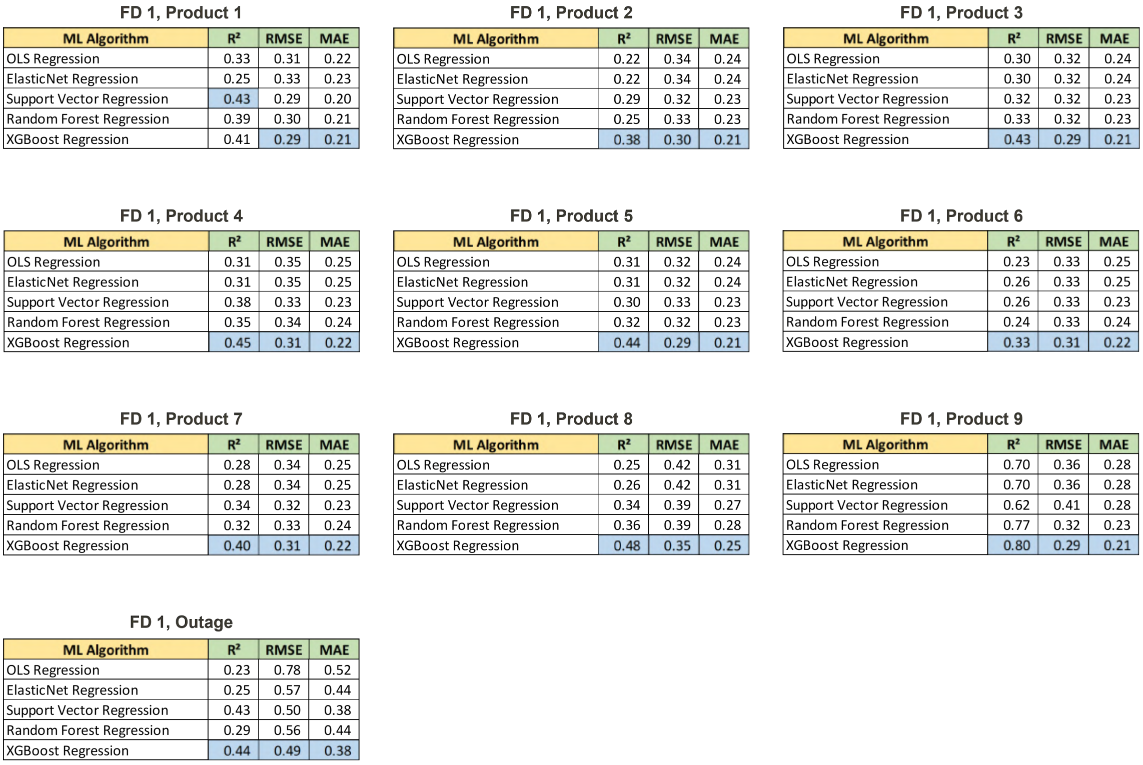


Figure H.41: FD 4, Outage

I

Appendix I: Machine Learning Algorithm Metrics of Different Products



FD 1, Product 1

ML Algorithm	R^2	RMSE	MAE
OLS Regression	0.33	0.31	0.22
ElasticNet Regression	0.25	0.33	0.23
Support Vector Regression	0.43	0.29	0.20
Random Forest Regression	0.39	0.30	0.21
XGBoost Regression	0.41	0.29	0.21

FD 1, Product 2

ML Algorithm	R^2	RMSE	MAE
OLS Regression	0.22	0.34	0.24
ElasticNet Regression	0.22	0.34	0.24
Support Vector Regression	0.29	0.32	0.23
Random Forest Regression	0.25	0.33	0.23
XGBoost Regression	0.38	0.30	0.21

FD 1, Product 3

ML Algorithm	R^2	RMSE	MAE
OLS Regression	0.30	0.32	0.24
ElasticNet Regression	0.30	0.32	0.24
Support Vector Regression	0.32	0.32	0.23
Random Forest Regression	0.33	0.32	0.23
XGBoost Regression	0.43	0.29	0.21

FD 1, Product 4

ML Algorithm	R^2	RMSE	MAE
OLS Regression	0.31	0.35	0.25
ElasticNet Regression	0.31	0.35	0.25
Support Vector Regression	0.38	0.33	0.23
Random Forest Regression	0.35	0.34	0.24
XGBoost Regression	0.45	0.31	0.22

FD 1, Product 5

ML Algorithm	R^2	RMSE	MAE
OLS Regression	0.31	0.32	0.24
ElasticNet Regression	0.31	0.32	0.24
Support Vector Regression	0.30	0.33	0.23
Random Forest Regression	0.32	0.32	0.23
XGBoost Regression	0.44	0.29	0.21

FD 1, Product 6

ML Algorithm	R^2	RMSE	MAE
OLS Regression	0.23	0.33	0.25
ElasticNet Regression	0.26	0.33	0.25
Support Vector Regression	0.26	0.33	0.23
Random Forest Regression	0.24	0.33	0.24
XGBoost Regression	0.33	0.31	0.22

FD 1, Product 7

ML Algorithm	R^2	RMSE	MAE
OLS Regression	0.28	0.34	0.25
ElasticNet Regression	0.28	0.34	0.25
Support Vector Regression	0.34	0.32	0.23
Random Forest Regression	0.32	0.33	0.24
XGBoost Regression	0.40	0.31	0.22

FD 1, Product 8

ML Algorithm	R^2	RMSE	MAE
OLS Regression	0.25	0.42	0.31
ElasticNet Regression	0.26	0.42	0.31
Support Vector Regression	0.34	0.39	0.27
Random Forest Regression	0.36	0.39	0.28
XGBoost Regression	0.48	0.35	0.25

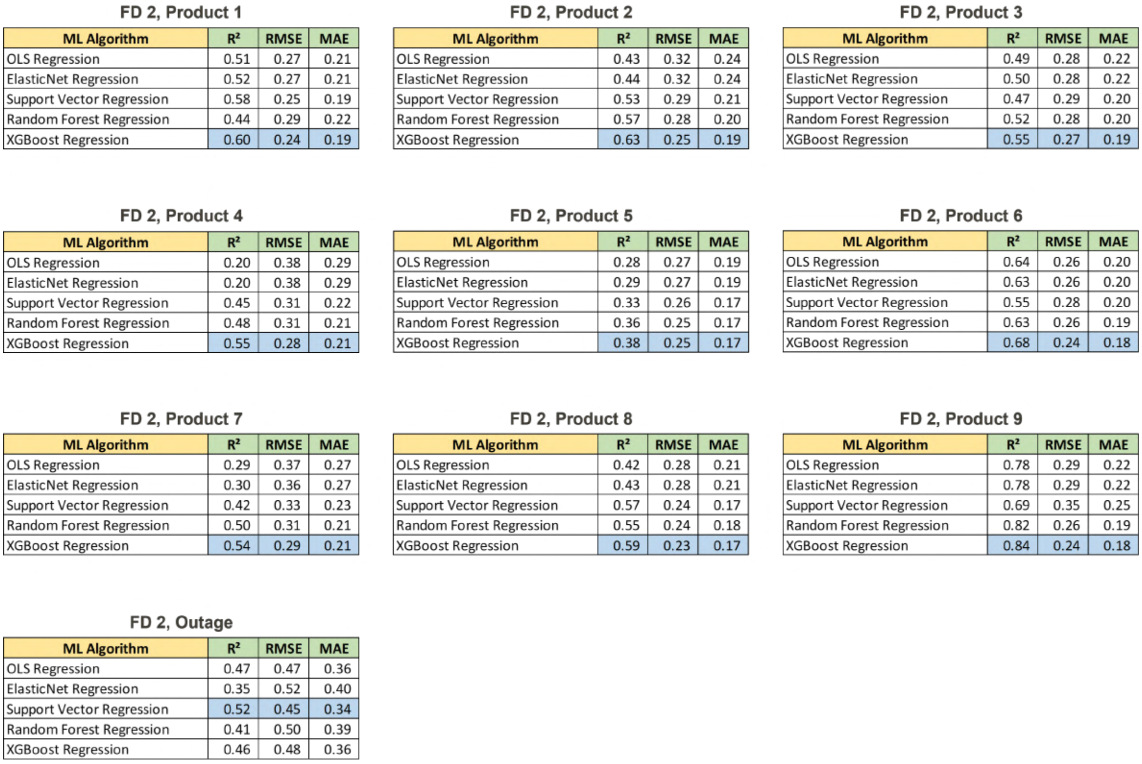
FD 1, Product 9

ML Algorithm	R^2	RMSE	MAE
OLS Regression	0.70	0.36	0.28
ElasticNet Regression	0.70	0.36	0.28
Support Vector Regression	0.62	0.41	0.28
Random Forest Regression	0.77	0.32	0.23
XGBoost Regression	0.80	0.29	0.21

FD 1, Outage

ML Algorithm	R^2	RMSE	MAE
OLS Regression	0.23	0.78	0.52
ElasticNet Regression	0.25	0.57	0.44
Support Vector Regression	0.43	0.50	0.38
Random Forest Regression	0.29	0.56	0.44
XGBoost Regression	0.44	0.49	0.38

Figure I.1: Machine Learning Algorithm Metrics of FD 1



FD 2, Product 1

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.51	0.27	0.21
ElasticNet Regression	0.52	0.27	0.21
Support Vector Regression	0.58	0.25	0.19
Random Forest Regression	0.44	0.29	0.22
XGBoost Regression	0.60	0.24	0.19

FD 2, Product 2

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.43	0.32	0.24
ElasticNet Regression	0.44	0.32	0.24
Support Vector Regression	0.53	0.29	0.21
Random Forest Regression	0.57	0.28	0.20
XGBoost Regression	0.63	0.25	0.19

FD 2, Product 3

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.49	0.28	0.22
ElasticNet Regression	0.50	0.28	0.22
Support Vector Regression	0.47	0.29	0.20
Random Forest Regression	0.52	0.28	0.20
XGBoost Regression	0.55	0.27	0.19

FD 2, Product 4

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.20	0.38	0.29
ElasticNet Regression	0.20	0.38	0.29
Support Vector Regression	0.45	0.31	0.22
Random Forest Regression	0.48	0.31	0.21
XGBoost Regression	0.55	0.28	0.21

FD 2, Product 5

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.28	0.27	0.19
ElasticNet Regression	0.29	0.27	0.19
Support Vector Regression	0.33	0.26	0.17
Random Forest Regression	0.36	0.25	0.17
XGBoost Regression	0.38	0.25	0.17

FD 2, Product 6

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.64	0.26	0.20
ElasticNet Regression	0.63	0.26	0.20
Support Vector Regression	0.55	0.28	0.20
Random Forest Regression	0.63	0.26	0.19
XGBoost Regression	0.68	0.24	0.18

FD 2, Product 7

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.29	0.37	0.27
ElasticNet Regression	0.30	0.36	0.27
Support Vector Regression	0.42	0.33	0.23
Random Forest Regression	0.50	0.31	0.21
XGBoost Regression	0.54	0.29	0.21

FD 2, Product 8

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.42	0.28	0.21
ElasticNet Regression	0.43	0.28	0.21
Support Vector Regression	0.57	0.24	0.17
Random Forest Regression	0.55	0.24	0.18
XGBoost Regression	0.59	0.23	0.17

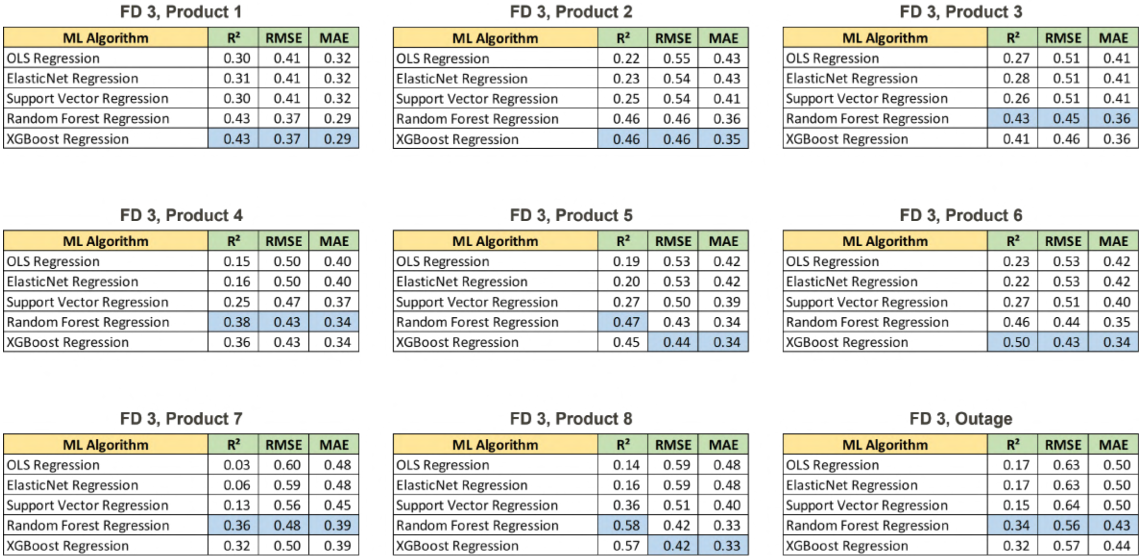
FD 2, Product 9

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.78	0.29	0.22
ElasticNet Regression	0.78	0.29	0.22
Support Vector Regression	0.69	0.35	0.25
Random Forest Regression	0.82	0.26	0.19
XGBoost Regression	0.84	0.24	0.18

FD 2, Outage

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.47	0.47	0.36
ElasticNet Regression	0.35	0.52	0.40
Support Vector Regression	0.52	0.45	0.34
Random Forest Regression	0.41	0.50	0.39
XGBoost Regression	0.46	0.48	0.36

Figure I.2: Machine Learning Algorithm Metrics of FD 2



FD 3, Product 1

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.30	0.41	0.32
ElasticNet Regression	0.31	0.41	0.32
Support Vector Regression	0.30	0.41	0.32
Random Forest Regression	0.43	0.37	0.29
XGBoost Regression	0.43	0.37	0.29

FD 3, Product 2

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.22	0.55	0.43
ElasticNet Regression	0.23	0.54	0.43
Support Vector Regression	0.25	0.54	0.41
Random Forest Regression	0.46	0.46	0.36
XGBoost Regression	0.46	0.46	0.35

FD 3, Product 3

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.27	0.51	0.41
ElasticNet Regression	0.28	0.51	0.41
Support Vector Regression	0.26	0.51	0.41
Random Forest Regression	0.43	0.45	0.36
XGBoost Regression	0.41	0.46	0.36

FD 3, Product 4

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.15	0.50	0.40
ElasticNet Regression	0.16	0.50	0.40
Support Vector Regression	0.25	0.47	0.37
Random Forest Regression	0.38	0.43	0.34
XGBoost Regression	0.36	0.43	0.34

FD 3, Product 5

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.19	0.53	0.42
ElasticNet Regression	0.20	0.53	0.42
Support Vector Regression	0.27	0.50	0.39
Random Forest Regression	0.47	0.43	0.34
XGBoost Regression	0.45	0.44	0.34

FD 3, Product 6

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.23	0.53	0.42
ElasticNet Regression	0.22	0.53	0.42
Support Vector Regression	0.27	0.51	0.40
Random Forest Regression	0.46	0.44	0.35
XGBoost Regression	0.50	0.43	0.34

FD 3, Product 7

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.03	0.60	0.48
ElasticNet Regression	0.06	0.59	0.48
Support Vector Regression	0.13	0.56	0.45
Random Forest Regression	0.36	0.48	0.39
XGBoost Regression	0.32	0.50	0.39

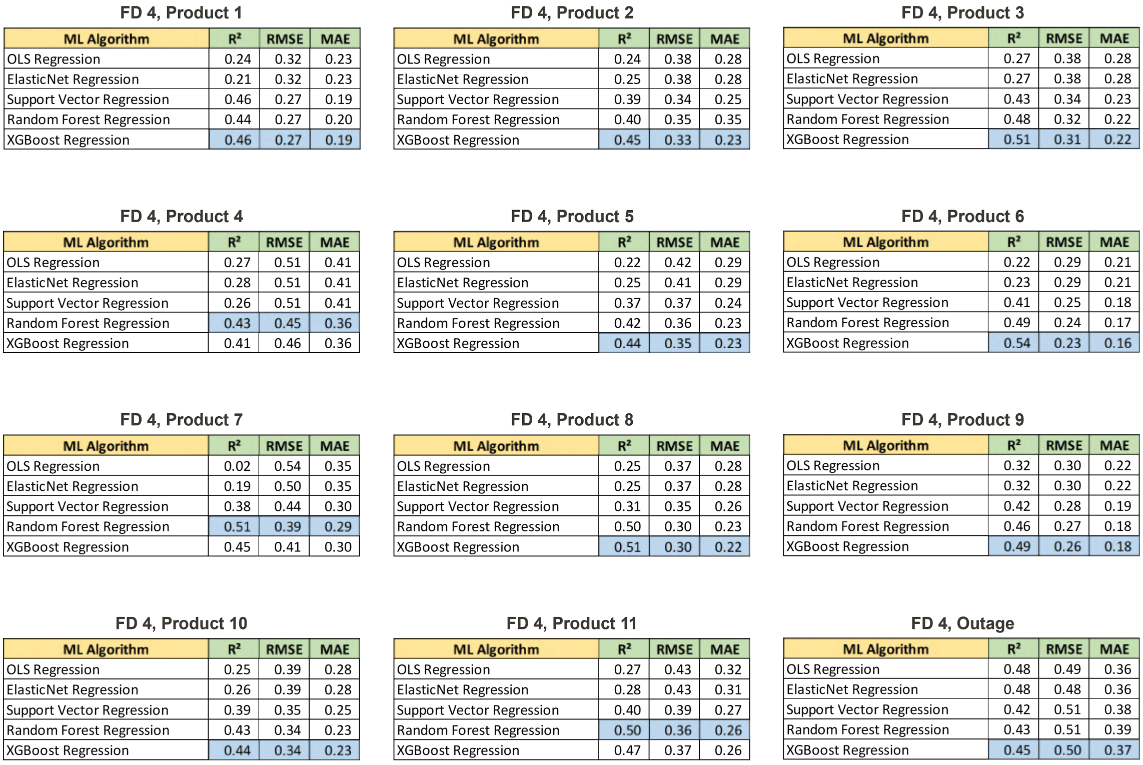
FD 3, Product 8

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.14	0.59	0.48
ElasticNet Regression	0.16	0.59	0.48
Support Vector Regression	0.36	0.51	0.40
Random Forest Regression	0.58	0.42	0.33
XGBoost Regression	0.57	0.42	0.33

FD 3, Outage

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.17	0.63	0.50
ElasticNet Regression	0.17	0.63	0.50
Support Vector Regression	0.15	0.64	0.50
Random Forest Regression	0.34	0.56	0.43
XGBoost Regression	0.32	0.57	0.44

Figure I.3: Machine Learning Algorithm Metrics of FD 3



FD 4, Product 1

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.24	0.32	0.23
ElasticNet Regression	0.21	0.32	0.23
Support Vector Regression	0.46	0.27	0.19
Random Forest Regression	0.44	0.27	0.20
XGBoost Regression	0.46	0.27	0.19

FD 4, Product 2

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.24	0.38	0.28
ElasticNet Regression	0.25	0.38	0.28
Support Vector Regression	0.39	0.34	0.25
Random Forest Regression	0.40	0.35	0.35
XGBoost Regression	0.45	0.33	0.23

FD 4, Product 3

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.27	0.38	0.28
ElasticNet Regression	0.27	0.38	0.28
Support Vector Regression	0.43	0.34	0.23
Random Forest Regression	0.48	0.32	0.22
XGBoost Regression	0.51	0.31	0.22

FD 4, Product 4

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.27	0.51	0.41
ElasticNet Regression	0.28	0.51	0.41
Support Vector Regression	0.26	0.51	0.41
Random Forest Regression	0.43	0.45	0.36
XGBoost Regression	0.41	0.46	0.36

FD 4, Product 5

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.22	0.42	0.29
ElasticNet Regression	0.25	0.41	0.29
Support Vector Regression	0.37	0.37	0.24
Random Forest Regression	0.42	0.36	0.23
XGBoost Regression	0.44	0.35	0.23

FD 4, Product 6

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.22	0.29	0.21
ElasticNet Regression	0.23	0.29	0.21
Support Vector Regression	0.41	0.25	0.18
Random Forest Regression	0.49	0.24	0.17
XGBoost Regression	0.54	0.23	0.16

FD 4, Product 7

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.02	0.54	0.35
ElasticNet Regression	0.19	0.50	0.35
Support Vector Regression	0.38	0.44	0.30
Random Forest Regression	0.51	0.39	0.29
XGBoost Regression	0.45	0.41	0.30

FD 4, Product 8

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.25	0.37	0.28
ElasticNet Regression	0.25	0.37	0.28
Support Vector Regression	0.31	0.35	0.26
Random Forest Regression	0.50	0.30	0.23
XGBoost Regression	0.51	0.30	0.22

FD 4, Product 9

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.32	0.30	0.22
ElasticNet Regression	0.32	0.30	0.22
Support Vector Regression	0.42	0.28	0.19
Random Forest Regression	0.46	0.27	0.18
XGBoost Regression	0.49	0.26	0.18

FD 4, Product 10

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.25	0.39	0.28
ElasticNet Regression	0.26	0.39	0.28
Support Vector Regression	0.39	0.35	0.25
Random Forest Regression	0.43	0.34	0.23
XGBoost Regression	0.44	0.34	0.23

FD 4, Product 11

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.27	0.43	0.32
ElasticNet Regression	0.28	0.43	0.31
Support Vector Regression	0.40	0.39	0.27
Random Forest Regression	0.50	0.36	0.26
XGBoost Regression	0.47	0.37	0.26

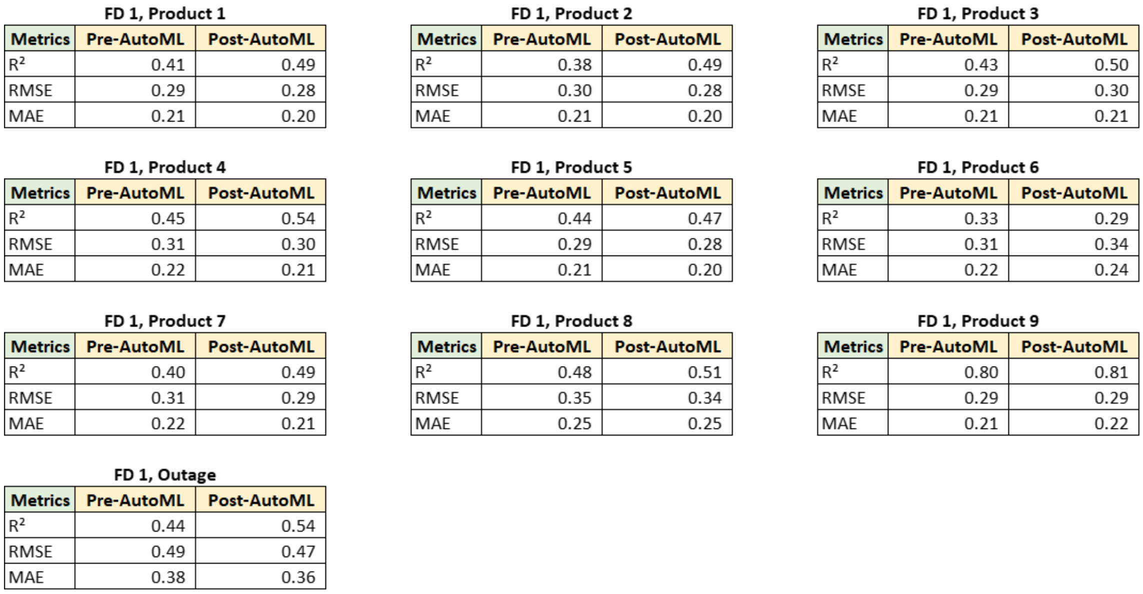
FD 4, Outage

ML Algorithm	R²	RMSE	MAE
OLS Regression	0.48	0.49	0.36
ElasticNet Regression	0.48	0.48	0.36
Support Vector Regression	0.42	0.51	0.38
Random Forest Regression	0.43	0.51	0.39
XGBoost Regression	0.45	0.50	0.37

Figure I.4: Machine Learning Algorithm Metrics of FD 4

J

Appendix J: Metrics Scores Comparison Pre- and Post-AutoML



FD 1, Product 1

Metrics	Pre-AutoML	Post-AutoML
R^2	0.41	0.49
RMSE	0.29	0.28
MAE	0.21	0.20

FD 1, Product 2

Metrics	Pre-AutoML	Post-AutoML
R^2	0.38	0.49
RMSE	0.30	0.28
MAE	0.21	0.20

FD 1, Product 3

Metrics	Pre-AutoML	Post-AutoML
R^2	0.43	0.50
RMSE	0.29	0.30
MAE	0.21	0.21

FD 1, Product 4

Metrics	Pre-AutoML	Post-AutoML
R^2	0.45	0.54
RMSE	0.31	0.30
MAE	0.22	0.21

FD 1, Product 5

Metrics	Pre-AutoML	Post-AutoML
R^2	0.44	0.47
RMSE	0.29	0.28
MAE	0.21	0.20

FD 1, Product 6

Metrics	Pre-AutoML	Post-AutoML
R^2	0.33	0.29
RMSE	0.31	0.34
MAE	0.22	0.24

FD 1, Product 7

Metrics	Pre-AutoML	Post-AutoML
R^2	0.40	0.49
RMSE	0.31	0.29
MAE	0.22	0.21

FD 1, Product 8

Metrics	Pre-AutoML	Post-AutoML
R^2	0.48	0.51
RMSE	0.35	0.34
MAE	0.25	0.25

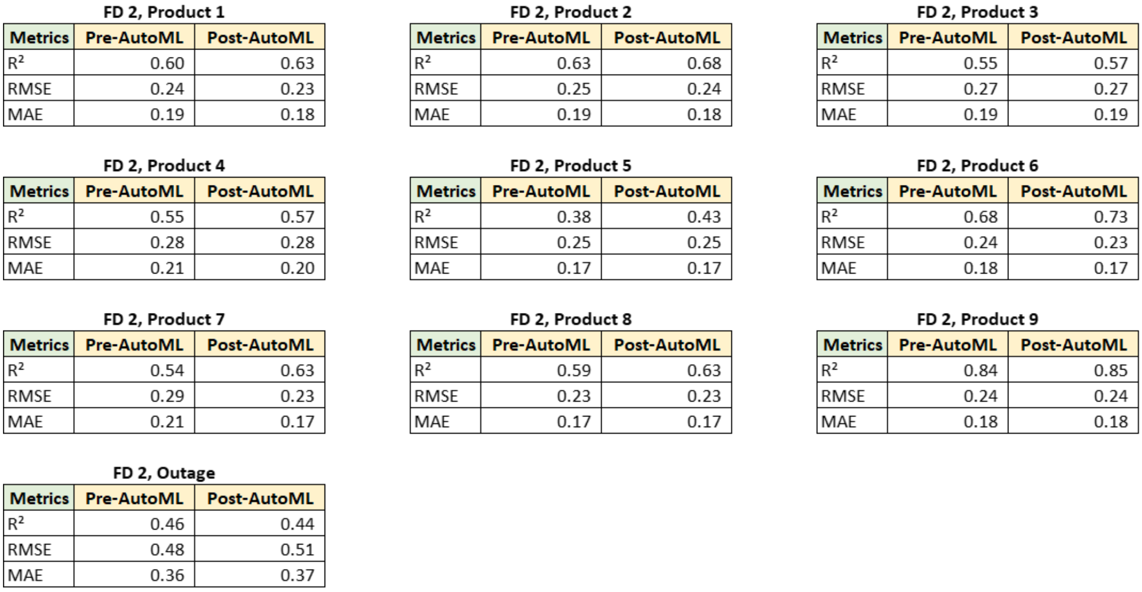
FD 1, Product 9

Metrics	Pre-AutoML	Post-AutoML
R^2	0.80	0.81
RMSE	0.29	0.29
MAE	0.21	0.22

FD 1, Outage

Metrics	Pre-AutoML	Post-AutoML
R^2	0.44	0.54
RMSE	0.49	0.47
MAE	0.38	0.36

Figure J.1: Metrics Scores Comparison Pre- and Post-AutoML in FD 1



FD 2, Product 1

Metrics	Pre-AutoML	Post-AutoML
R^2	0.60	0.63
RMSE	0.24	0.23
MAE	0.19	0.18

FD 2, Product 2

Metrics	Pre-AutoML	Post-AutoML
R^2	0.63	0.68
RMSE	0.25	0.24
MAE	0.19	0.18

FD 2, Product 3

Metrics	Pre-AutoML	Post-AutoML
R^2	0.55	0.57
RMSE	0.27	0.27
MAE	0.19	0.19

FD 2, Product 4

Metrics	Pre-AutoML	Post-AutoML
R^2	0.55	0.57
RMSE	0.28	0.28
MAE	0.21	0.20

FD 2, Product 5

Metrics	Pre-AutoML	Post-AutoML
R^2	0.38	0.43
RMSE	0.25	0.25
MAE	0.17	0.17

FD 2, Product 6

Metrics	Pre-AutoML	Post-AutoML
R^2	0.68	0.73
RMSE	0.24	0.23
MAE	0.18	0.17

FD 2, Product 7

Metrics	Pre-AutoML	Post-AutoML
R^2	0.54	0.63
RMSE	0.29	0.23
MAE	0.21	0.17

FD 2, Product 8

Metrics	Pre-AutoML	Post-AutoML
R^2	0.59	0.63
RMSE	0.23	0.23
MAE	0.17	0.17

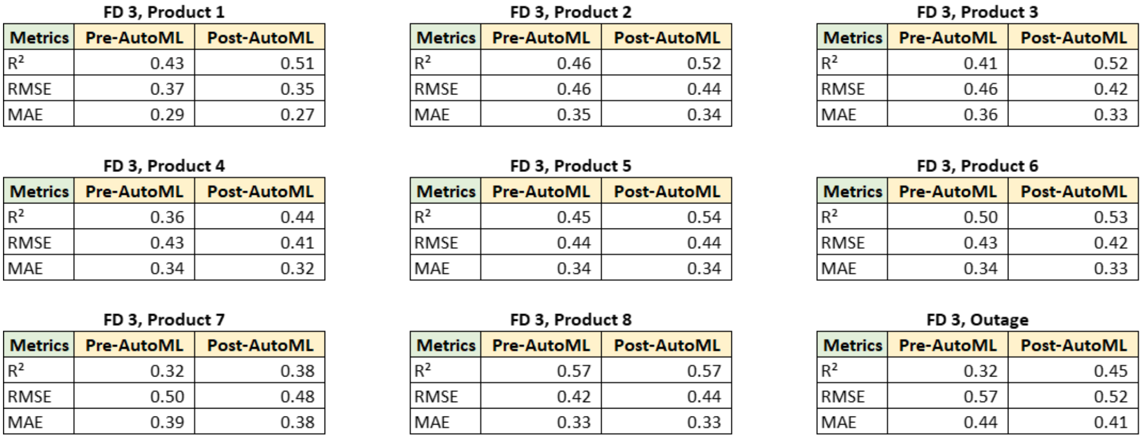
FD 2, Product 9

Metrics	Pre-AutoML	Post-AutoML
R^2	0.84	0.85
RMSE	0.24	0.24
MAE	0.18	0.18

FD 2, Outage

Metrics	Pre-AutoML	Post-AutoML
R^2	0.46	0.44
RMSE	0.48	0.51
MAE	0.36	0.37

Figure J.2: Metrics Scores Comparison Pre- and Post-AutoML in FD 2



FD 3, Product 1

Metrics	Pre-AutoML	Post-AutoML
R^2	0.43	0.51
RMSE	0.37	0.35
MAE	0.29	0.27

FD 3, Product 2

Metrics	Pre-AutoML	Post-AutoML
R^2	0.46	0.52
RMSE	0.46	0.44
MAE	0.35	0.34

FD 3, Product 3

Metrics	Pre-AutoML	Post-AutoML
R^2	0.41	0.52
RMSE	0.46	0.42
MAE	0.36	0.33

FD 3, Product 4

Metrics	Pre-AutoML	Post-AutoML
R^2	0.36	0.44
RMSE	0.43	0.41
MAE	0.34	0.32

FD 3, Product 5

Metrics	Pre-AutoML	Post-AutoML
R^2	0.45	0.54
RMSE	0.44	0.44
MAE	0.34	0.34

FD 3, Product 6

Metrics	Pre-AutoML	Post-AutoML
R^2	0.50	0.53
RMSE	0.43	0.42
MAE	0.34	0.33

FD 3, Product 7

Metrics	Pre-AutoML	Post-AutoML
R^2	0.32	0.38
RMSE	0.50	0.48
MAE	0.39	0.38

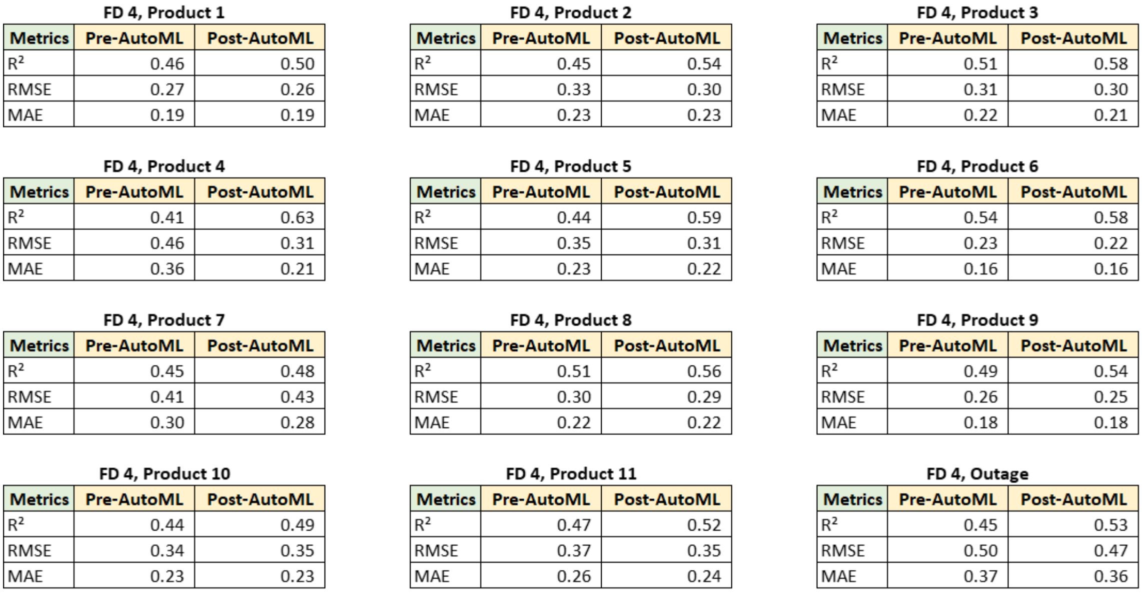
FD 3, Product 8

Metrics	Pre-AutoML	Post-AutoML
R^2	0.57	0.57
RMSE	0.42	0.44
MAE	0.33	0.33

FD 3, Outage

Metrics	Pre-AutoML	Post-AutoML
R^2	0.32	0.45
RMSE	0.57	0.52
MAE	0.44	0.41

Figure J.3: Metrics Scores Comparison Pre- and Post-AutoML in FD 3



FD 4, Product 1

Metrics	Pre-AutoML	Post-AutoML
R²	0.46	0.50
RMSE	0.27	0.26
MAE	0.19	0.19

FD 4, Product 2

Metrics	Pre-AutoML	Post-AutoML
R²	0.45	0.54
RMSE	0.33	0.30
MAE	0.23	0.23

FD 4, Product 3

Metrics	Pre-AutoML	Post-AutoML
R²	0.51	0.58
RMSE	0.31	0.30
MAE	0.22	0.21

FD 4, Product 4

Metrics	Pre-AutoML	Post-AutoML
R²	0.41	0.63
RMSE	0.46	0.31
MAE	0.36	0.21

FD 4, Product 5

Metrics	Pre-AutoML	Post-AutoML
R²	0.44	0.59
RMSE	0.35	0.31
MAE	0.23	0.22

FD 4, Product 6

Metrics	Pre-AutoML	Post-AutoML
R²	0.54	0.58
RMSE	0.23	0.22
MAE	0.16	0.16

FD 4, Product 7

Metrics	Pre-AutoML	Post-AutoML
R²	0.45	0.48
RMSE	0.41	0.43
MAE	0.30	0.28

FD 4, Product 8

Metrics	Pre-AutoML	Post-AutoML
R²	0.51	0.56
RMSE	0.30	0.29
MAE	0.22	0.22

FD 4, Product 9

Metrics	Pre-AutoML	Post-AutoML
R²	0.49	0.54
RMSE	0.26	0.25
MAE	0.18	0.18

FD 4, Product 10

Metrics	Pre-AutoML	Post-AutoML
R²	0.44	0.49
RMSE	0.34	0.35
MAE	0.23	0.23

FD 4, Product 11

Metrics	Pre-AutoML	Post-AutoML
R²	0.47	0.52
RMSE	0.37	0.35
MAE	0.26	0.24

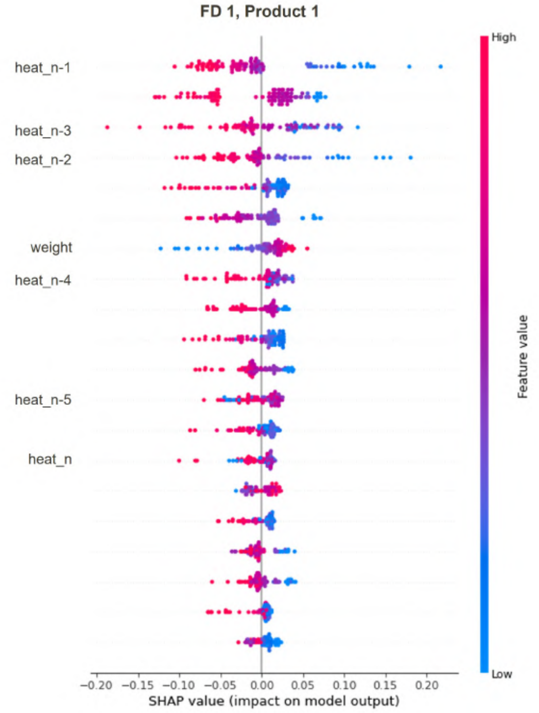
FD 4, Outage

Metrics	Pre-AutoML	Post-AutoML
R²	0.45	0.53
RMSE	0.50	0.47
MAE	0.37	0.36

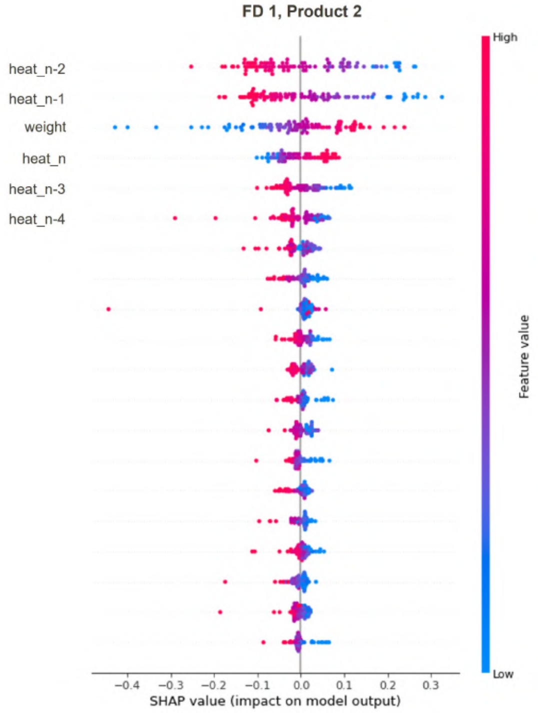
Figure J.4: Metrics Scores Comparison Pre- and Post-AutoML in FD 4

K

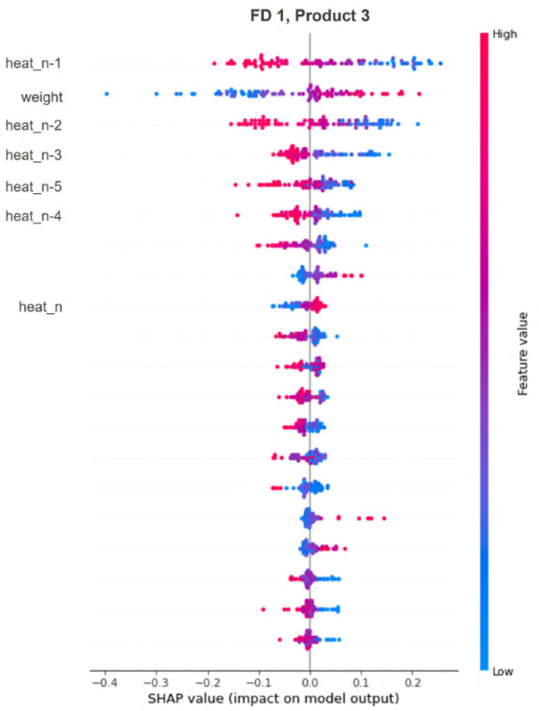
Appendix K: SHAP Plots of the XGBoost Models



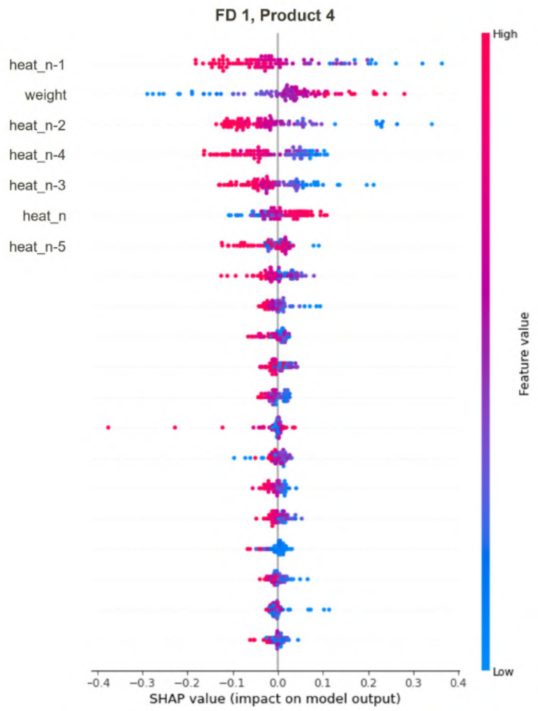
FD 1, Product 1



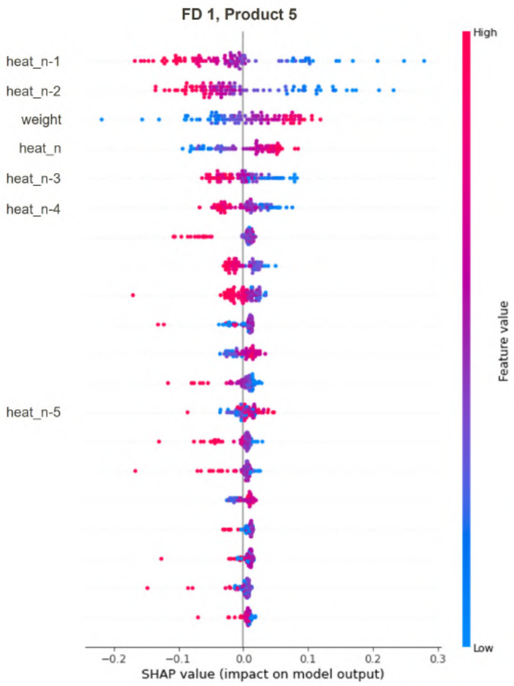
FD 1, Product 2

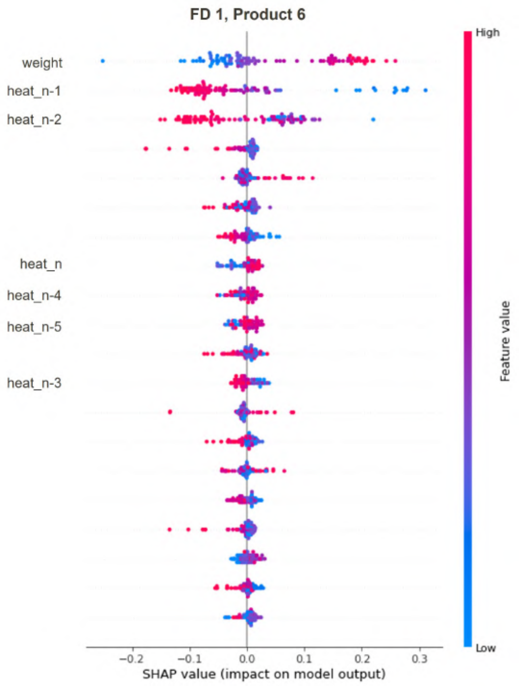


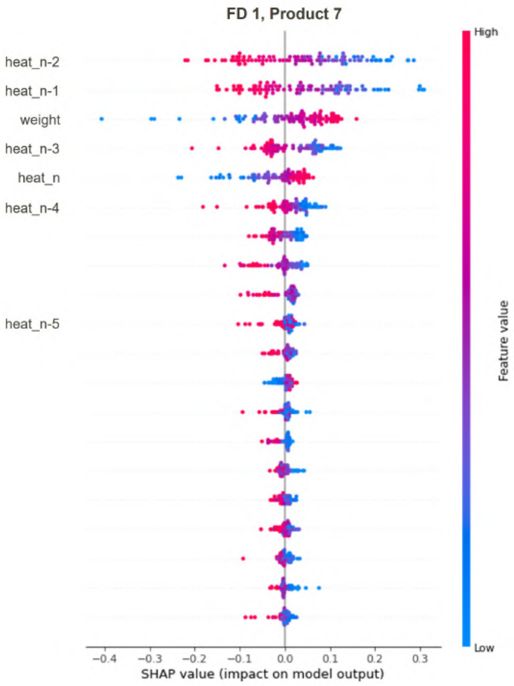
FD 1, Product 3

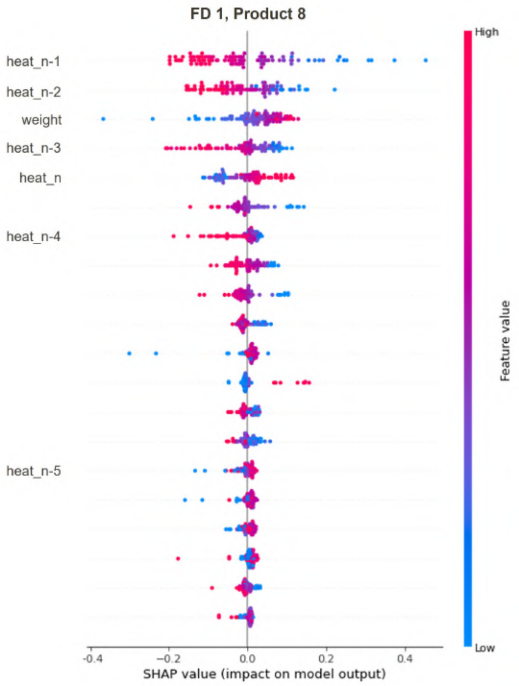


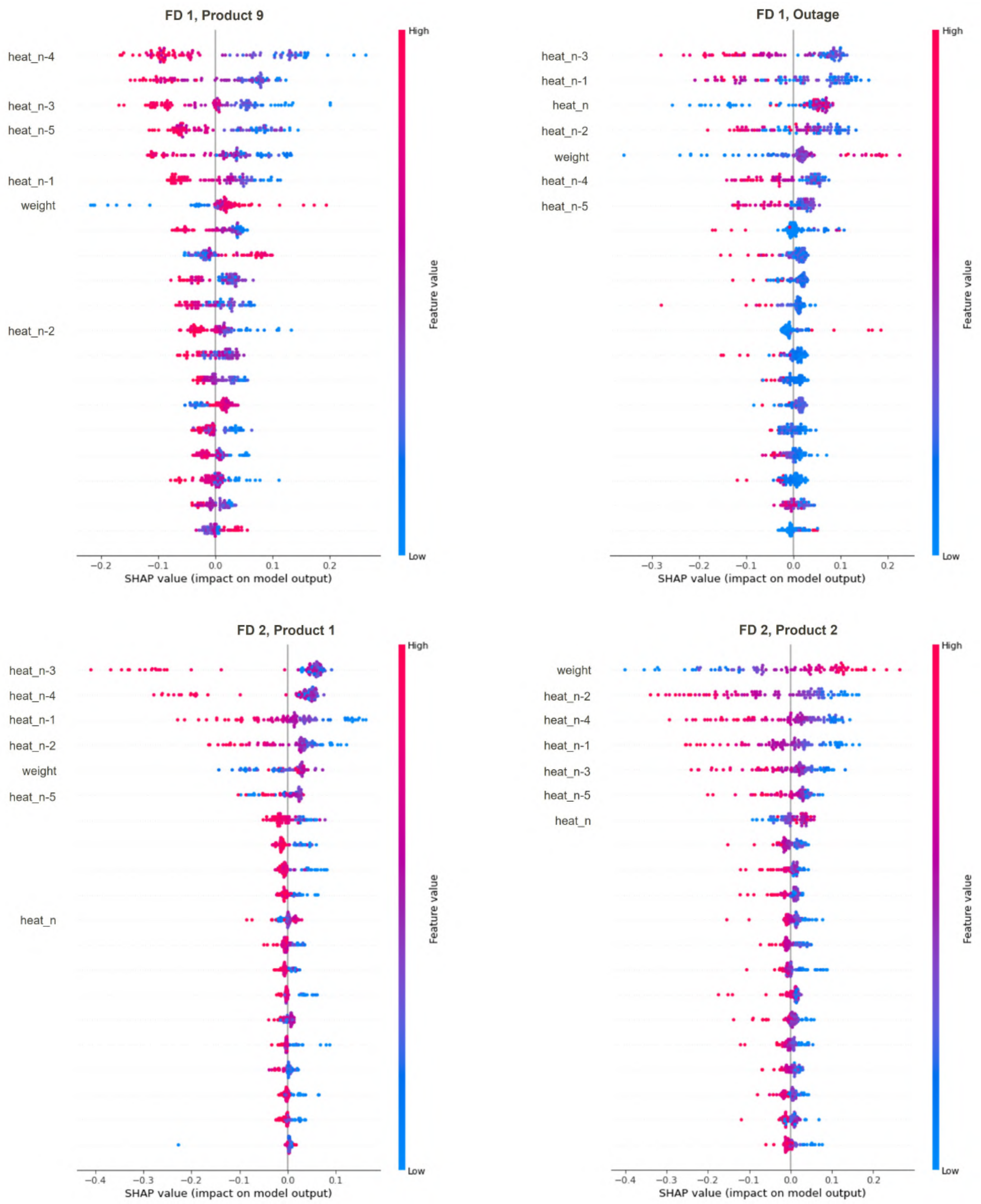
FD 1, Product 4

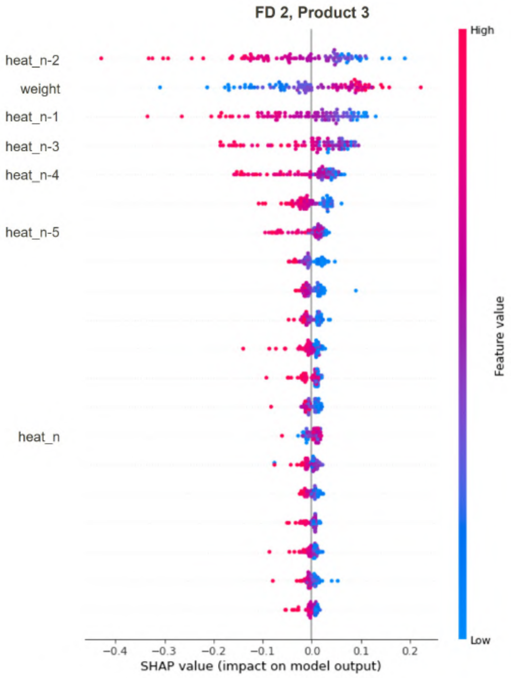

FD 1, Product 5
heat_n-1
heat_n-2
weight
heat_n
heat_n-3
heat_n-4
heat_n-5
−0.2
−0.1
0.0
0.1
0.2
0.3
SHAP value (impact on model output)
High
Low
Feature value

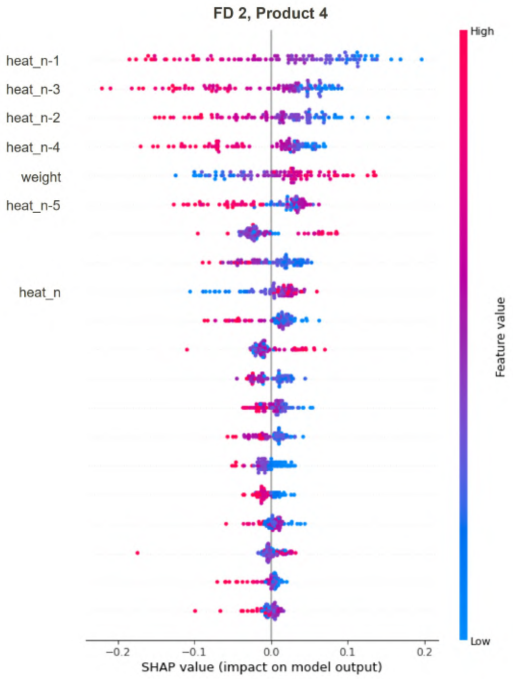

FD 1, Product 6
weight
heat_n-1
heat_n-2
heat_n
heat_n-4
heat_n-5
heat_n-3
−0.2
−0.1
0.0
0.1
0.2
0.3
SHAP value (impact on model output)
High
Low
Feature value

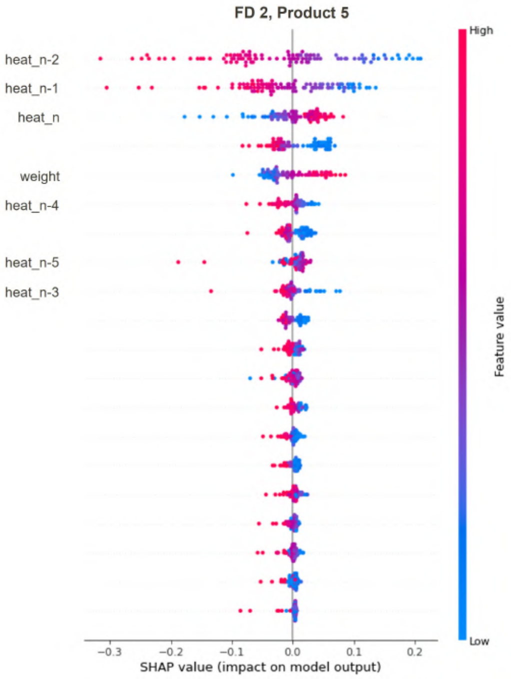

FD 1, Product 7
heat_n-2
heat_n-1
weight
heat_n-3
heat_n
heat_n-4
heat_n-5
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
SHAP value (impact on model output)
High
Low
Feature value

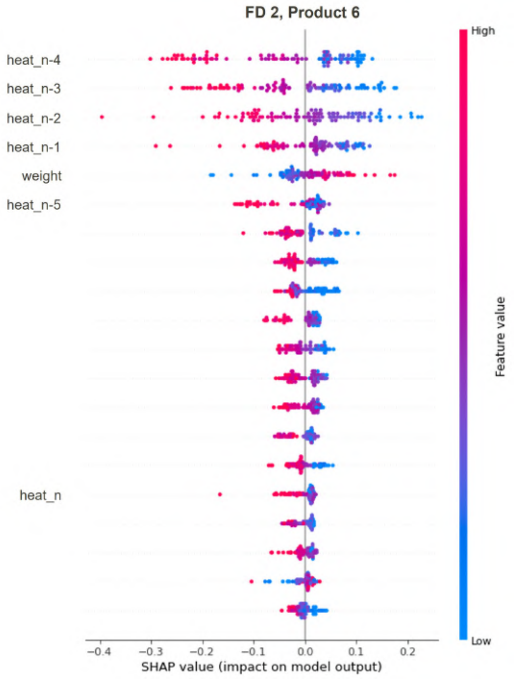

FD 1, Product 8
heat_n-1
heat_n-2
weight
heat_n-3
heat_n
heat_n-4
heat_n-5
-0.4
-0.2
0.0
0.2
0.4
SHAP value (impact on model output)
High
Low
Feature value

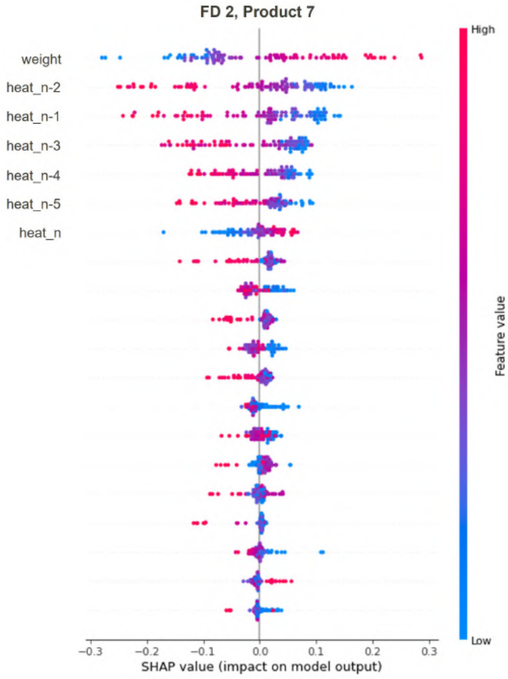

FD 1, Product 9
heat_n-4
heat_n-3
heat_n-5
heat_n-1
weight
heat_n-2
−0.2
−0.1
0.0
0.1
0.2
SHAP value (impact on model output)
High
Feature value
Low
FD 1, Outage
heat_n-3
heat_n-1
heat_n
heat_n-2
weight
heat_n-4
heat_n-5
−0.3
−0.2
−0.1
0.0
0.1
0.2
SHAP value (impact on model output)
High
Feature value
Low
FD 2, Product 1
heat_n-3
heat_n-4
heat_n-1
heat_n-2
weight
heat_n-5
heat_n
−0.4
−0.3
−0.2
−0.1
0.0
0.1
SHAP value (impact on model output)
High
Feature value
Low
FD 2, Product 2
weight
heat_n-2
heat_n-4
heat_n-1
heat_n-3
heat_n-5
heat_n
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
SHAP value (impact on model output)
High
Feature value
Low

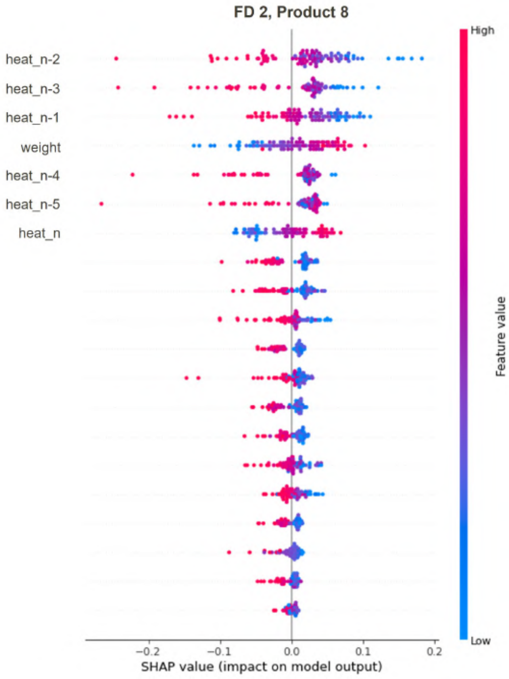

FD 2, Product 3
heat_n-2
weight
heat_n-1
heat_n-3
heat_n-4
heat_n-5
heat_n
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
SHAP value (impact on model output)
High
Feature value
Low

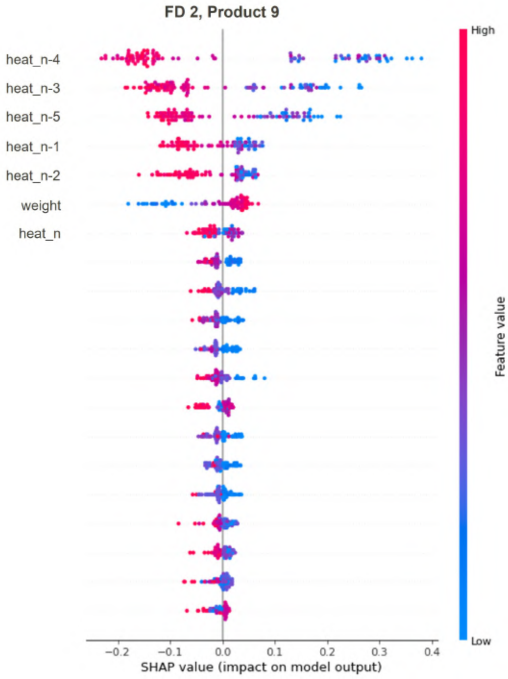

FD 2, Product 4
heat_n-1
heat_n-3
heat_n-2
heat_n-4
weight
heat_n-5
heat_n
−0.2
−0.1
0.0
0.1
0.2
SHAP value (impact on model output)
High
Feature value
Low

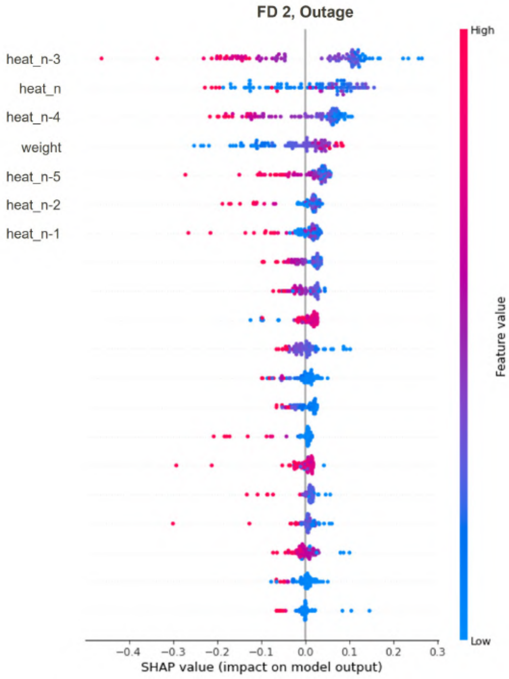

FD 2, Product 5
heat_n-2
heat_n-1
heat_n
weight
heat_n-4
heat_n-5
heat_n-3
−0.3
−0.2
−0.1
0.0
0.1
0.2
SHAP value (impact on model output)
High
Feature value
Low

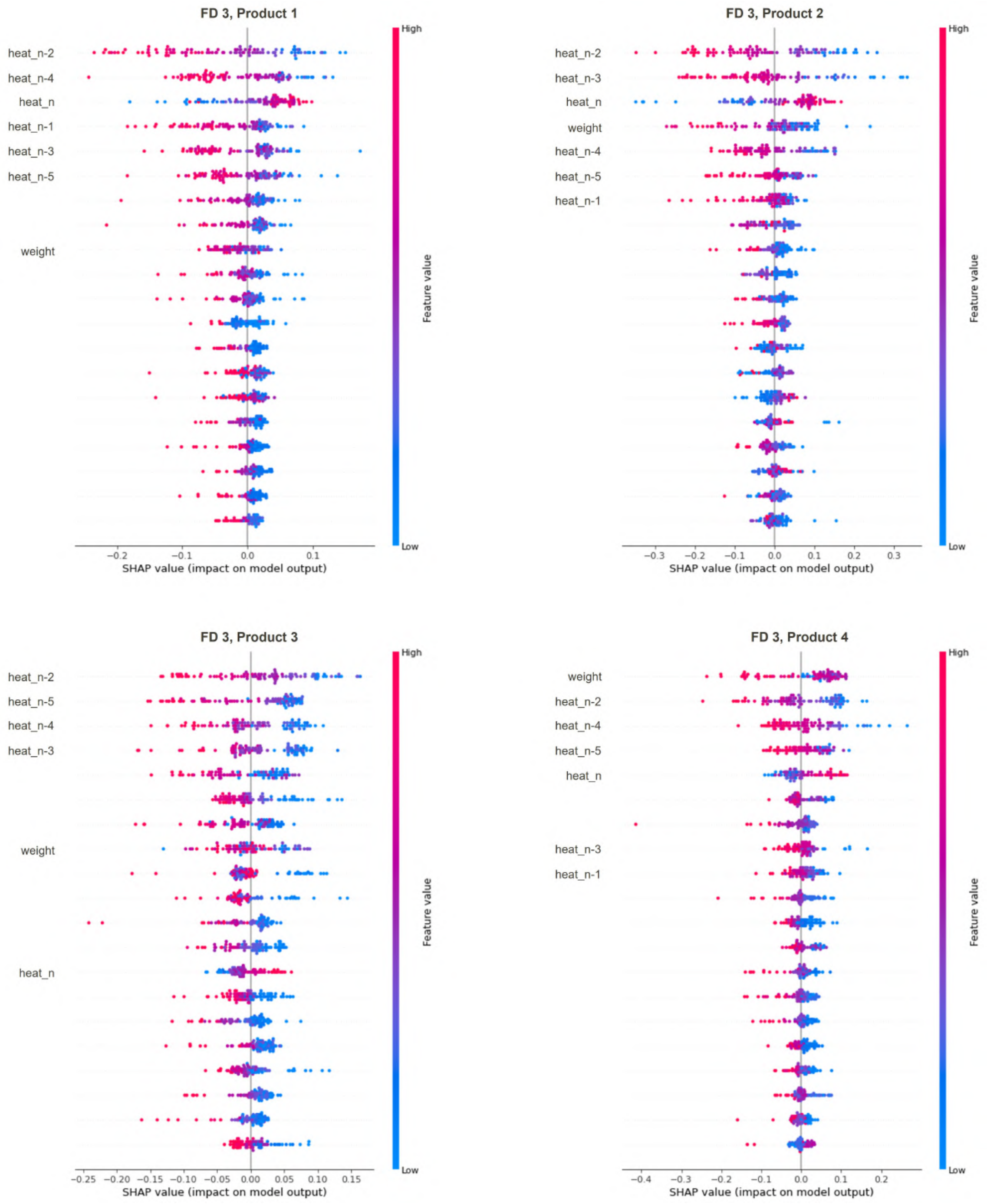

FD 2, Product 6
heat_n-4
heat_n-3
heat_n-2
heat_n-1
weight
heat_n-5
heat_n
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
SHAP value (impact on model output)
High
Feature value
Low

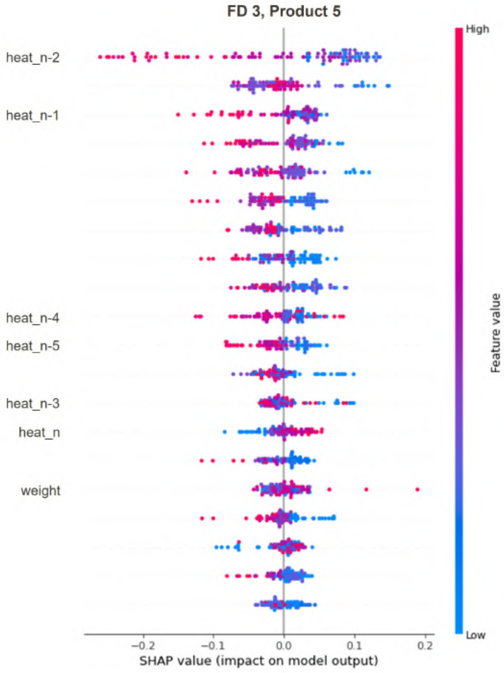

FD 2, Product 7
weight
heat_n-2
heat_n-1
heat_n-3
heat_n-4
heat_n-5
heat_n
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
SHAP value (impact on model output)
High
Feature value
Low

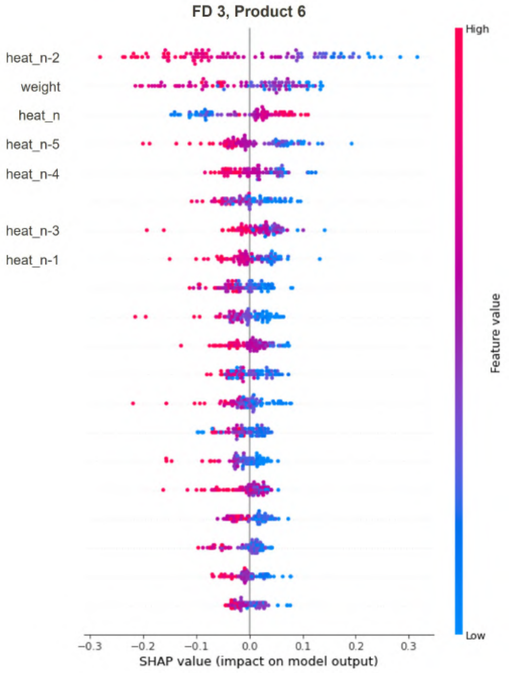

FD 2, Product 8
heat_n-2
heat_n-3
heat_n-1
weight
heat_n-4
heat_n-5
heat_n
−0.2
−0.1
0.0
0.1
0.2
SHAP value (impact on model output)
High
Feature value
Low

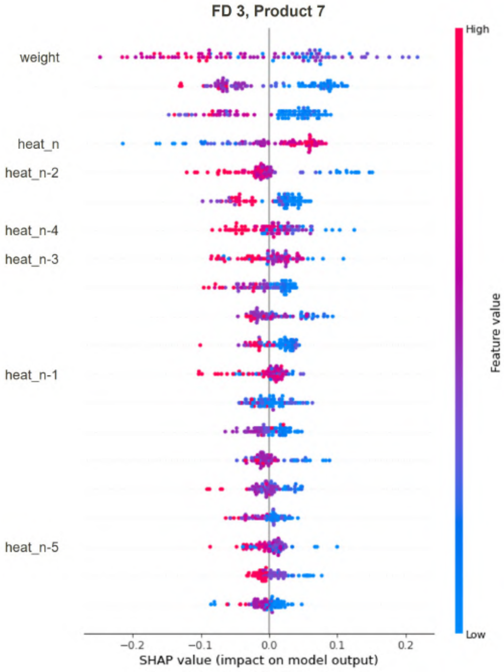

FD 2, Product 9
heat_n-4
heat_n-3
heat_n-5
heat_n-1
heat_n-2
weight
heat_n
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
SHAP value (impact on model output)
High
Feature value
Low

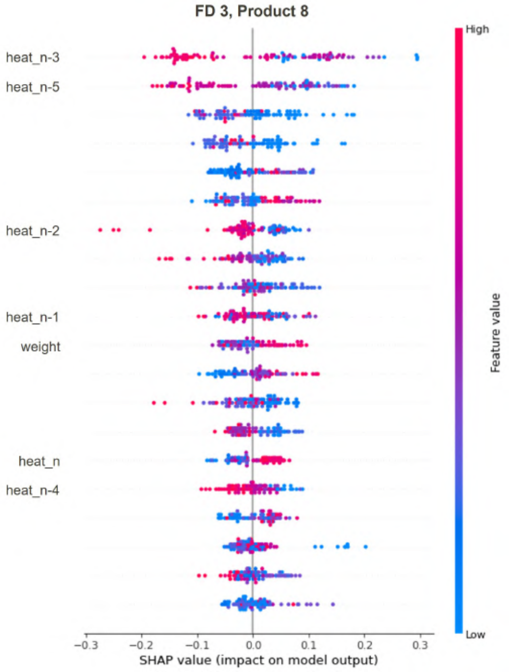

FD 2, Outage
heat_n-3
heat_n
heat_n-4
weight
heat_n-5
heat_n-2
heat_n-1
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
SHAP value (impact on model output)
High
Feature value
Low

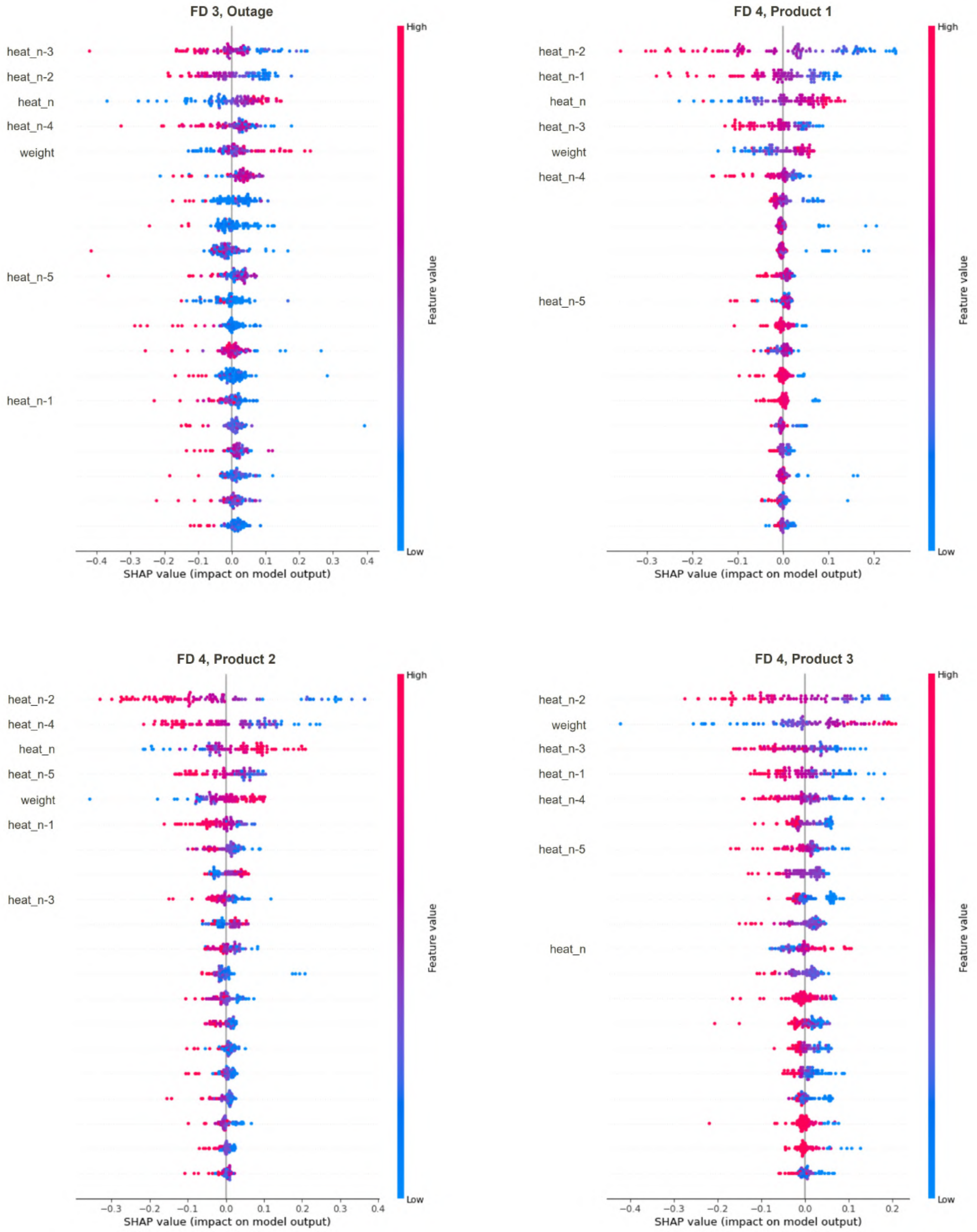

FD 3, Product 1
heat_n-2
heat_n-4
heat_n
heat_n-1
heat_n-3
heat_n-5
weight
−0.2
−0.1
0.0
0.1
SHAP value (impact on model output)
High
Feature value
Low
FD 3, Product 2
heat_n-2
heat_n-3
heat_n
weight
heat_n-4
heat_n-5
heat_n-1
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
SHAP value (impact on model output)
High
Feature value
Low
FD 3, Product 3
heat_n-2
heat_n-5
heat_n-4
heat_n-3
weight
heat_n
−0.25
−0.20
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
SHAP value (impact on model output)
High
Feature value
Low
FD 3, Product 4
weight
heat_n-2
heat_n-4
heat_n-5
heat_n
heat_n-3
heat_n-1
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
SHAP value (impact on model output)
High
Feature value
Low

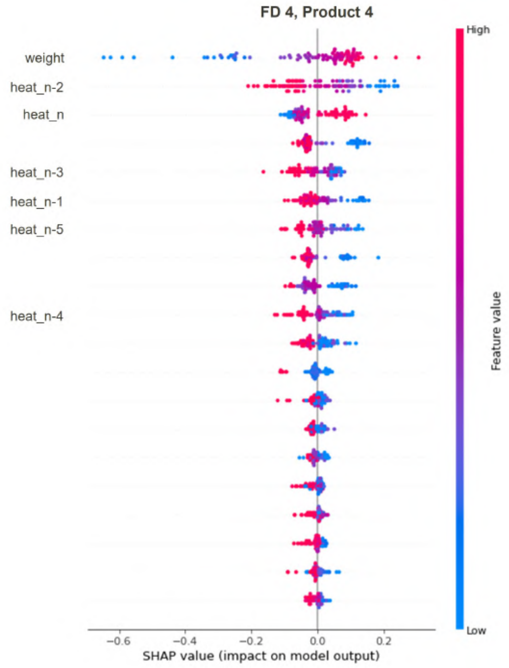

FD 3, Product 5
heat_n-2
heat_n-1
heat_n-4
heat_n-5
heat_n-3
heat_n
weight
High
Low
Feature value
−0.2
−0.1
0.0
0.1
0.2
SHAP value (impact on model output)

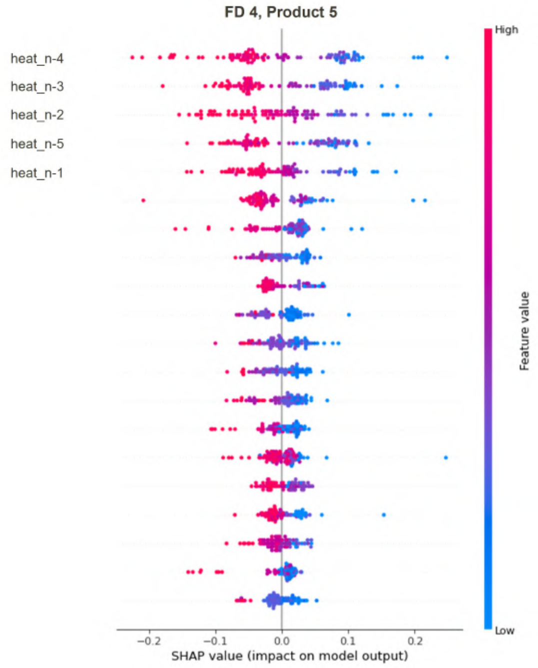

FD 3, Product 6
heat_n-2
weight
heat_n
heat_n-5
heat_n-4
heat_n-3
heat_n-1
High
Low
Feature value
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
SHAP value (impact on model output)

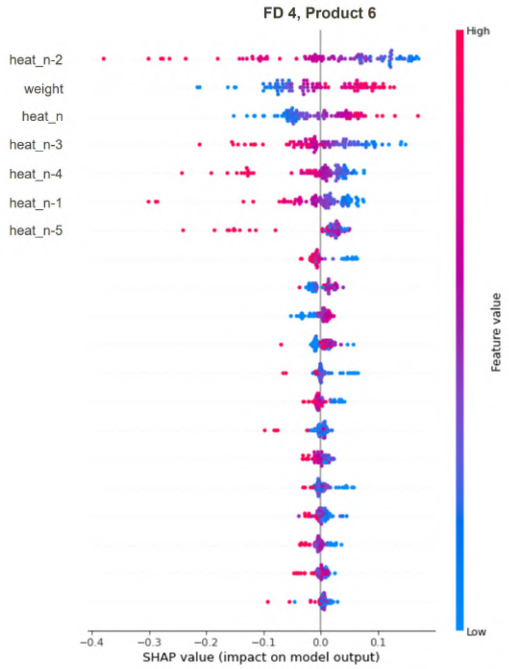

FD 3, Product 7
weight
heat_n
heat_n-2
heat_n-4
heat_n-3
heat_n-1
heat_n-5
High
Low
Feature value
−0.2
−0.1
0.0
0.1
0.2
SHAP value (impact on model output)

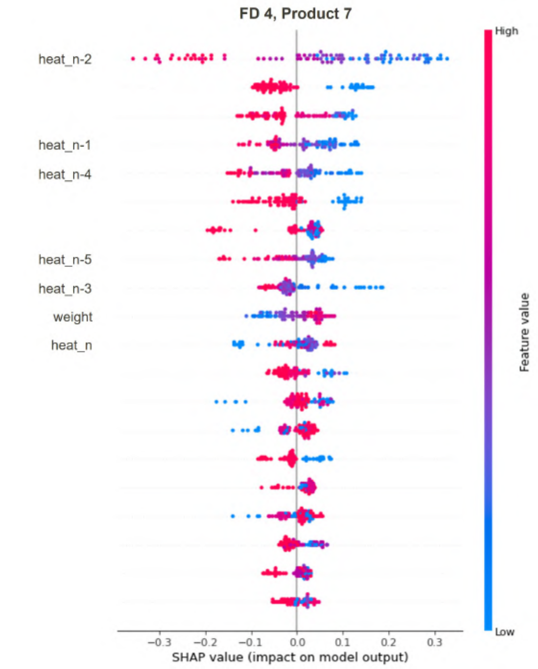

FD 3, Product 8
heat_n-3
heat_n-5
heat_n-2
heat_n-1
weight
heat_n
heat_n-4
High
Low
Feature value
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
SHAP value (impact on model output)


FD 3, Outage
heat_n-3
heat_n-2
heat_n
heat_n-4
weight
heat_n-5
heat_n-1
−0.4 −0.3 −0.2 −0.1 0.0 0.1 0.2 0.3 0.4
SHAP value (impact on model output)
High
Feature value
Low
FD 4, Product 1
heat_n-2
heat_n-1
heat_n
heat_n-3
weight
heat_n-4
heat_n-5
−0.3 −0.2 −0.1 0.0 0.1 0.2
SHAP value (impact on model output)
High
Feature value
Low
FD 4, Product 2
heat_n-2
heat_n-4
heat_n
heat_n-5
weight
heat_n-1
heat_n-3
−0.3 −0.2 −0.1 0.0 0.1 0.2 0.3 0.4
SHAP value (impact on model output)
High
Feature value
Low
FD 4, Product 3
heat_n-2
weight
heat_n-3
heat_n-1
heat_n-4
heat_n-5
heat_n
−0.4 −0.3 −0.2 −0.1 0.0 0.1 0.2
SHAP value (impact on model output)
High
Feature value
Low

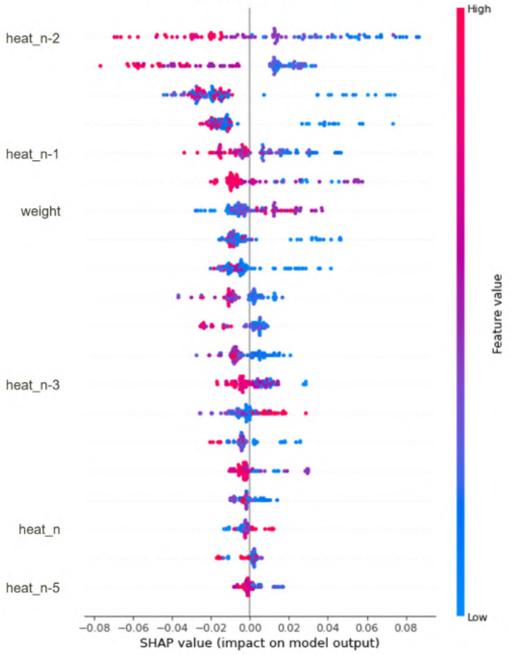

FD 4, Product 4
weight
heat_n-2
heat_n
heat_n-3
heat_n-1
heat_n-5
heat_n-4
High
Low
Feature value
−0.6
−0.4
−0.2
0.0
0.2
SHAP value (impact on model output)


FD 4, Product 5
heat_n-4
heat_n-3
heat_n-2
heat_n-5
heat_n-1
High
Low
Feature value
−0.2
−0.1
0.0
0.1
0.2
SHAP value (impact on model output)

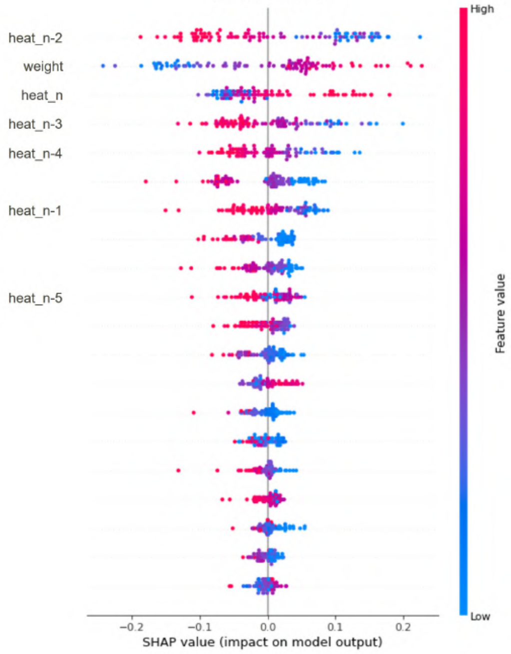

FD 4, Product 6
heat_n-2
weight
heat_n
heat_n-3
heat_n-4
heat_n-1
heat_n-5
High
Low
Feature value
−0.4
−0.3
−0.2
−0.1
0.0
0.1
SHAP value (impact on model output)


FD 4, Product 7
heat_n-2
heat_n-1
heat_n-4
heat_n-5
heat_n-3
weight
heat_n
High
Low
Feature value
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
SHAP value (impact on model output)

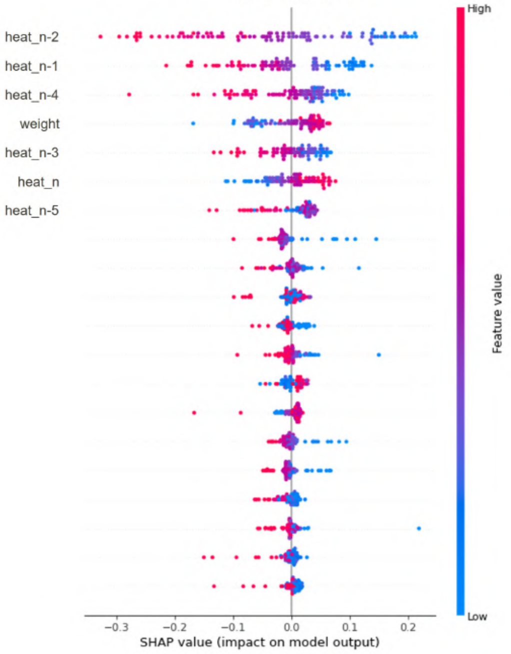
FD 4, Product 8



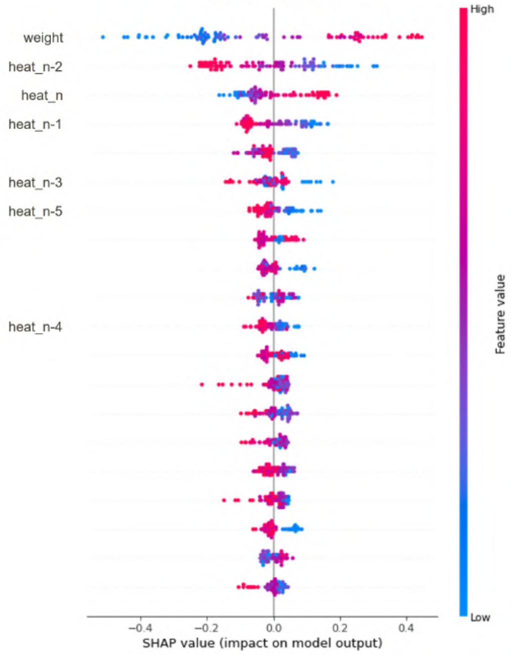
FD 4, Product 9



FD 4, Product 10



FD 4, Product 11



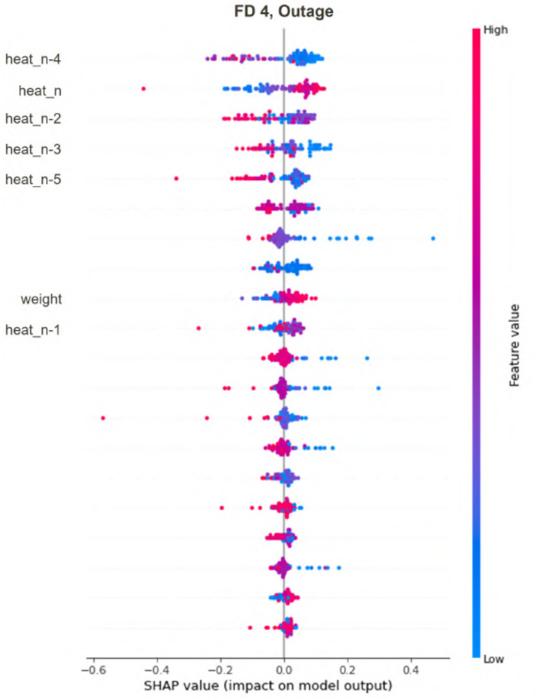


Figure K.1: SHAP Plots of the XGBoost Models

L

Appendix L: Metrics Scores of ANFIS Models in All Products

Table L.1: Metrics Scores of ANFIS Models in All Products

Freeze Dryer	Product ID	R^2 (train)	R^2 (test)	RMSE (train)	RMSE (test)	MAE (train)	MAE (test)
FD 1	Product 1	0.65	0.69	6.99	6.71	5.24	4.94
	Product 2	0.65	0.67	8.39	8.17	6.30	6.21
	Product 3	0.61	0.61	8.94	8.84	6.79	6.83
	Product 4	0.73	0.74	8.10	8.07	5.96	5.94
	Product 5	0.69	0.65	6.90	7.31	5.09	5.29
	Product 6	0.66	0.70	8.19	8.02	6.01	6.30
	Product 7	0.74	0.75	10.22	9.91	7.91	7.63
	Product 8	0.40	0.40	10.43	10.00	8.03	7.74
	Product 9	0.58	0.64	10.33	9.99	7.92	7.85
	Outage	0.51	0.51	15.90	15.40	12.56	12.29
FD 2	Product 1	0.69	0.78	6.95	5.70	5.70	4.04
	Product 2	0.76	0.77	9.17	9.02	6.85	6.66
	Product 3	0.68	0.70	9.69	9.14	7.22	6.77
	Product 4	0.62	0.64	11.55	11.25	9.06	8.88
	Product 5	0.72	0.71	8.12	8.12	6.09	6.18
	Product 6	0.71	0.73	11.14	10.25	8.21	7.68
	Product 7	0.72	0.73	7.92	8.04	5.99	6.13
	Product 8	0.55	0.54	7.71	7.71	6.02	5.99
	Product 9	0.85	0.83	9.16	9.28	6.97	6.62
	Outage	0.46	0.52	17.94	17.35	13.94	13.26
FD 3	Product 1	0.42	0.48	10.80	10.27	8.33	8.03
	Product 2	0.16	0.14	13.01	12.80	9.82	10.17
	Product 3	0.33	0.27	5.19	5.18	4.03	3.99
	Product 4	0.45	0.49	8.74	8.47	6.14	6.07
	Product 5	0.26	0.26	13.44	14.28	10.51	10.94
	Product 6	0.32	0.32	7.83	7.61	5.96	5.88
	Product 7	0.24	0.19	7.71	6.64	5.45	5.04
	Product 8	0.32	0.33	11.93	11.69	9.39	8.84
	Outage	0.19	0.19	18.76	18.50	15.19	14.93
FD 4	Product 1	0.66	0.68	9.93	9.71	7.43	7.43
	Product 2	0.44	0.37	15.57	15.93	12.42	12.77
	Product 3	0.54	0.49	11.80	11.88	9.33	9.57
	Product 4	0.56	0.62	12.59	11.34	9.70	8.68
	Product 5	0.62	0.57	10.99	11.18	8.11	7.74
	Product 6	0.54	0.54	11.21	10.61	8.38	8.01
	Product 7	0.63	0.62	15.69	15.20	12.26	11.39
	Product 8	0.63	0.60	8.48	8.17	6.42	6.62
	Product 9	0.65	0.65	9.19	9.08	7.02	6.81
	Product 10	0.33	0.31	12.67	13.48	9.98	10.74
	Product 11	0.77	0.80	12.96	11.69	10.05	9.08
	Outage	0.46	0.42	17.53	17.80	14.18	14.30

M

Appendix M: Implementation Result of the Solution in Outages

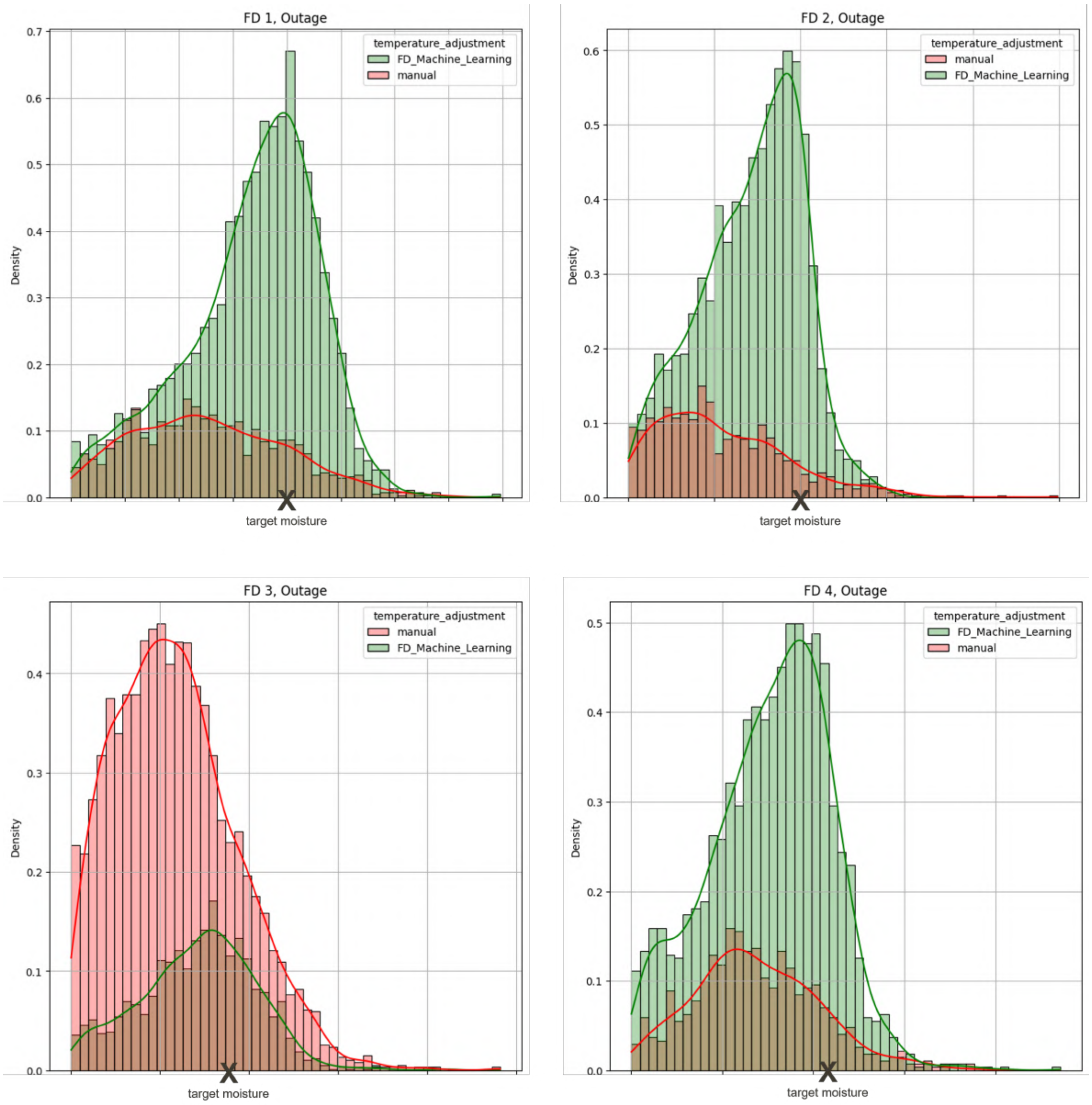


Figure M.1: Implementation Result in Every FD Machine (Outage)