A Work Project, presented as part of the requirements for the Award of a Master's degree in Business Analytics from the Nova School of Business and Economics.

FIELD LAB JERÓNIMO MARTINS – INTELLIGENT MANAGEMENT INFORMATION SYSTEMS FOR NOVA & GO STORE

AUTOMATING FOOD QUALITY AND ENERGY EFFICIENCY MONITORING IN RETAIL 4.0

MARIANA ALEXANDRA FREITAS FILIPE

Work project carried out under the supervision of:

Professor Qiwei Han

Mr. Rui Tomás

25/01/2023

Acknowledgments: To Professor Qiwei Han, for sharing his experience and for providing such valuable guidance. To the Jerónimo Martins group and, in particular, Mr. Rui Tomás, for the opportunity to work on a real-world problem. To my colleagues, Maria and Wander, for all the video calls and the countless all-nighters. And to my family and friends, for their patience and support throughout this journey.

Abstract: Taking advantage of the latest technologies, the retail industry has drastically evolved in recent years. Launched by Jerónimo Martins, Nova & Go store is an example of a revolutionary supermarket. With the purpose of further improving the Lab store's daily operations, IoT data collected by smart ovens was analyzed to automate the monitoring of the food batches produced, both in terms of food quality and energy efficiency. Due to the lack of labeled data available, a semi-supervised approach was tested and applied. Later, the multi-class classification's results were disclosed using the Microsoft package, *InterpretML*.

Keywords: Automation, Energy Efficiency, Food Quality, Internet of Things, Retail 4.0

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

INDEX

1. Introduction

In today's world, data is viewed, by most organizations, as a strategic asset *(Delloite, 2021).* In the words of Clive Humby, a British mathematician, it can even be described as "the new oil of the 21st century", due to its tremendous potential and value when realized through business applications. However, just like oil, data needs to be refined. Raw data is worthless, if valuable insights cannot be extracted from it *(McKinsey, 2018).* Rather, it's through the acquisition of information that companies are able to respond proactively and intentionally to market conditions. Accordingly, in 2011, Peter Sondergaard, a Gartner's board member, corrected Humby's famous quote by stating: "Information is the oil […] and analytics is the combustion engine. Today, analytics is the new oil."

Among a wide range of industries, Big Data has opened up new opportunities, and Retail is no exception. In fact, the concept of Retail 4.0 was introduced in 2010, with the intention of describing the sector transformation based on Industry 4.0 technologies, such as Internet of Things, Cloud Computing or Artificial Intelligence *(Har et al, 2022).* Technology has enabled the retail sector to create a brand-new shopping experience *(Gazzola et al., 2022).* From self-service kiosks to e-commerce platforms or AR-based tools aimed to enhance customer experience, this new era is characterized by innovation.

An example of an innovative project aligned with this new trend is a cashier-less store created by Jerónimo Martins (JM), the retail leader in Portugal. In October 2019, JM launched Nova & Go store, at the Nova School of Business and Economics campus. Taking advantage of a group of customers predominantly made up of young students, unlike what happens in the regular stores, this project aims to test new technologies and new products. Being primarily visited by Gen Z customers, it is an ideal opportunity to experiment with innovative technologies and new business models. Apart from traditional groceries, the store offers various freshly prepared foods and other novel products, all offered with the students in mind.

Customers' experience in this Lab store starts with the download of the Pingo Doce & Go Nova app, available for both iOS and Android. After registering, a QR code should be displayed to enter the store. Once inside, purchases can be recorded in two ways: either through NFC technology or by pointing the cell phone's camera at the product code. Later, to be able to pay, customers can associate their bank card to the app or, alternatively, use the payment towers. Yet, even in the latter case, only credit or debit cards are accepted, not notes or coins.

During this process, data is generated at various points. So, in order to handle the large volume of data produced by the store, Management Information Systems (MIS) must be incorporated into its daily activities. Essentially, MIS collects and processes data from all sources within an organization to assist managers in their decision-making processes *(Berisha-Shaqiri, 2014)*. As a result, JM has access to a detailed database that helps it explore the habits of its customers and analyze critical aspects of its processes.

Based on the exceptional opportunities provided by the Lab store, JM has identified one of the supermarket's biggest challenges to be developed throughout this thesis. Essentially, the data collected from the store's semi-automated ovens will be used to classify each batch both in terms of food quality and energy efficiency. Such an improvement can both increase customer trust and decrease costs.

Lastly, it should be noted that Design Science Research (DSR) methodology was chosen to structure the present thesis, since it has proven to offer outstanding benefits when it comes to solving real-life problems within organizations *(Hevner et al., 2004).* In the appendix, it is possible to find the different steps of this approach (*Fig. 1 in the appendix*).

1. Problem Statement and Motivation

At Nova & Go store, hundreds of ready-made meals are sold every day. Currently, the store is equipped with three ovens, which secure the preparation of a wide range of food products, including bakery items, chicken, burgers, and other takeaway choices. The truth is that the customer base of this revolutionary Lab store cannot be compared to any other regular store within the group. Apart from the lower average age and the propensity to tech savviness, the typical client's mission is not limited to shopping for groceries. Rather, the store is primarily used by students for immediate consumption who, due to their fast-paced routine, are looking for fast and effective alternatives for their class breaks. The store's singularity justifies, then, a greater investment in takeaway solutions *(Salgueiro, 2019).*

However, serving ready-to-go meals entails an additional level of responsibility. For instance, if chicken is not properly cooked, it may contain bacteria such as salmonella, resulting in health problems for the customer *(Centers for Disease Control and Prevention, 2022)*. Due to this, all poultry products must be cooked at 74 degrees Celsius or higher (*Assistant Secretary for Public Affairs (ASPA), 2022)*. Even in cases where undercooking does not have such severe negative consequences, food quality is, nevertheless, an inherent concern to any food provider. Positive client perceptions of the supermarket's food options play an important role in building trust and loyalty, improving the chances of a repeat customer, and promoting word-of-mouth marketing. On the other hand, an adverse incident can damage the brand's reputation, exacerbated by the fact that people often recall the bad experiences more vividly and for longer *(J. Knutson, 1988).* Furthermore, to offer takeaway services, Jerónimo Martins (JM) incurs additional electricity costs, either for the cooking process itself or for later heating food recipes that have been prepared in advance. Yet, considering the global energetical crisis that the modern world faces, following Russia's invasion of Ukraine last February, energy expenditures have increasingly become a matter of concern worldwide. *(International Energy Agency, 2022).* Unquestionably,

with the rapid ascension of prices, energy efficiency is now more important than ever, and companies are deeply worried about the future and the crisis' repercussions *(Euronews, 2022),* which are expected to be worse than in the 1970s *(Weisenthal & Alloway, 2022).*

Therefore, to ensure that clients are always provided with safe and pleasant meals while at the same time minimizing energy losses, both food quality and energy consumption must be monitored throughout the preparation of the food batches. As of now, staff members must examine the oven's screen to assess each batch and make their own judgment based on their experience, which requires time and is subjective.

2. Objectives of a Solution

However, if Nova & Go store's unique characteristics require a new approach and pose new challenges for the store's management, the amount of data collected also allows new solutions to be developed. In fact, the semi-automated ovens are connected to a database, which receives updates every 60 seconds or whenever something changes in the process - for instance, when the door is opened, or the cooking mode is revised. This is a good example of an Internet of Things (IoT) application.

Several advantages would arise if the store's staff was able to automatically monitor the preparation of each batch. First, as aforementioned, if successfully implemented, damaged meals and energy inefficiencies would be identified and addressed promptly. As a result, food quality would always be assured, which is crucial for the brand's reputation, and costs would be reduced. Moreover, automation would allow employees to concentrate on other tasks, which may increase their performance, since they would have one less thing to worry about.

Finally, by collecting and analyzing all this information, it would be possible to determine what is lacking in each batch process. Or, to put it another way, what are the external factors that lead a given batch to be of poor quality or inefficient in energetical terms?

By identifying what is failing, the store manager can act accordingly and correct it. This promotes better waste management, which will save time and resources. For instance, by reducing the frequency of faulty batches, which cannot be sold, the amount of food wasted would decrease. Furthermore, implementing better practices could reduce the amount of energy spent unnecessarily.

In this sense, by combining all the accumulated data, an algorithm could be developed to automate batch vigilance and classification. In fact, the problem identified can be summarized in one question:

> ***Research Question:*** *How can Machine Learning be used to monitor and detect faults in food preparation processes and spot energy inefficiencies?*

3. Literature Review

To gain a deeper understanding of the problem at hand, this section provides a brief overview of the existing research and solutions in this field.

3.1. Internet of Things (IoT)

In 1999, Kevin Ashton introduced the concept of Internet of Things (IoT), by proposing radio-frequency identification (RFID) as a means of tracking products throughout the supply chain *(Wang et al. 2015).* Despite the absence of a well-established definition, the term can be described as a collection of devices that, when connected to the internet, are able to communicate and exchange data with each other. Over the past few years, IoT devices have grown rapidly. *Kaspersky* reported that, in 2019, 61% of global companies had already incorporated IoT into their businesses. Moreover, the trend is expected to continue, as Gartner predicts there will be 25 billion IoT connections worldwide by 2025. In the future, more and more objects are expected to extend their connectivity spectrum anywhere in the world, as wireless objects become more common (*Kellmereit & Obodovski, 2013; Brynjolfsson & McAfee, 2018*).

By connecting data, people, processes, and objects virtually, new opportunities arise. In fact, different industries can benefit from IoT applications, supporting operations and optimizing processes. A special focus will be placed on Retail and how it can benefit from Industry 4.0. Data can be used to improve inventory management or personalize and enhance customers' experience, for example. Sensors are able to track customers' journeys inside the store, aiming to better the positioning of the merchandise *(Gregory, 2015).* Similarly, customers can locate products using a mobile application that works with an Indoor Positioning System *(Hicks et al., 2013).* RFID technologies can boost inventory tracking *(Gregory, 2015).* And sensors and Big Data can even help ensure shelf availability *(Vargheese et al., 2014).* As far as possibilities go, there is a vast array to choose from.

3.2. Applications in the research field

Specifically, in the areas of energy efficiency and food quality monitoring, research has also been conducted in the past. Sensors incorporated into food packaging have proved to enable the monitoring of the product's quality and safety, measuring the number of pathogen agents, gases, temperature, humidity, and storage period *(Popa et al., 2019).* Alternatively, a sensor system using optical fibre was used to control the cooking process in a large-scale industrial oven, through pattern recognition and artificial neural networks *(O'Farrell et al., 2005).* Particularly, this approach has been applied to poultry products. In addition, a multi-class Support Vector Machine (SVM) algorithm was used to classify hyperspectral images of premium honey products according to their quality *(Phillips & Abdulla, 2021).*

As for power consumption, an example is ERINOKS. Using induction technology, this IoT application measures the weight of the food and instantly adjusts the power accordingly, maximizing energy efficiency *(Tanrıseven et al, 2019).* Moreover, monitoring real-time energy data can provide insights into abnormal consumption patterns and quantify energy efficiency gaps *(Shee Tan et al., 2017).*

In fact, anomalies have been studied for centuries. In *Hawkins' (2014)* terminology, an anomaly (outlier) is an observation that deviates significantly from other observations, leading to suspicions that it was generated by a different mechanism. According to another definition, anomalies are patterns in data that do not fit the expected or so-called normal behavior *(Chandola et al., 2009).* Such concept may be appropriate for the issue presented, since energy inefficiencies and poor-quality meals are associated with an atypical behavior captured by the variables considered.

Generally, anomaly detection is a one-class classification task *(Bergman & Hoshen, 2020).* However, since both food quality and energy efficiency must be evaluated simultaneously, a supervised multi-class classification algorithm should be displayed instead. As part of this type of problem, each input example is assigned a single class label. This can be accomplished through a variety of models, such as neural networks, support vector machine or decision trees, for example *(Aly, 2005).* Ultimately, any Machine Learning algorithm should be chosen based on the type of output desired, the number of features, and the type and size of data.

4. Design and Development

4.1. Data Understanding

Data was collected on 1,653 batches prepared on the ovens of the Nova & Go store, between February 1, 2022, and July 28, 2022. This is the most recent data gathered by the company, and it corresponds roughly to the second semester of the previous academic year (2021/22).

For the purpose of this thesis, only the ten most prepared food recipes were selected and analyzed *(Table 1 in the appendix).* Thus, the initial dataset comprised a total of 49,914 rows and 25 columns, split into 5 numerical and 20 categorical features. Each row corresponded to a new update sent by one of the IoT devices.

4.2. Data Curation & Feature Engineering

Following an initial inspection of the data, some adjustments were made to ensure an accurate and easily understandable dataset. The first step was to remove batches with a total duration of 0 seconds, as these were erroneous units that were immediately canceled. Similarly, the target temperature for one of the recipes, "*frango"*, was set at 88 degrees for all batches, except for some rows. These cases were corrected, as well. Furthermore, to resolve the issue of a few batches being incorrectly divided and assigned distinct IDs, the temperature, door status and duration at the beginning of each batch were examined.

All variables recorded in seconds were converted to minutes, to facilitate the understanding of the features and the comparison between different food recipes.

Additionally, departing from each batch's start date, 7 new columns were added, aiming to provide more details about the timeline and the utilization of each oven. The following variables were created: *Date, Year, Month, Day, Weekday, Time,* and *Hour.*

Moreover, the number of times the chamber's door was opened during the preparation of each batch was not a sufficient indicator. Even if the door was opened only once on a given batch, it was important to consider the length of time that it remained open. This would certainly have an impact on the baking process and was calculated as another feature.

Similarly, to measure energy efficiency, the total amount of energy spent was a critical factor. Nevertheless, since this is highly dependent on the duration of each batch, the energy spent per minute was added too.

Average and extreme temperatures may also be of interest. Yet, a word of caution is necessary. It is possible, as explained above, for the updates to be sent in an irregular manner, regardless of the predefined 60-second interval. Naturally, this results in unevenly spaced time series. So, before computing statistical measures such as the average, all timestamps must be equally distant from each other. To achieve this, a resampling function was applied to the *Start Offset* column,

aiming to increase the time series frequency, through *upsampling*. Then, a linear interpolation method was used, with the goal of filling all the empty values in the newly added rows. Simply, by connecting the known data points with a straight line, the unknown entries can be estimated. As the temperature variation is not significant in a single second, such an approach seems appropriate in this situation.

Yet, average temperatures can be affected greatly by large deviations, thus making it necessary to take into account the time during the cooking process when the temperature was above a certain threshold.

As a final step, some non-relevant and redundant columns were removed. In the end, the cleaned dataset included a total of 1,520 batches, disclosed in 49,877 rows and 31 columns. A detailed description of all variables can be found in the appendix *(Table 2 in the appendix).*

4.3. Exploratory Data Analysis

Next, using data visualization as a starting point, some conclusions were drawn. Firstly, the month in which more data was collected was May *(Fig. 2 in the appendix).* It is possible, however, that this does not reflect a lower number of batches prepared in other months, but rather a compromised connection between the device and the database. For instance, there is no data available for March. Furthermore, looking at the intra-week variation, on average, the beginning and end of the week produce more batches than the other weekdays *(Fig. 3 in the appendix).* Similarly, the busiest hour of the day, measured by the average number of batches prepared per hour, corresponds to 7 a.m. and, in general, the morning period *(Fig. 4 in the appendix).*

With respect to how each one of the three ovens is utilized, "Restaurante Grande" is used with a lesser variety of meal options *(Fig. 5 in the appendix).* Plus, most of the food recipes are exclusive to a given chamber, with a few exceptions. "*pao de quejo PD"*, *"croquetes PD"* and *"frango"* can be prepared in more than one oven. Moreover, "Padaria" holds the leading position as producer, supplying 72% of all batches *(Fig. 6 in the appendix).* The reason for this can be

ascribed to the fact that the food recipes prepared there have a shorter duration, on average *(Fig. 7 in the appendix)* and are typically sold throughout the day rather than more frequently at specific times, such as lunch time *(Fig. 8 in the appendix).* In contrast, “Restaurante Grande” is usually used until a certain hour and for longer batches.

Indeed, focusing on the duration, the longest food recipes are “*frango*” and “*burguer congelado*”, which take approximately 62 and 42 minutes, respectively *(Fig. 9 in the appendix).* The two products represent 95% of the production of "Restaurante Grande", the device with the lowest production volume *(Fig. 5 in the appendix).*

Considering energy consumption, the most expensive food recipe is by far "*frango*", perhaps due to the target temperature that must be reached and the longer average duration *(Fig. 10 in the appendix).* Thus, since this variable has a significant correlation with time, it would be more appropriate to analyze the energy spent per minute instead. Although *“frango”* remains in second place, surprisingly, “*pao*” consumes the most energy per minute *(Fig. 11 in the appendix).* This may already indicate some inefficiencies and poor practices in the preparation of this recipe. In contrast, *“pao de queijo PD”* is the less energy intensive.

Further, to determine if some ovens are inherently more energetically efficient than others, the three recipes prepared in more than one device were compared *(Fig. 12 in the appendix).* In fact, it appears that, for the same service, “Restaurant Grande” consumes more energy per minute than “Restaurante Pequeno”. However, a sample of two recipes is not sufficient to make a solid judgment, and a further analysis with more meals should be conducted.

As part of measuring energy efficiency, one important factor to consider was the opening of the door during the preparation of a batch. As a matter of fact, 17.6% of all exceptional updates sent by the ovens were a result of opening the door, which represents almost a fifth *(Fig. 13 in the appendix).* This reveals the severity of the problem. Most of the batches have had the door opened at least once during the cooking process *(Fig. 14 in the appendix)*. Specifically, *“frango”*

has the highest average number, with 2.37 times per batch *(Fig. 15 in the appendix)*. Yet, more than the number of times the door was opened, it is important to consider the total duration for which it remained open. Again, *"frango"* holds the lead, registering an average of 26.6 minutes per batch *(Fig. 16 in the appendix).* As to why this occurs, one possibility could be that the staff members often try to understand the baking status to avoid undercooking. For this purpose, the time at which the door was opened, for *"frango"*, was considered. To eliminate the duration bias in the timing of the door opening, only batches with a duration of approximately 60 minutes were selected. Unexpectedly, most of the cases occur in the first 10 minutes of the batch's preparation, which suggests that it's the pre-heating that motivates more investigation *(Fig. 17 in the appendix).*

Upon analyzing the data regarding the temperature, one may wonder if the chamber temperature differs from the temperature of the food, which is measured using a thermometer. This may be due to energy dispersion. *"frango"* is the food recipe with the largest variation *(Fig. 18 in the appendix).* Likewise, it is also the recipe that achieves a broader range of temperatures, throughout the cooking process *(Fig. 19 in the appendix).*

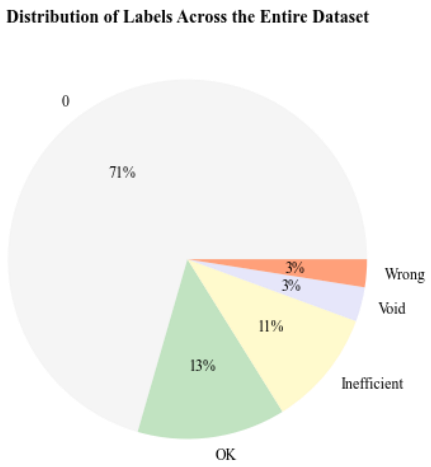
Finally, a supervised model requires labels to be trained. To address the problem at hand, four distinct labels were assigned by JM: OK, Inefficient, Void, and Wrong. A description can be found below *(Table 3).*

Table 3 – Description of the Labels, provided by Jerónimo Martins

Label	Class	Description
Inefficient	0	The batch was properly prepared but inefficient
OK	1	The batch was properly prepared and efficient
Void	2	The batch was very short
Wrong	3	The batch was not properly prepared

However, only 29.47% of the observations were labeled, with OK and Inefficient being the most frequently observed classes *(Fig. 20)*.

Fig. 20 – Distribution of Labels Across the Entire Dataset, in percentage (%)



"misto pastelaria novo", "pao", and "frango" were the food recipes that had the greatest number of labeled batches *(Fig. 21 in the appendix)*. This makes sense since these recipes were also among the top 5 most-produced meals *(Fig. 22 in the appendix)*. Regarding the factors that influence the classification, it is evident that Inefficient and Wrong Batches tend to leave the door open for a longer period of time, on average *(Fig. 23 in the appendix)*. Similarly, OK batches usually have a higher average temperature *(Fig. 24 in the appendix),* and Void units register a shorter duration *(Fig. 25 in the appendix)*.

4.4. Model

a) Feature Selection

When training a model, large datasets do not necessarily result in better results. In fact, if the number of features is significantly greater than the optimal, accuracy may even be reduced *(Kohavi & H.John, 1998)*. For that reason, only the most relevant variables were kept for the modeling phase. To account for the differences between recipes' requirements regarding duration and temperature, the *CC Name* feature was considered. Furthermore, *Duration Minutes* is a determinant indicator for Void batches. Equivalently, it might be related to unnecessarily long

batches, leading to wasted energy, or OK meals, which need to be cooked over a given period of time, otherwise, they cannot be properly prepared.

It was also important to measure both the number of times that the oven's door is open, *Open Door Count*, and the amount of time that it remains open, *Open Door Minutes,* in order to assess inefficiency. Complementarily, *Energy Consumption* measures the total amount of energy consumed during the preparation of each batch. In place of *Energy Spent per Minute*, this variable was used, since the former, while providing us with important insights regarding the previous section, was highly dependent upon the duration. Yet, *Duration* is key to the success of the label Void, and thus cannot be removed.

Lastly, it was necessary to evaluate the temperature registered during the batch's preparation in accordance with the recipe requirements. In light of this, the average temperature, *Avg Temperature*, and the maximum temperature reached, *Max Temperature*, should also be input into the model. Yet, although the average was a useful statistical measure, it is highly affected by variations. For instance, a batch cannot be properly cooked if it is prepared at a low temperature for the majority of the time, and then reaches an extremely high temperature, causing the average to be distorted. For that reason, the total amount of time that the temperature was above a certain threshold, *Minutes Above Threshold*, was also considered. The threshold was determined by the average temperature of batches labeled as OK for a given recipe, and it varies for different recipes. As a result, not only was the average temperature taken into consideration, but also for how long the temperature was above a certain target.

b) A Semi-Supervised Approach for Handling Unlabeled Data

During the training process of supervised Machine Learning algorithms, a large amount of accurately labeled data is required, so that the model is able to generalize its results when presented with unseen data. However, data labeling is often a time-consuming and expensive task *(IBM Cloud Education, 2021).* In fact, in the dataset received, only 29.47% of the batches

were labeled *(Fig. 20 in the appendix).* Yet, even if dropping the unlabeled observations seems to be an effective solution and simplifies computation, it may lead to an overfitting problem. A semi-supervised approach is recommended instead, in these cases.

Indeed, it is possible to expand the number of labeled points available for training. The combination of labeled and unlabeled data allows three different algorithms to be applied and compared: Label Propagation, Label Spreading, and Self-Training Classifier.

Label Propagation (LP) assumes that closer data points have similar class labels. It creates a linked graph of the data, in which nearby points have a greater connection and weight, thus increasing the likelihood of a particular label propagating. Using the probabilities determined by the process above, new labels are assigned to unlabeled points. Labels keep being updated, for multiple iterations, until convergence is reached *(D'Sa et al., 2020).*

Similarly, Label Spreading (LS) is also a graph-based algorithm based on distance and similarity matrices (sometimes called affinity matrices) *(Chen & Wang, 2017).* In Scikit-learn documentation, the differences between the two approaches are explained by adjustments to the similarity matrix that is used to graph the label distributions, as well as clamping effects. For instance, while LP uses the raw similarity matrix, LS uses the symmetric normalized graph Laplacian matrix in its calculations. Simply, it has a regularization function, thereby making it more robust to noise.

Another difference is that LS employs a soft clamping, controlled by a hyperparameter α, whereas LP employs a hard clamping ($\alpha = 0$). Clamping describes the degree to which the labeled data can be influenced by its neighbors, resulting in the reclassification of the original points. Consequently, a value near 1 allows most of the original labeled data to be altered.

Alternatively, the Self-Training Classifier (STC) algorithm starts by using the labeled data to train a classifier, which is then used to make predictions on unlabeled data *(Didaci & Roli, 2006).*

Later, the pseudo-labels with the highest confidence are added to the training set and the process is repeated through multiple iterations.

To determine which algorithm best fitted the dataset studied, the accuracy scores of the three algorithms were compared, since pseudo-labels should be as reliable as possible to avoid harming the model's performance in later steps. An algorithm's accuracy can be understood as the probability that the prediction is correct *(Grandini, Bagli & Visani, 2020).*

Moreover, to maximize the algorithm's performance, the right combination of hyperparameters was determined before the comparison. In addition, due to the wide range of values, features were scaled. To ensure a fair assessment, only the labeled data points were selected for the evaluation. After that, 30% of the labels were retained, while the remaining ones were masked. The threshold of 30% was chosen to make the experiment as similar as possible to the dataset.

Looking at the results, LS had the best performance, achieving an accuracy of 76.1%. This means that the algorithm was capable of accurately predicting approximately 76% of the masked data. In contrast, LP and STC registered a score of 62.1% and 57.1%, respectively.

It should also be noted that the best estimator included a clamping factor of 0.1, which means that 10% of the original label distribution was replaced. This may indicate the presence of some noise in the input labels. In spite of this, the fact that alpha was still smaller than the default hyperparameter for Label Spreading, which is set as 0.2, may hint that there is no cause for concern. As a matter of fact, the accuracy score is decreased when alpha is increased, *ceteris paribus*.

Furthermore, an analysis of the confusion matrix can provide us with additional insights *(Fig. 26 in the appendix).* This metric can be defined as a representation of the number of occurrences between the true/actual label, and the predicted classification *(Grandini et al., 2020).* It was evident that, compared to the other labels, the algorithm failed to spot the poor food-quality batches, as 21 incorrectly prepared batches were not identified. In contrast, out of 13 labels

predicted as Wrong, only 6 of them were accurate. This can be problematic, as mistakenly classified batches may be wasted, while undetected wrong units may be sold to customers. That's why inputting more labels, especially from this class label, is extremely important as a future step, to ensure the model's reliability.

c) Training the Model

In order to determine which algorithm was best suited to the dataset presented, a number of different models were tested. A listing of all experiments, along with the accuracy score obtained, can be found in the appendix *(Table 4 in the appendix).* It should be noted that, during this initial analysis, all models were trained using the default hyperparameters. As a result, XGBoost proved to be the most suitable classifier.

In numerous data mining competitions, Extreme Gradient Boosting (XGBoost) has shown to provide exceptional results on a wide range of classification problems *(Chen & Guestrin, 2016)*. Being an optimized distributed gradient boosting library, as described in the XGBoost documentation, it combines a set of weak learners together to create a single strong classification model. Also, it is trained in an additive manner, using the learned mistakes to improve the output's performance, in each iteration *(Chen & Guestrin, 2016).* That is, each tree added considers the previous prediction value for the given input data and then maximizes the prediction gain *(Le, Oktian & Kim, 2022).*

Moreover, one of the most important advantages introduced by XGBoost is its scalability, making its running time more than ten times faster than other models through parallel learning *(Chen & Guestrin, 2016; Zou et. al, 2022).* In addition, the algorithm is also capable of handling missing data and fine-tuning an ample range of hyperparameters *(Le, Oktian & Kim, 2022).*

Based on the advantages offered by this algorithm and the preliminary results obtained, this thesis will focus on XGBoost. In fact, some studies have shown that it is superior to other algorithms in handling tabular data *(Chakraborty & Elzarka, 2019).*

Before setting up the model, however, to solve the unlabeled data issue, Label Spreading was now applied to the entire dataset, using the best hyperparameters determined in the previous step. The experiment returned an accuracy of approximately 76% with only 30% of the data labeled, as in this case, which appears to be a reasonable estimate.

As a result, a new distribution of class labels was created (*Table 5 in the appendix*). Now, the most frequent class corresponded to OK, with 881 batches classified, followed by Inefficient, and lastly, Wrong and Void, with only 79 and 74 observations, respectively. However, this reflects a moderately imbalanced dataset, according to Google Developers, where the least populated class has a 1:8 ratio for the most recurrent label. Therefore, the classifier would tend to assign all data to the majority class, which is usually the less important one *(Kotsiantis et al., 2006).* In fact, due to the lack of food quality, Wrong labeled batches cannot be sold. This is the class with the highest risk of misclassification, and one of the least frequent.

Consequently, to rebalance the class distribution, a random sampling approach was enforced on the training dataset. After applying a Random Oversampling technique, rows within minority classes were randomly duplicated until all labels had the same number of data points. In contrast, a Random Undersampling approach would eliminate rows from the majority classes at random, until all classes had 74 observations. Since there was only a limited amount of data for the purpose of the current thesis, the first approach was preferred.

Following the abovementioned adjustments, a fine-tuning process was conducted. For instance, the maximum depth of each tree, *max_depth*, was used to control the degree of overfitting to the model. A higher value allows the model to learn the specificities of a particular sample.

The optimal combination of hyperparameters was, then, determined through Randomized Search, available in the scikit-learn library. This technique involves testing a random mix of hyperparameters, for a predetermined number of iterations, until the best estimator is found

(Akyol, 2022). Indeed, when limited computational resources are available, it has been proven that searching over a wider range has resulted in better *results (Bergstra & Bengio, 2012).*

d) Results and Discussion

i) Model Evaluation

Once the hyperparameters have been tuned to ensure optimal performance, the model must be evaluated using a variety of multi-class metrics. In fact, the use of disparate methods allows for different perspectives and, consequently, a deeper level of judgment to be reached.

The confusion matrix is, once more, used as a starting point, since it summarizes all the classifier's predictions and behavior *(Grandini, Bagli, & Visani, 2020).* It can be seen that, although 29 batches were not properly identified, the model correctly classifies 275 others *(Fig. 27 in the appendix).* This is a clear indication of a positive result.

In fact, to provide an overall evaluation of multi-class classification, accuracy is among the most widely used metrics. As stated before, it calculates the percentage of correct predictions made by the model. Based on the data provided, a score of 90.5% was achieved.

However, as the dataset is unbalanced towards labels OK and Inefficient, this score heavily depends on the model's performance on these classes, as a result of its weight *(Luque et. al, 2019).* Alternatively, macro-average precision and recall can be calculated, by computing each metric independently for every class and then averaging the results. Precision can be defined as the proportion of correct predictions, divided by the total number of predictions for a given class *(Grandini, Bagli, & Visani, 2020).* Meanwhile, Recall measures the proportion of data points from a given class that were detected. Based on the results, macro-averaged precision was greater than recall, although by a relatively small margin, with scores of approximately 85% and 83%, respectively. The fact that both values are lower than their weighted-averaged scores, confirms the dataset's imbalance and the distortion of overall metrics.

This is also visible on the classification report, where each metric's score varies on the class considered *(Fig. 28 in the appendix).* By looking at the F1-score, which combines both precision and recall, the Wrong class performs relatively worse, reaching only 62%, compared to approximately 90% for all other labels. Similarly, even if macro-averaged precision was equal to 85%, it is evident that there was a discrepancy across classes.

Overall, labels OK, Inefficient, and Void are able to accomplish fairly good results. In fact, the precision score of 0.93 indicates that almost all batches classified as Void and OK were correct predictions. Additionally, a large percentage of inefficient batches were detected, as evidenced by an 88% recall rate. Yet, it appears that the model did not fully understand the characteristics of the Wrong class, which is the most poorly predicted label. As a matter of fact, when evaluating the label-spreading algorithm, this behavior was already evident. This can be justified by the limited amount of labeled data available, which is likely to have an impact on the model in two ways. The first concern is that it may compromise the propagation of labels, as Wrong was the less frequent label, making it difficult to absorb the class's essence. Plus, for the modeling phase, the fact that this label has a very small number of rows, as a result of an imbalanced dataset, creates the need for oversampling, potentially resulting in overfitting, since data points from minority classes are repeated *(Kotsiantis, Kanellopoulos & Pintelas, 2006).*

Thus, to dive into the performance differences between classes, the Area Under the Curve (AUC) can be estimated, assessing how well the classifier can distinguish between labels *(Döring, 2018).* Simply, it computes the likelihood that a random data point from one class has a lower probability of belonging to another class than a random observation from the opposite class *(Hand & Till, 2001).* In general, a value close to 1 indicates a high level of separability. In this multi-class classification setting, a One-vs-One (OvO) approach was employed. The results show that, registering the lowest scores, Wrong batches are difficult to separate from Void and Inefficient units. In contrast, Void batches are undoubtedly distinct from Inefficient

and OK meals, achieving a value of 1. Despite this, all values are 0.75 or higher, resulting in an average of 0.94. This reflects how accurate the classifier was in predicting each class. Later, further insights can be gained by examining the most important features that may affect the model's prediction.

ii) Model Interpretability

It is important for people who will be affected by a Machine Learning algorithm to trust it, no matter how well the model performs in its evaluation. In fact, trust is a key factor in the success of any deployment and plays a crucial role in determining whether or not people will rely on automated systems *(Lee, 2021).* Furthermore, in critical areas, such as medicine, the criminal justice system, or the financial markets, it is even more decisive to understand how models make decisions or exhibit a given behavior *(Lipton, 2018).* In a similar manner, controlling food quality and determining whether a batch is fit for human consumption carries some responsibility, since misclassification can result in health-related issues.

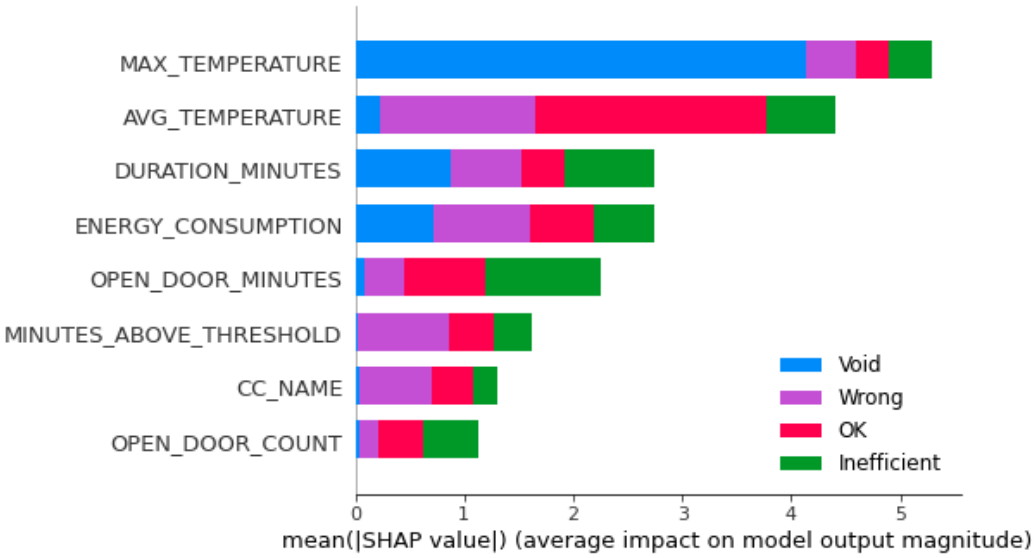
Toward this end, Microsoft has developed a free open-source package, *InterpretML*, which aims to help business decision-makers gain a solid understanding of models' overall behavior, through a collection of techniques *(Microsoft, 2020).* In addition, this toolkit can also be used for debugging and examining individual predictions. The goal is to ensure Intelligibility, one of Microsoft's six principles of artificial intelligence (AI), which states that AI systems should be understandable, monitorable, and responsive to human input *(Microsoft, 2020).* This has proven to be beneficial, in terms of improving the model's performance by identifying errors and fixing them, increasing trust, and therefore making the adoption process easier for end-users, as well as uncovering potential unfair practices.

In this sense, although XGBoost is defined as a glass-box model, meaning that it is inherently interpretable, it is still advantageous to employ explainability tools to have both global and local explanations. Microsoft offers two methods for this purpose: SHAP and LIME.

The former, Shapley Additive Explanations (SHAP), uses Shapley values to explain a model's output. Based on a game theory approach, it measures how each feature contributes to the model's prediction *(Smith & Alvarez, 2021).* Essentially, SHAP values calculate the mean marginal contribution of having a certain feature value, across all possible values in the feature space. It can be used either for analyzing individual predictions or, rather, for global explainability.

Taking an overall look at feature importance, it is evident that *Max Temperature* and *Avg Temperature*, as well as *Duration Minutes*, are among the most important features for the model's classifications *(Fig. 29)*.

Fig. 29 – SHAP Summary Plot



Contrary to this, the number of times the oven's door was opened appears to have little impact. In fact, due to the limited number of unique values, it may be difficult to differentiate between classes based on this. The type of food recipe seems to be of little relevance, too.

Specifically, the Void label is highly affected by *Max Temperature* and *Duration Minutes*. This makes sense, since interrupted batches are characterized by a shortened duration, which results

in a lower temperature being achieved. The lower the duration and maximum temperature achieved, the greater the effect on prediction *(Fig. 30 in the appendix).* Rather, the Inefficient class is most impacted by *Open Door Minutes*, where larger values push the prediction toward this label *(Fig. 31 in the appendix).* This class is also associated with longer durations and lower average temperatures, due to the dispersion of energy caused by the opening of the door. Correspondingly, OK batches are defined by high values of *Avg Temperature* and *Minutes Above Threshold*, along with modest values of *Open Door Minutes (Fig. 32 in the appendix).* This indicates that the cooking process was conducted in an appropriate manner. Finally, batches classified as Wrong depend mostly on temperature-related features *(Fig. 33 in the appendix).* Particularly, contributing to this category are small average temperatures and the length of time it was above a threshold, which suggests that these meals were often undercooked, leading to a poor-quality batch. This is the opposite of an OK unit.

Diving into the confusion between the three classes, it may be convenient to evaluate individual classifications as part of the interpretability analysis. Even if it does not provide a universal rule, examining concrete examples of incorrect predictions can help identify the root cause of the problem.

Keeping this in mind, another framework for achieving intelligibility is Local Interpretable Model-Agnostic Explanations (LIME), which focuses exclusively on local interpretation. It can be applied to any classifier or regressor, by approximating it locally with an interpretable model. Plus, in essence, it generates a new dataset composed of perturbed samples and their corresponding predictions, to test what happens when data is varied and how this affects the outcome. When training the interpretable model, the algorithm weighs the sampled instances by the proximity to the target instance *(Molnar, 2022).*

In general, it is evident that all labels are primarily affected by the batch's temperature and duration, which are the most relevant features for the model's overall behavior. The difference

relies on the values taken. While Void units have a short duration and low temperatures recorded, Inefficient batches are lasting, despite the low temperatures. At the same time, Wrong batches seem to correspond to undercooked meals with low temperatures, as well. Thus, commonly in these three labels, the average temperature is reduced, regardless of whether the cause is energy dispersion or short duration. Moreover, since the Wrong label is solely a function of food quality, it is largely influenced by temperature. Such ambiguity in other aspects allows it to display certain misleading traits, such as excessive *Open Door Minutes* and *Duration* values, in some cases.

To illustrate this, a random batch was chosen from the test set *(Fig. 34 in the appendix).* According to the model, there is a 52% confidence level that this data point belongs to the Wrong class, when in fact it corresponds to an Inefficient production. The features that lead the model to a bad allocation are an extremely long duration of about 2 hours and a low average temperature. In fact, when a batch is prepared in an inefficient manner, where the cooking time is prolonged, but the door is open most of the time, it is difficult to determine whether the food was properly cooked or not.

Besides, considering an example of misidentification involving the Void class, another row was randomly selected *(Fig. 35 in the appendix).* Based on the algorithm's judgment, there is a 98% probability that the meal was interrupted, but it was in reality of low quality. Curiously, only the fact that the door was opened once pushes the label towards Wrong, as Void batches are typically valued as 0. This confirms the intuition that relying only on the quality of food and having ambiguous measures of inefficiency can lead to some problems, and it may explain the confusion between the three classes.

5. Challenges & Limitations

The model built produces fairly good results, with an accuracy rate of almost 91%. In fact, although the Wrong class performed relatively poorly, the label still managed to achieve an F1 score of 62%, which seems reasonable. Yet, there are certainly limitations, which can be turned into opportunities for improvement and, for that reason, should be addressed.

Firstly, as explained before, the limited amount of labeled data compromises the performance of the model from the very beginning of the pipeline. Based on the results of the above experiment, a portion of 30% of the labeled data could provide a 76% accuracy rate, using a semi-supervised solution. However, the results can be further improved if more batches were labeled. This way, the algorithm would have access to a greater number of examples, allowing it to better absorb the essence of each class and perform a more sustained propagation. Later, feeding the model with labels that are more confidently distributed would benefit its training and performance. It is, however, important to note that labeling efforts should be directed primarily at minority groups, at least as a first step, to ensure that all classes are represented and treated equally.

Additionally, the ambiguity in class Wrong and the fact that labels do not mutually exclude one another complicate the classification task. Thus, the definition and number of labels should be reviewed. Rather, adding more features would be another alternative for improving differentiation. In fact, other types of data, such as humidity, are collected by the devices.

Finally, even IoT systems are susceptible to flaws, which can be caused, for instance, by a bad connection or a misuse of the device on the part of the employee. Therefore, some errors were discovered in the oven's database. A more concerning example relates to batches that were incorrectly split and assigned different IDs. Consequently, each batch's data is at risk, as it may be unreliable and influence the prediction. Taking steps to address this issue will ensure that the classifier uses good-quality data, promoting a successful outcome.

6. Recommendation for Future Steps

In the future, Jerónimo Martins plans to prototype this pilot project, so that batch surveillance and process optimization can be performed close to real-time. To accomplish this, it is necessary to settle and run the trained algorithm on a server, which, in turn, should be connected to the database where the oven updates are stored. Upon implementation, if a batch of food is not properly prepared, an alert should be sent to the device, so that staff members are informed and can act accordingly. Later, to ensure optimal performance, the model should be retrained periodically in light of new data collected.

With this in mind, a partnership with Amazon Web Services (AWS) was established. In fact, a few initial meetings have already taken place to implement the model in an agile manner. Apart from Cloud computing, AWS provides other solutions that can facilitate all steps of a Data Science project, from data curation to deployment. Among the most advantageous built-in tools for the present case is Amazon SageMaker Ground Truth service, which provides a wide range of options for labeling the data in a less costly and time-consuming manner. Nevertheless, further discussions with AWS technicians are necessary to better understand the different options and tailor them to meet the business' needs. Plus, with extended computational resources, the model's performance can also be potentially improved. For instance, a broader range of hyperparameters can be tested, when fine-tuning.

7. Conclusion

In *Hopping's (2000)* view, the history of retail industry reflects the progress of technology. In fact, retailers have been able to streamline and accelerate their processes over the years by implementing technology, from ATM card payments to RFID systems.

Nova & Go store exemplifies this increased ability to modernize and improve the shopping experience. Among other things, this store can gather various types of data, from sales, in-store

product placement, inventory levels or even the cooking process of takeaway meals. For the purpose of this thesis, the latter type of data was analyzed.

To constantly ensure the quality and safety of ready-to-go meals, the surveillance of the store's food batches was automated. Previously, a staff member would need to inspect the oven's screen manually, which was a time-consuming, subjective process that could result in human error. Alternatively, an algorithm for multi-class classification was now developed, based on the data collected from the store's semi-automated ovens. As a result of incorporating the model into the store's daily activities, an alert could be sent whenever the food quality is not adequate. Further, an analysis of the inefficient batches would allow an understanding of what are the poor practices in place. Leaving the door open is, for example, a significant risk factor. Identifying the primary reasons for this action can help reduce energy waste and, in turn, reduce costs.

8. References

Akyol, Kemal. 2022. “Coronary artery disease classification with support vector machines tuned via randomized search cross-validation.” 2022. https://assets.researchsquare.com/files/rs-1551634/v1/cf26adaf-67c7-42c1-808b-5415f6718487.pdf?c=1650308370.

Aly, Mohamed. 2005. “Survey on Multiclass Classification Methods.” Academia.edu. October 13, 2005. https://www.academia.edu/29135582/Survey_on_Multiclass_Classification_Methods.

Assistant Secretary for Public Affairs (ASPA). 2022. “Cook to a Safe Minimum Internal Temperature.” FoodSafety.gov. November 2, 2022. https://www.foodsafety.gov/food-safety-charts/safe-minimum-internal-temperatures.

Bergman, Liron, and Yedid Hoshen. 2020. “Classification-Based Anomaly Detection for General Data.” ArXiv.org. May 5, 2020. https://arxiv.org/abs/2005.02359.

Bergstra, James, and Yoshua Bengio. 2012. “Random Search for Hyper-Parameter Optimization.” 2012. https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf.

Berisha-Shaqiri, Aferdita. 2014. “Management Information System and Decision-Making.” 2014.https://www.researchgate.net/publication/287205806_Management_Information_System_and_Decision-Making.

Brynjolfsson, Erik, and Andrew McAfee. 2018. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. Vancouver, B.C.: Langara College.

Centers for Disease Control and Prevention. 2022. “Chicken and Food Poisoning.” Centers for Disease Control and Prevention. Centers for Disease Control and Prevention. October 31, 2022. https://www.cdc.gov/foodsafety/chicken.html.

Chakraborty, Debaditya, and Hazem Elzarka. 2017. “Advanced Machine Learning Techniques for Building Performance.” 2017. https://www.researchgate.net/profile/Debaditya-Chakraborty/publication/326542245_Advanced_machine_learning_techniques_for_building_performance_simulation_a_comparative_analysis/links/6016581492851c2d4d06bdec/Advanced-machine-learning-techniques-for-building-performance-simulation-a-comparative-analysis.pdf.

Chandola, Varun, Arindam Banerjee, and Vipin Kumar. 2009. “Anomaly Detection: A Survey.” University of Illinois Urbana-Champaign. Association for Computing Machinery (ACM). July 1, 2009. https://experts.illinois.edu/en/publications/anomaly-detection-a-survey.

Chen, Tianqi, and Carlos Guestrin. 2016. “XGBoost: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.” ACM Conferences. August 1, 2016. https://dl.acm.org/doi/10.1145/2939672.2939785.

Chen, Xu, and Tao Wang. 2017. “Combining Active Learning and Semi-Supervised Learning by Using Selective Label Spreading: Semantic Scholar.” Undefined. January 1, 2017. https://www.semanticscholar.org/paper/Combining-Active-Learning-and-Semi-Supervised-by-Chen-Wang/5fb5226137f9ddc8455358beaaadc5e39a57adc1.

Deloitte. 2011. “Data as a Strategic Asset.” Deloitte United States. 2011. https://www2.deloitte.com/us/en/pages/consulting/articles/data-strategic-asset.html.

Didaci, Luca, and Fabio Roli. 2006. “Using Co-Training and Self-Training in Semi-Supervised Multiple Classifier Systems.” SpringerLink. Springer Berlin Heidelberg. January 1, 2006. https://link.springer.com/chapter/10.1007/11815921_57.

d'Sa, Ashwin Geet, Irina Illina, Dominique Fohr, Dietrich Klakow, and Dana Ruiter. 2020. “Label Propagation-Based Semi-Supervised Learning for Hate Speech Classification.” Inria. October 12, 2020. https://hal.inria.fr/hal-02964065.

Döring, Matthias. 2018. “Performance Measures for Multi-Class Problems.” Data Science Blog: Understand. Implement. Succed. December 4, 2018. https://www.datascienceblog.net/post/machine-learning/performance-measures-multi-class-problems/.

Euronews. 2022. “Famílias e Empresas Europeias Asfixiadas Pela Crise Energética.” Euronews. August 22, 2022. https://pt.euronews.com/2022/08/22/familias-e-empresas-europeias-asfixiadas-pela-crise-energetica.

Gazzola, Patrizia, Daniele Grechi, Ilaria Martinelli, and Roberta Pezzetti. 2022. “The Innovation of the Cashierless Store.” MDPI. Multidisciplinary Digital Publishing Institute. February 11, 2022. https://www.mdpi.com/2071-1050/14/4/2034/htm.

Grandini, Margherita, Enrico Bagli, and Giorgio Visani. 2020. “Metrics for Multi-Class Classification: An Overview.” ArXiv.org. August 13, 2020. https://arxiv.org/abs/2008.05756v1.

Gregory, Jonathan. 2015. “Conexxus.” 2015. https://www.conexxus.org/sites/default/files/book_files/Accenture-The-Internet-Of-Things.pdf.

Hand, David, and Robert Till. 2001. “ A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems.” 2001. https://link.springer.com/content/pdf/10.1023/A:1010920819831.pdf.

Har, Loh Li, Umi Kartini Rashid, Seah Choon Sen, Loh Yin Xia, and Lee Te Chuan. 2022. “Revolution of Retail Industry: From Perspective of Retail 1.0 to 4.0.” Procedia Computer Science. Elsevier. March 8, 2022. https://www.sciencedirect.com/science/article/pii/S1877050922003714?fr=RR-2&ref=pdf_download&rr=777a9fa62f08866f.

Hawkins, D. 2014. *Identification of Outliers*. Springer.

Hevner, Alan R., Salvatore T. March, Jinsoo Park, and Sudha Ram. 2004. “Design Science in Information Systems Research.” University of Arizona. Management Information Systems Research Center. 2004. https://experts.arizona.edu/en/publications/design-science-in-information-systems-research.

Hicks, Dylan, Byron J. Gao, Hannah M. Bowles, and Kevin Mannix. 2013. “SmartMart: IoT-Based In-Store Mapping for Mobile Devices.” Innovating the Retail Industry: an IoT Approach. January 1, 2013. https://web.archive.org/web/20190429035546id_/https://eudl.eu/pdf/10.4108/icst.collaboratecom.2013.254116.

Hopping, Dan. 2000. “Technology in Retail: Semantic Scholar.” Undefined. January 1, 2000. https://www.sciencedirect.com/science/article/pii/S0160791X99000421?via=ihub.

IBM Cloud Education. 2021. “What Is Data Labeling?” IBM. 2021. https://www.ibm.com/cloud/learn/data-labeling.

International Energy Agency. 2022. “Global Energy Crisis – Topics.” IEA. 2022. https://www.iea.org/topics/global-energy-crisis.

J.Knutson, Bonnie. 1988. “Ten Laws of Customer Satisfaction.” The Cornell Hotel and Restaurant Administration Quarterly. No longer published by Elsevier. December 1988. https://www.sciencedirect.com/science/article/pii/S0010880488800107.

Kaspersky. 2021. “Things Just Got Real: 61% of Businesses Already Use IOT Platforms despite Security Risks.” Www.kaspersky.com. May 26, 2021. https://www.kaspersky.com/about/press-releases/2020_things-just-got-real-61-of-businesses-already-use-iot-platforms-despite-security-risks.

Kellmereit, Daniel, and Daniel Obodovski. 2013. *The Silent Intelligence: The Internet of Things*. San Francisco, CA: DnD Ventures.

Kohavi, Ron, and George H.John. 1998. “Wrappers for Feature Subset Selection.” Artificial Intelligence. Elsevier. May 19, 1998. https://www.sciencedirect.com/science/article/pii/S000437029700043X.

Kotsiantis, S., D. Kanellopoulos, and P. Pintelas. 2006. “Handling Imbalanced Datasets: A Review: Semantic Scholar.” Undefined. January 1, 2006. https://www.semanticscholar.org/paper/Handling-imbalanced-datasets%3A-A-review-Kotsiantis-Kanellopoulos/95dfdc02010b9c390878729f459893c2a5c0898f.

Le, Thi-Thu-Huong, Yustus Eko Oktian, and Howon Kim. 2022. “XGBoost for Imbalanced Multiclass Classification-Based Industrial Internet of Things Intrusion Detection Systems.” MDPI. Multidisciplinary Digital Publishing Institute. July 16, 2022. https://www.mdpi.com/2071-1050/14/14/8707.

Lee, Ernesto. 2021. “(PDF) How Do We Build Trust in Machine Learning Models?” 2021. https://www.researchgate.net/publication/350785349_How_do_we_build_trust_in_machine_learning_models.

Lipton, Zachary C. 2018. “The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability Is Both Important and Slippery.: Queue: Vol 16, No 3.” Queue. 2018. https://dl.acm.org/doi/10.1145/3236386.3241340.

Luque, Amalia, Alejandro Carrasco, Alejandro Martín, and Ana de las Heras. 2019. “The Impact of Class Imbalance in Classification Performance Metrics Based on the Binary Confusion Matrix.” Pattern Recognition. Pergamon. February 28, 2019. https://www.sciencedirect.com/science/article/pii/S0031320319300950.

Mckinsey. 2018. “Achieving Business Impact with Data - Mckinsey & Company.” 2018. https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20analytics/our%20insights/achieving%20business%20impact%20with%20data/achieving-business-impact-with-data_final.ashx.

Microsoft. 2020. “InterpretML: A Toolkit for Understanding Machine Learning Models.” 2020. https://www.microsoft.com/en-us/research/uploads/prod/2020/05/InterpretML-Whitepaper.pdf.

Molnar, Christoph. 2022. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Munich, Germany: Christoph Molnar.

O'Farrell, M., E. Lewis, C. Flanagan, W. Lyons, and N. Jackman. 2005. “Combining Principal Component Analysis with an Artificial Neural Network to Perform Online Quality Assessment of Food as It Cooks in a Large-Scale Industrial Oven: Semantic Scholar.” Undefined. January 1, 2005. https://www.semanticscholar.org/paper/Combining-principal-component-analysis-with-an-to-a-O'Farrell-Lewis/88c4db22c5af7732ba7a3349fcaa7d3b6b2e3bd7.

Phillips, Tessa, and W. Abdulla. 2021. “Developing a New Ensemble Approach with Multi-Class Svms for Manuka Honey Quality Classification: Semantic Scholar.” Undefined. January 1, 2021. https://www.semanticscholar.org/paper/Developing-a-new-ensemble-approach-with-multi-class-Phillips-Abdulla/db356799965487987d0f6021c67247fb5a79dc8c.

Popa, Alexandru, Mihaela Hnatiuc, Mirel Paun, Oana Geman, D. Jude Hemanth, Daniel Dorcea, Le Hoang Son, and Simona Ghita. 2019. “An Intelligent IOT-Based Food Quality Monitoring Approach Using Low-Cost Sensors.” MDPI. Multidisciplinary Digital Publishing Institute. March 13, 2019. https://www.mdpi.com/2073-8994/11/3/374.

Salgueiro, Maria. 2019. “Pingo Doce & Go: Experimentámos o Novo Conceito De Supermercado Sem Filas.” NiT. 2019. https://www.nit.pt/comida/gourmet-e-vinhos/ja-fomos-ao-novo-pingo-doce-sem-filas-ficamos-impressionados.

Smith, Matthew, and Francisco Alvarez. 2021. “Identifying Mortality Factors from Machine Learning Using Shapley Values – a Case of covid19.” Expert Systems with Applications. Pergamon. March 11, 2021. https://www.sciencedirect.com/science/article/pii/S0957417421002736?casa_token=xz1qMW-g2P8AAAAA%3AAJDY2Jy7F_Po_3iuGBQzTt92R-wuev74X-9H4b1hq86NatORQUjRPNt8N8XolWYxWTyFReqdT3A.

Tanrıseven, Sercan, Niyazi Uğur, Alper Kanak, and Salih Ergün. 2019. “ERINOKS: EneRgy-Efficient INduction-Based Food Processing for Optimized KitchenS.” 2019. https://ieeexplore.ieee.org/document/8702571/.

Shee Tan , Yee, Yen TingNg, and Jonathan Sze ChoongLow. 2017. “Internet-of-Things Enabled Real-Time Monitoring of Energy Efficiency on Manufacturing Shop Floors.” Procedia CIRP. Elsevier. April 19, 2017. https://www.sciencedirect.com/science/article/pii/S2212827116314111?ref=pdf_download&fr=RR-2&rr=7771da13d89a6683.

Vargheese, Rajesh, and Hazim Dahir. 2014. “An IoT/IoE Enabled Architecture Framework for Precision on Shelf Availability: Enhancing Proactive Shopper Experience.” 2014. https://ieeexplore.ieee.org/document/7004418.

Wang, Pan, Ricardo Valerdi, Shangming Zhou, and Ling Li. 2015. “Introduction: Advances in IOT Research and Applications - Information Systems Frontiers.” SpringerLink. Springer US. March 11, 2015. https://link.springer.com/article/10.1007/s10796-015-9549-2.

Weisenthal, Joe, and Tracy Alloway. 2022. “Nouriel Roubini Predicts a Crisis 'Worse' than the 1970s.” Bloomberg.com. Bloomberg. October 20, 2022. https://www.bloomberg.com/news/articles/2022-10-20/nouriel-roubini-predicts-a-crisis-worse-than-the-1970s?leadSource=uverify+wall.

Zou, Miao, Wu-Gui Jiang, Qing-Hua Qin, Yu-Cheng Liu, and Mao-Lin Li. 2022. “Optimized XGBoost Model with Small Dataset for Predicting Relative Density of Ti-6al-4v Parts Manufactured by Selective Laser Melting.” MDPI. Multidisciplinary Digital Publishing Institute. August 1, 2022. https://www.mdpi.com/1996-1944/15/15/5298.

9. Appendix

Appendix Tables

Table 1 – Top 10 of Most Prepared Food Recipes, in Total Number of Batches

	Food Recipe	Total Number of Batches	Frequency
0	MISTO PASTELARIA NOVO	1041	0.27
1	PAO	929	0.24
2	PAO DE QUEIJO PD	423	0.11
3	SOPA	234	0.06
4	FRANGO	183	0.05
5	PASTEL DE NATA	175	0.05
6	BURGUER VEGGIE	152	0.04
7	BURGUER CONGELADO	149	0.04
8	SALGADOS P.D	130	0.03
9	CROQUETES P.D	104	0.03

Table 2 – Description of the Cleaned Dataset's Features

Feature	Description
START_OFFSET	Total time that passed since the beggining of the batch, in seconds
DOOR_STATUS	Whether the door was opened or closed, at the moment of the update
CHAMBER_TEMPERATURE	Temperature of the chamber, in Celsius degrees
FOOD_TEMPERATURE_SET	Target temperature that should be achieved, in Celsius degrees
FOOD_TEMPERATURE	Temperature of the food, measured by a thermomether, in Celsius degrees
CC_BATCH	Identifier of the batch
START_DATE	Datetime when the batch was produced
ENERGY_CONSUMPTION	Total energy spent, in kWh
DURATION	Total duration, in seconds
OPEN_DOOR_COUNT	Total amount of times the door was opened
FOOD_TEMP_TARGET_VALUE_REACHED	Whether the target temperature was reached or not
LABEL	Classification of the batch
NAME	Identifier of the device
CC_NAME	Type of food recipe
START_OFFSET_MINUTES	Total time that passed since the beggining of the batch, in minutes
OPEN_DOOR_SECONDS	Total amount of time the door was left open, in seconds
OPEN_DOOR_MINUTES	Total amount of time the door was left open, in minutes
DATE	Date when the batch was produced
TIME	Time of the day when the batch was produced
HOUR	Hour of the day when the batch was produced
WEEKDAY	Weekday when the batch was produced
DAY	Day of the month when the batch was produced
MONTH	Month when the batch was produced
DURATION_MINUTES	Total duration, in minutes
ENERGY_SPENT_PER_MINUTE	Total energy spent per minute, in kWh
MAX_TEMPERATURE	Maximum food temperature achieved, in Celsius degrees
MIN_TEMPERATURE	Minimum food temperature achieved, in Celsius degrees
AVG_TEMPERATURE	Average food temperature
AVG_CHAMBER_TEMPERATURE	Average chamber temperature, in Celsius degrees
SECONDS_ABOVE_THRESHOLD	Total time the food temperature was above the threshold, in seconds
MINUTES_ABOVE_THRESHOLD	Total time the food temperature was above the threshold, in minutes

Table 3 – Description of the Labels, provided by Jerónimo Martins

Label	Class	Description
Inefficient	0	The batch was properly prepared but inefficient
OK	1	The batch was properly prepared and efficient
Void	2	The batch was very short
Wrong	3	The batch was not properly prepared

Table 4 – Other Models Tested, and the Accuracy Score obtained

Model	Accuracy Score
Logistic Regression	0.67
SVC	0.80
Decision Tree	0.86
XGBoost	0.89

Table 5 – Comparison of the Distribution of Labels before and after Label Spreading

Label	Class	Before	After
Unknown	-1	1072	0
Inefficient	0	160	486
OK	1	203	881
Void	2	47	74
Wrong	3	38	79

Appendix Figures

Fig. 1 – Data Science Research

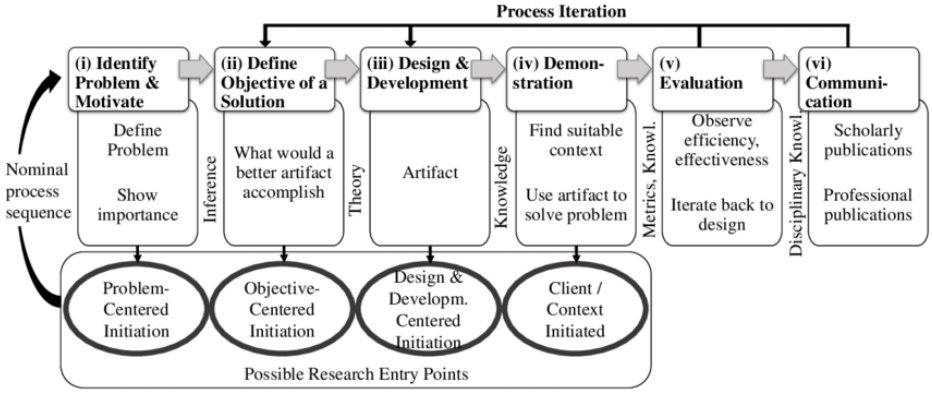


Fig. 2 – Total Number of Batches Prepared, by Month

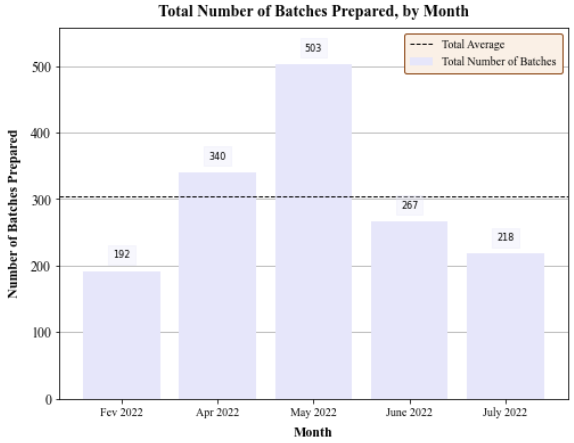


Fig. 3 – Average Number of Batches Prepared, by Weekday

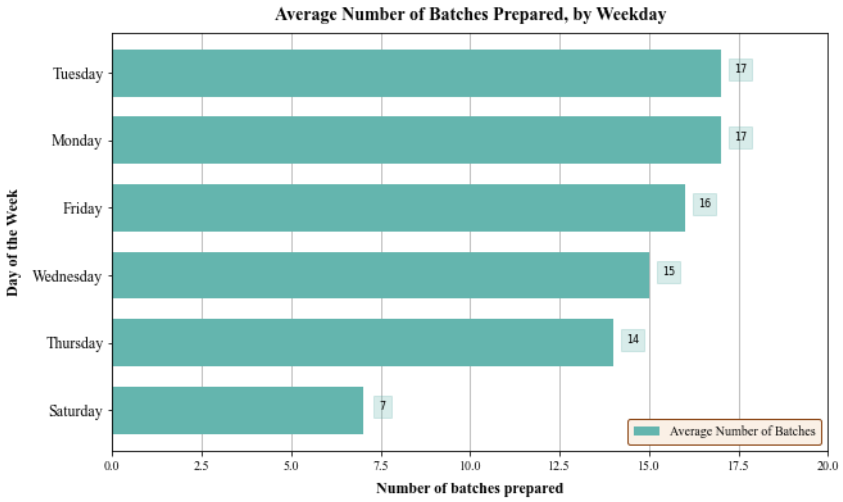


Fig. 4 – Average Number of Batches Prepared, by Hour

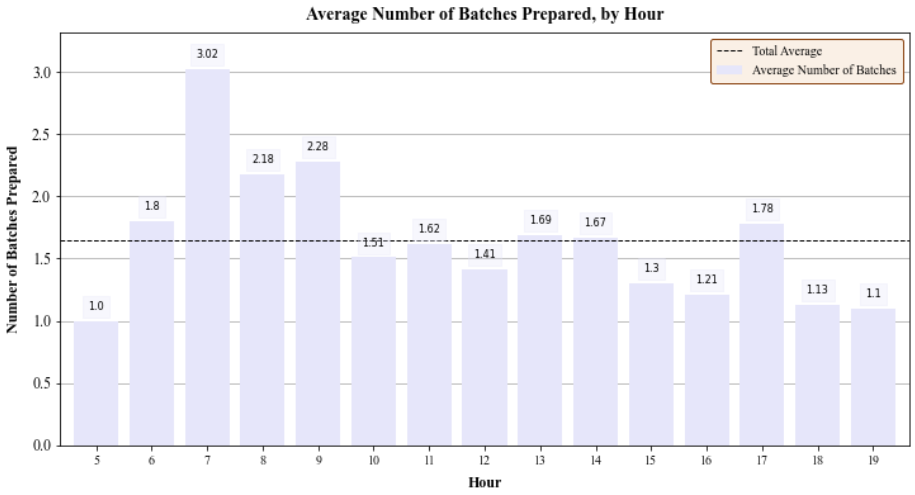


Fig. 5 – Percentage of Chamber Utilization, by Food Recipe

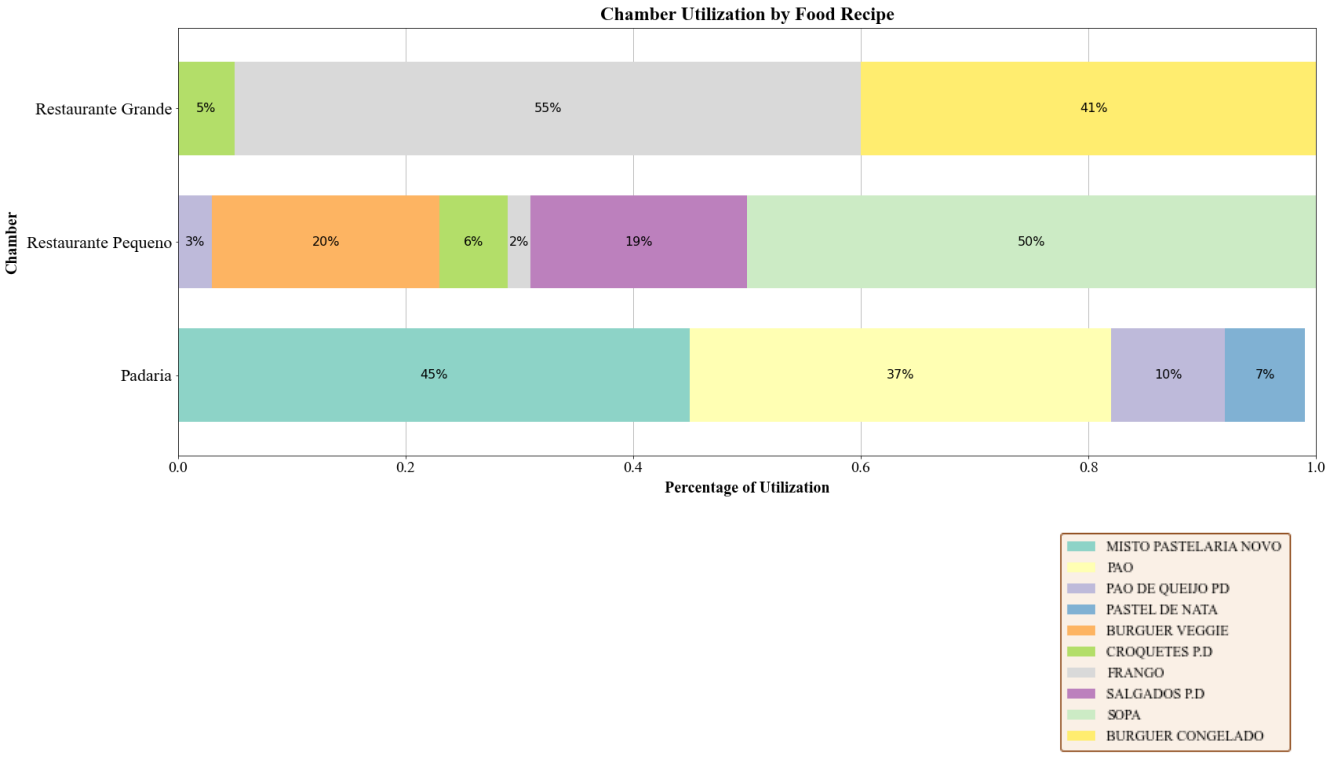


Fig. 6 – Percentage of Batches Produced, by Chamber

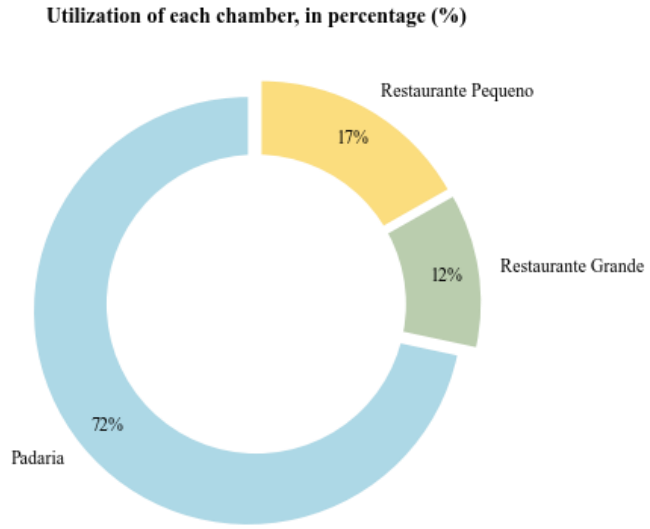


Fig. 7 – Average Duration by Food Recipe and by Chamber, in Minutes

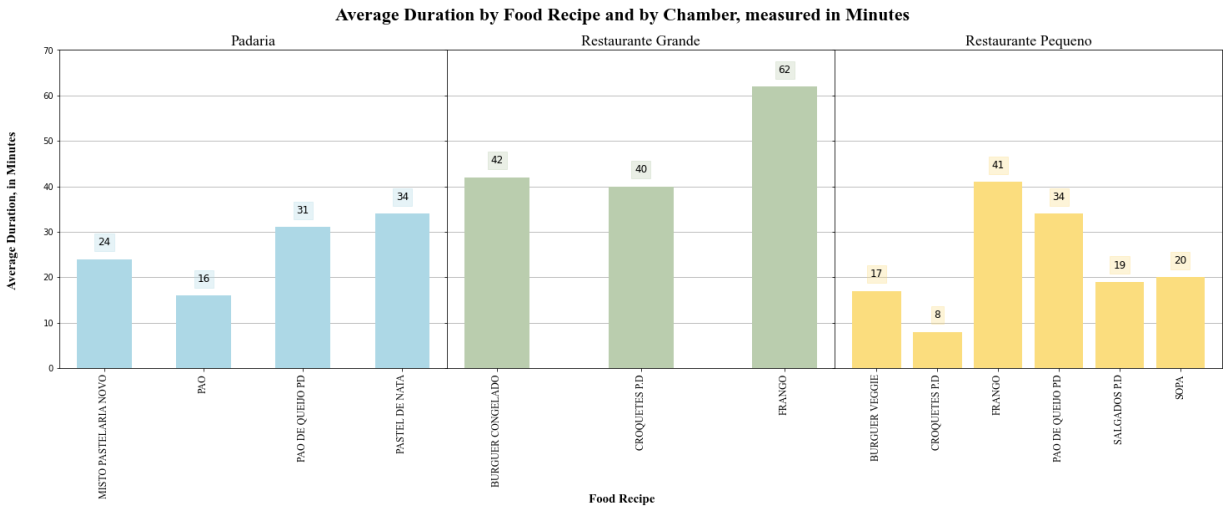


Fig. 8 – Total Number of Batches Prepared, by Hour and by Chamber

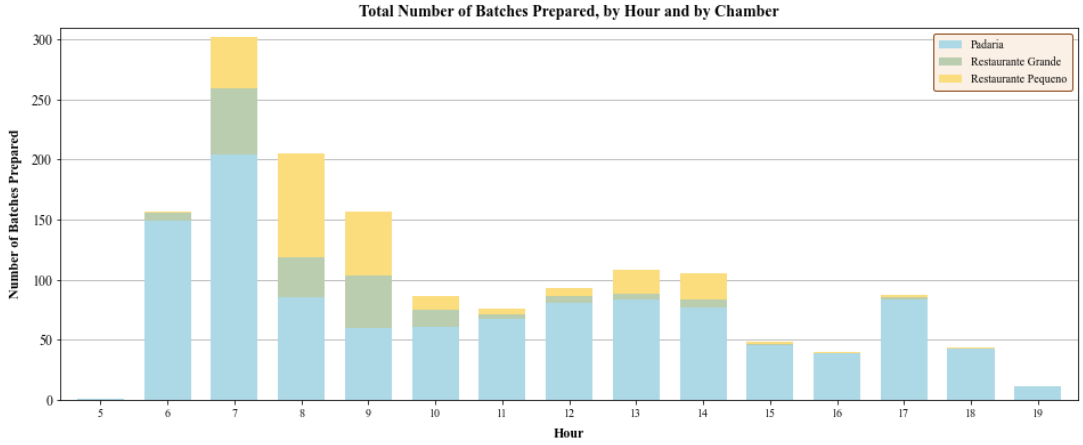


Fig. 9 – Average Duration per Batch, in Minutes, by Food Recipe

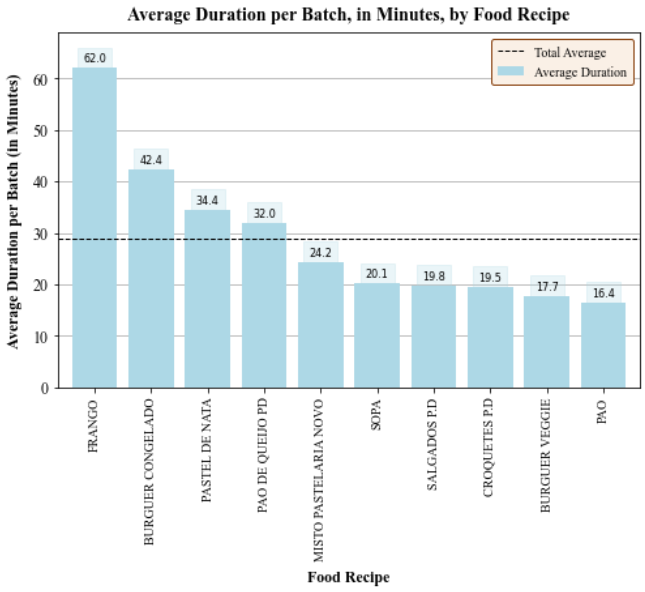


Fig. 10 – Average Energy Consumption per Batch, in kWh, by Food Recipe

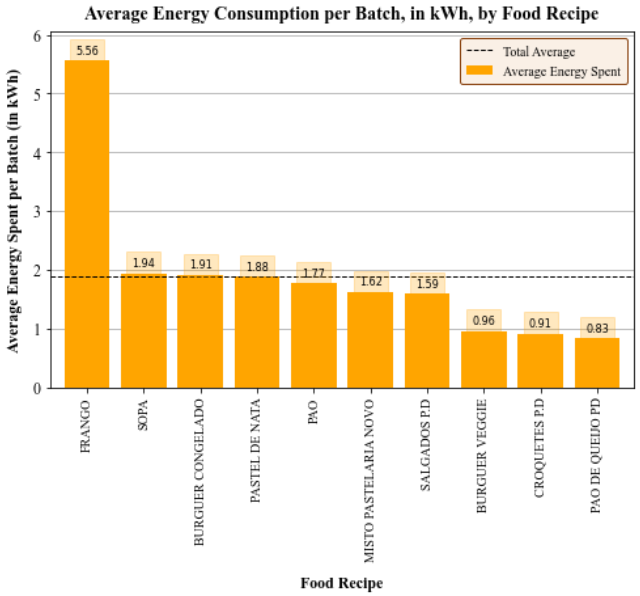


Fig. 11 – Average Energy Spent per Minute, in kWh, by Food Recipe

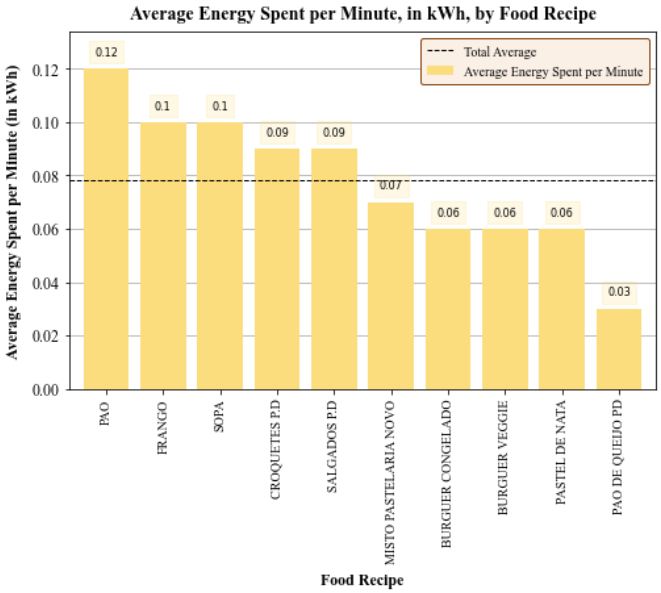


Fig. 12 – Average Energy Spent per Minute, in kWh, by Food Recipe and by Chamber

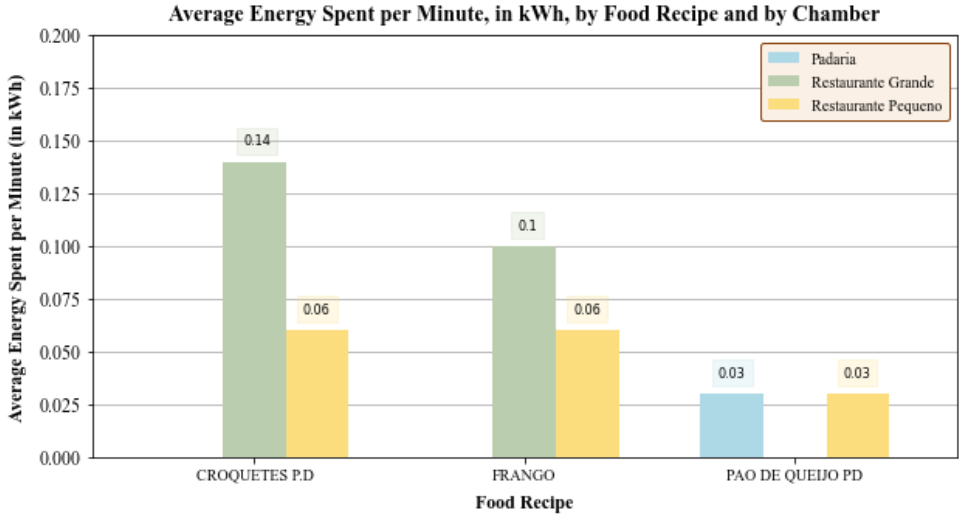


Fig. 13 – Possible Motives for an Exceptional Update, in percentage (%)

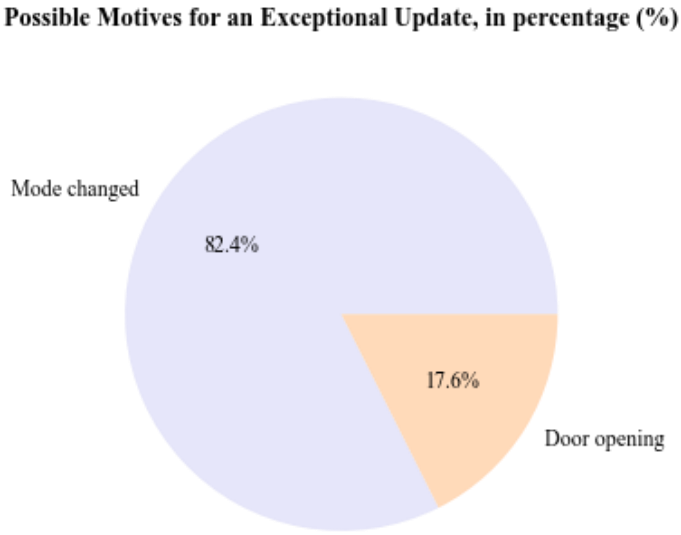


Fig. 14 – Distribution of the Number of Times the Door was Opened per Batch

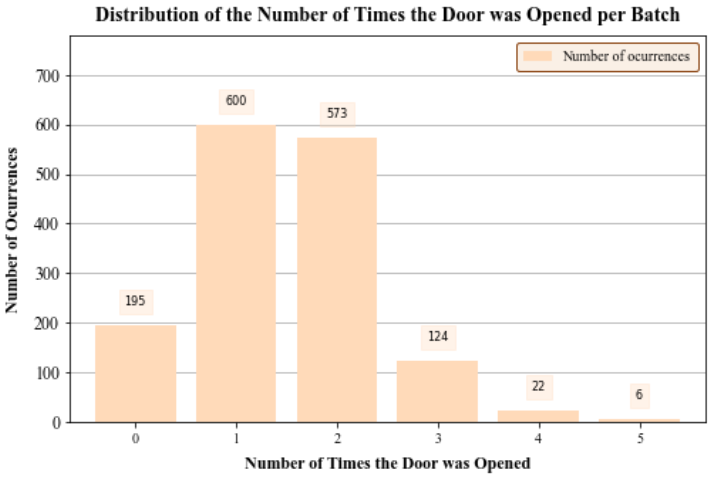


Fig. 15 – Average Number of Times the Door was Opened, by Food Recipe

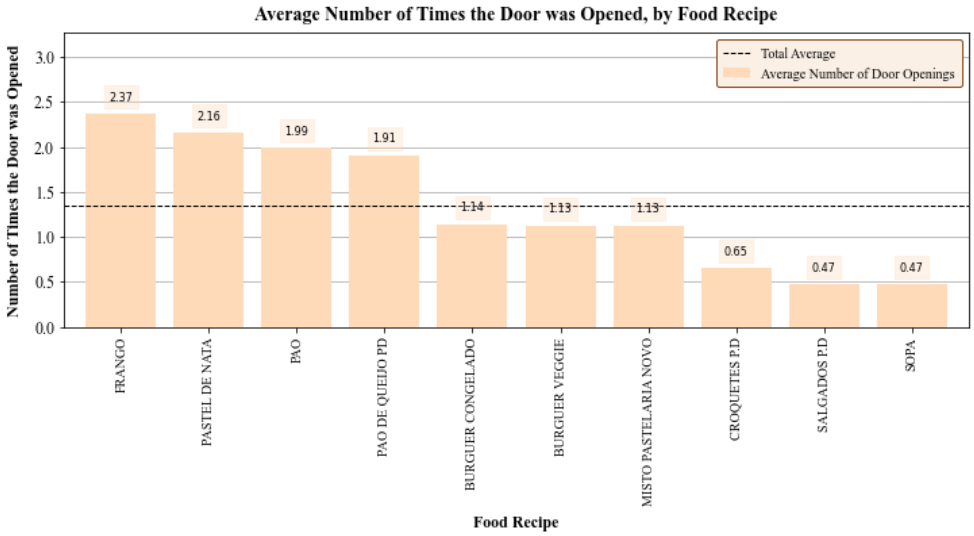


Fig. 16 – Average Total Time the Door was Opened, in Minutes, by Food Recipe

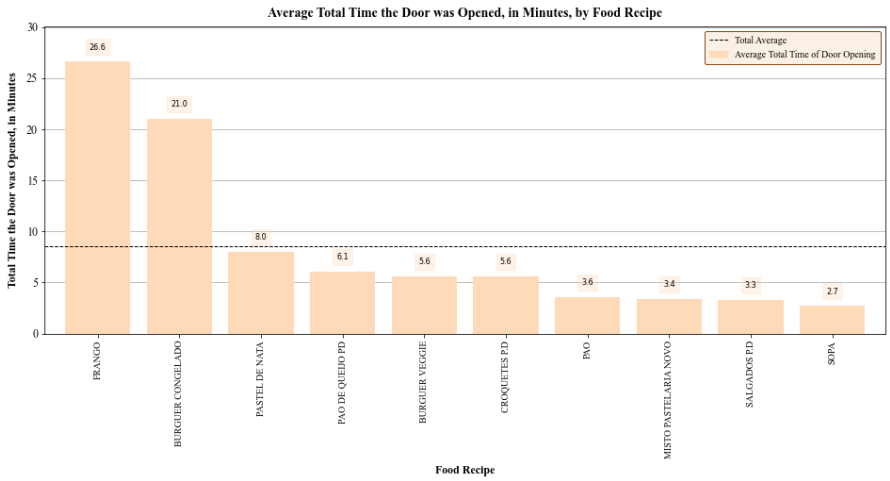


Fig. 17 – Number of Updates Sent Due to Opening the Door, by Start Offset

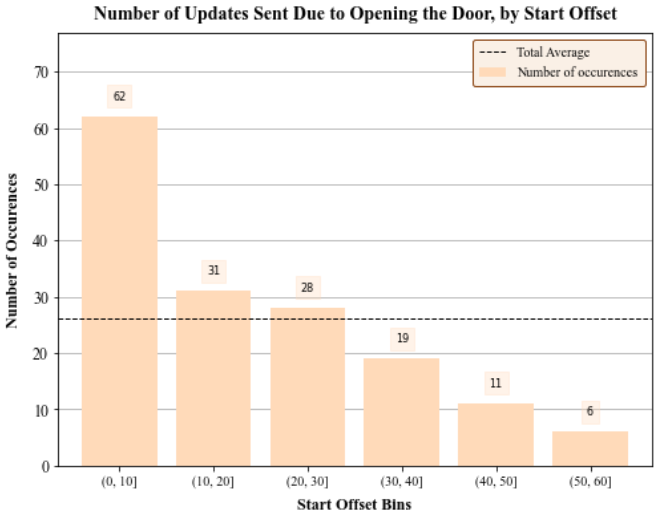


Fig. 18 – Average Chamber and Food Temperature, by Food Recipe

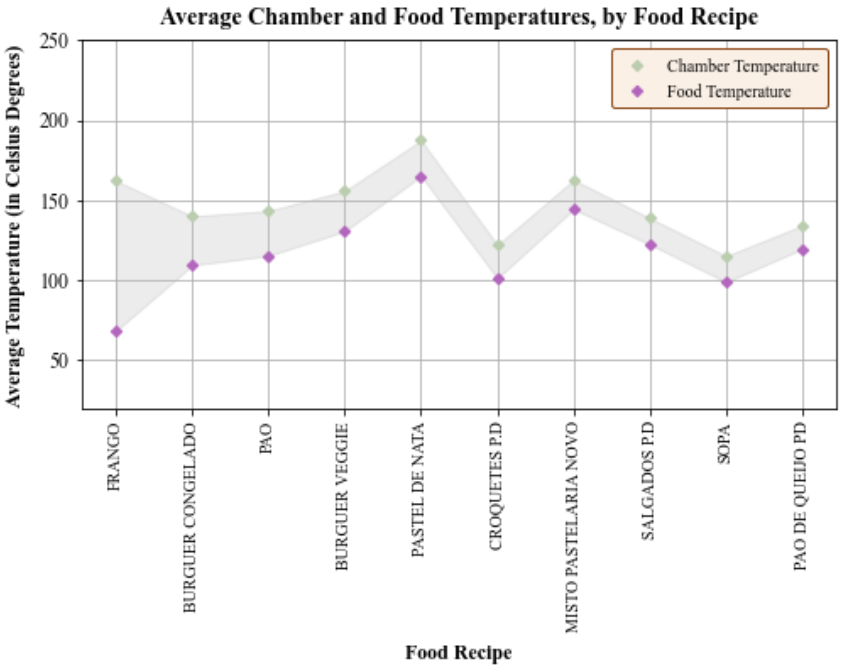


Fig. 19 – Average Maximum and Minimum Temperature, by Food Recipe

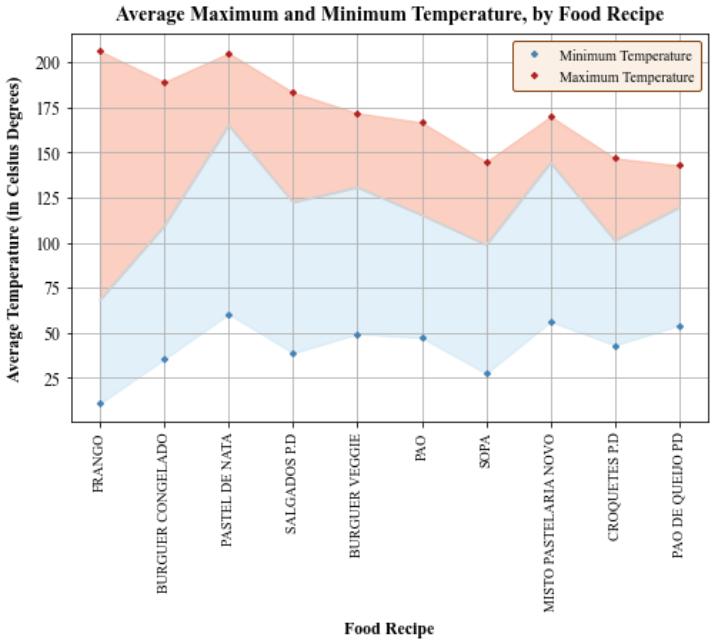


Fig. 20 – Distribution of Labels Across the Entire Dataset, in percentage (%)

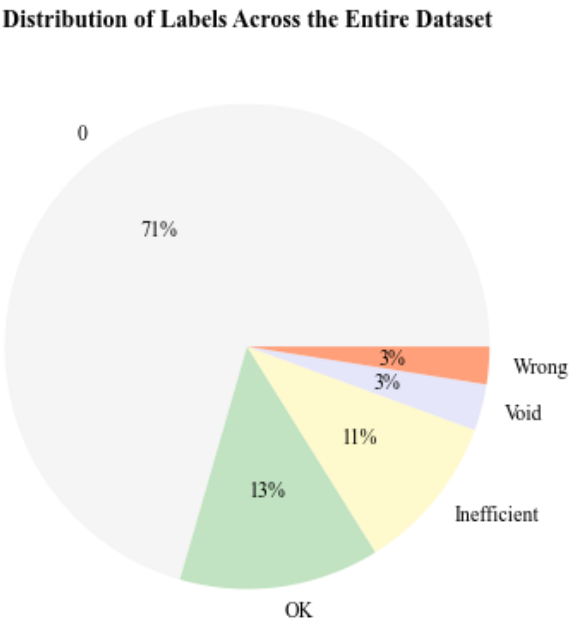


Fig. 21 – Number of Distinct Labels, by Food Recipe

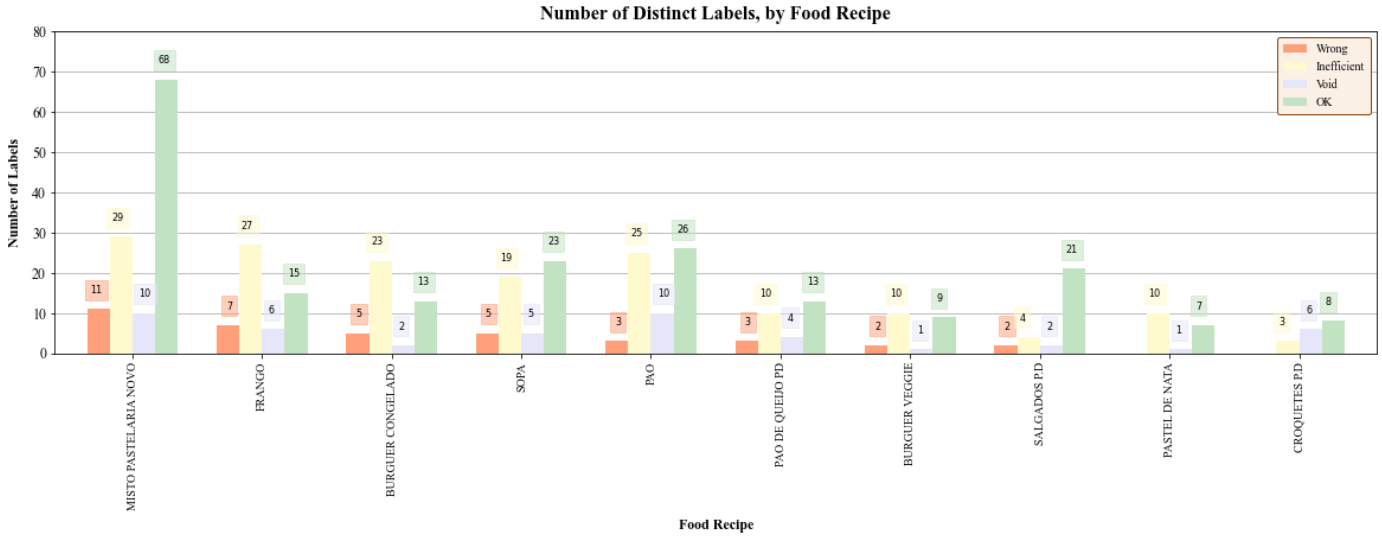


Fig. 22 – Top 5 Most Produced Food Recipes, in Number of Batches

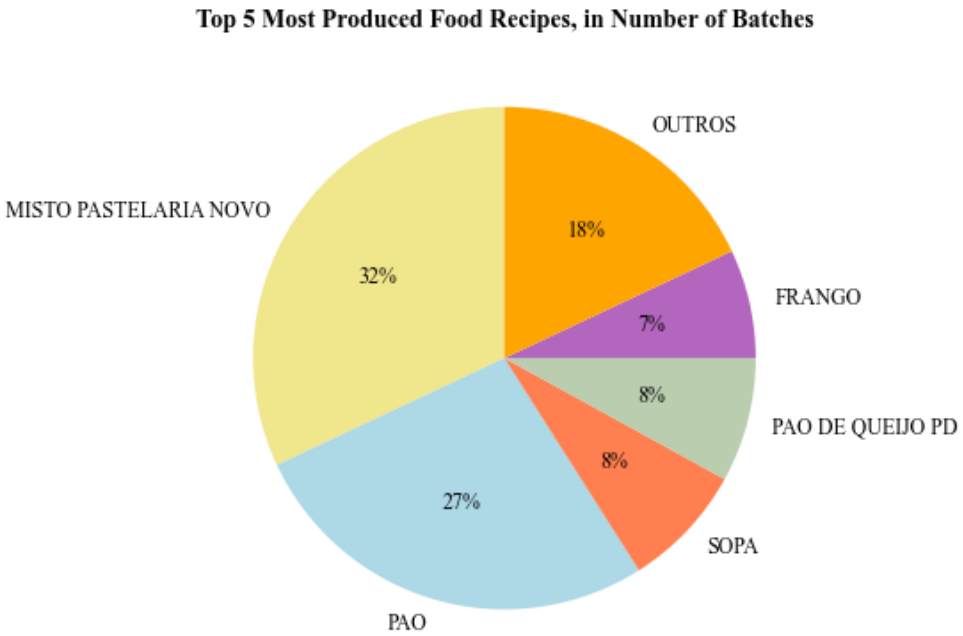


Fig. 23 – Average Total Time that the Door was Opened, in Minutes, by Recipe and Label

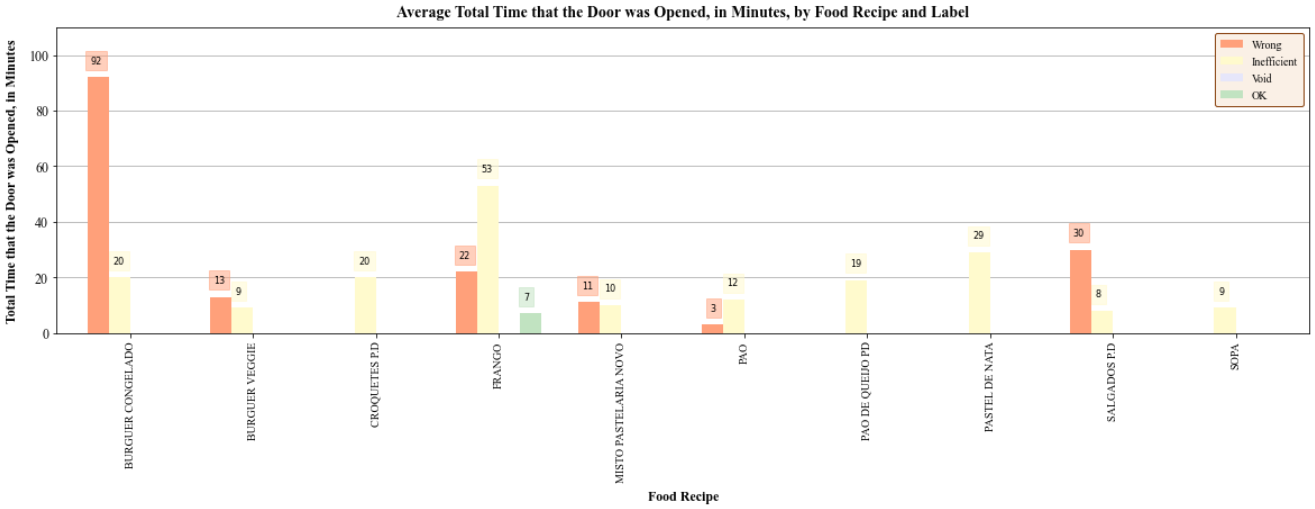


Fig. 24 – Average Food Temperature, by Food Recipe and Label

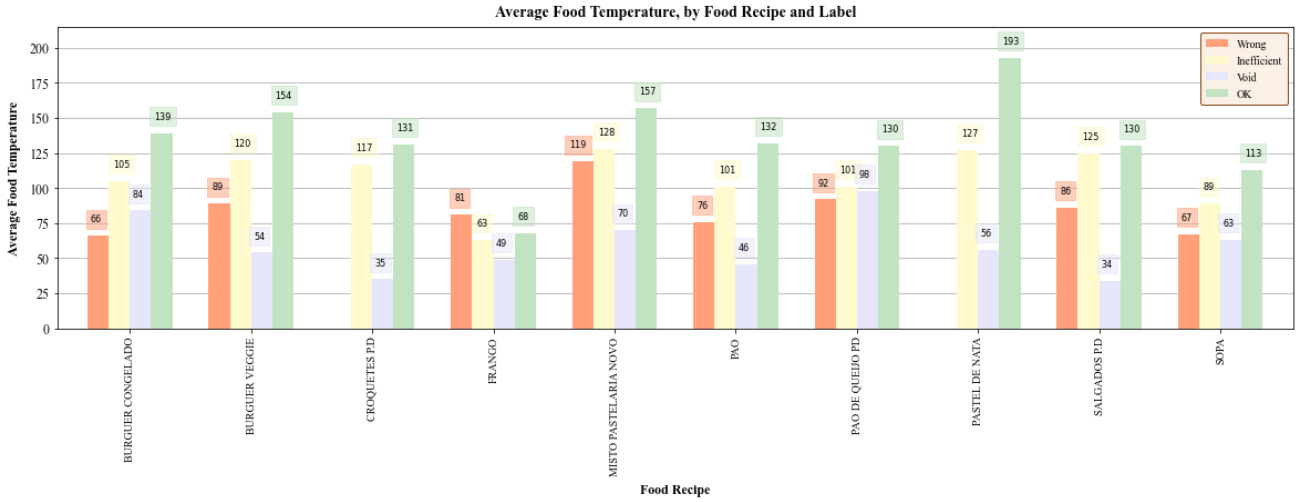


Fig. 25 – Average Total Duration in Minutes, by Food Recipe and Label

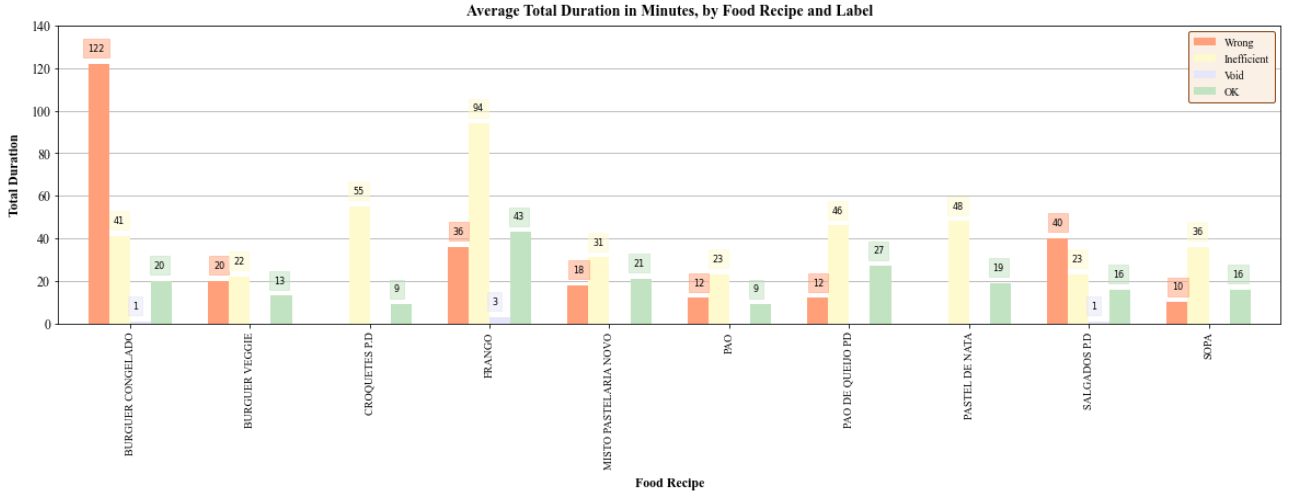


Fig. 26 – Confusion Matrix for the Label Spreading experiment with 30% of labeled data

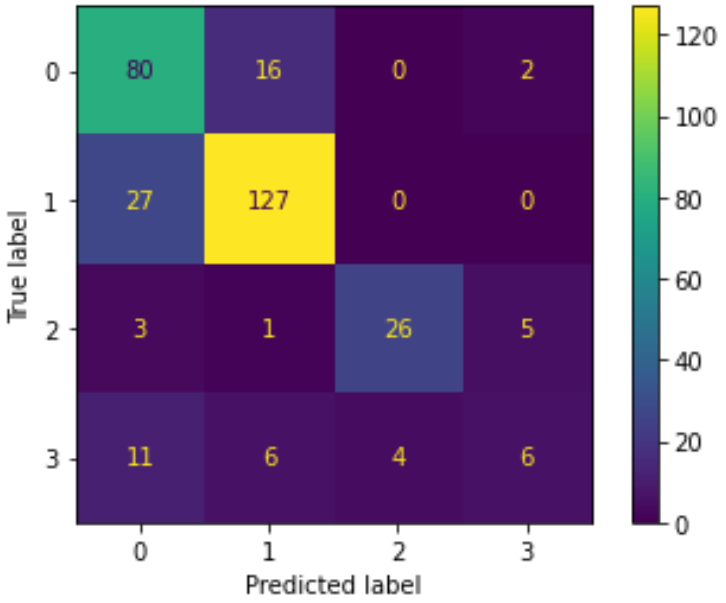


Fig. 27 – Confusion Matrix for the fine-tuned Model (using the Test Set)

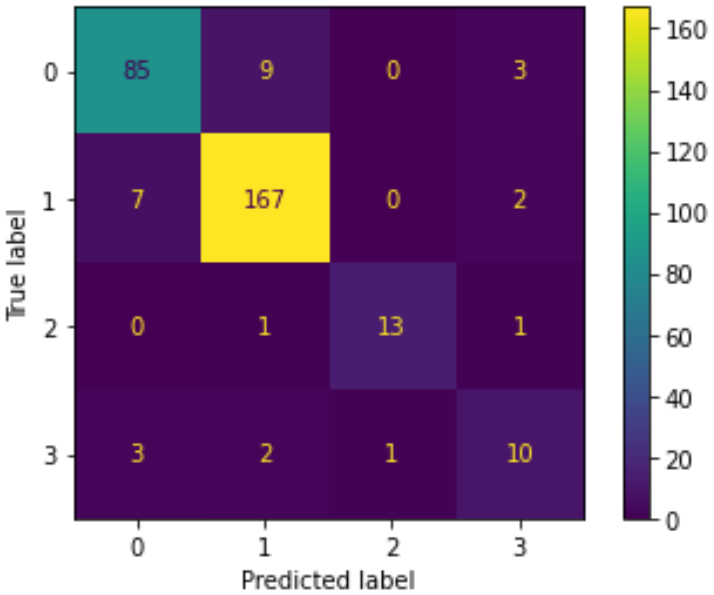


Fig. 28 – Classification Report obtained for the fine-tuned Model (using the Test Set)

	precision	recall	f1-score	support
0	0.89	0.88	0.89	97
1	0.93	0.95	0.94	176
2	0.93	0.87	0.90	15
3	0.62	0.62	0.62	16
accuracy			0.90	304
macro avg	0.85	0.83	0.84	304
weighted avg	0.90	0.90	0.90	304

Fig. 29 – SHAP Summary Plot

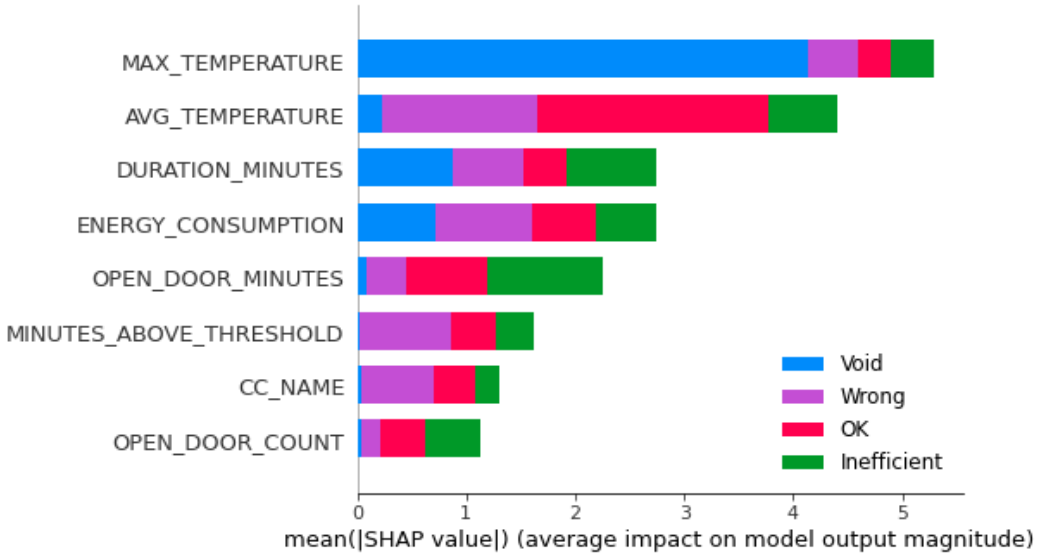


Fig. 30 – SHAP Summary Plot for Class 2 (Void)

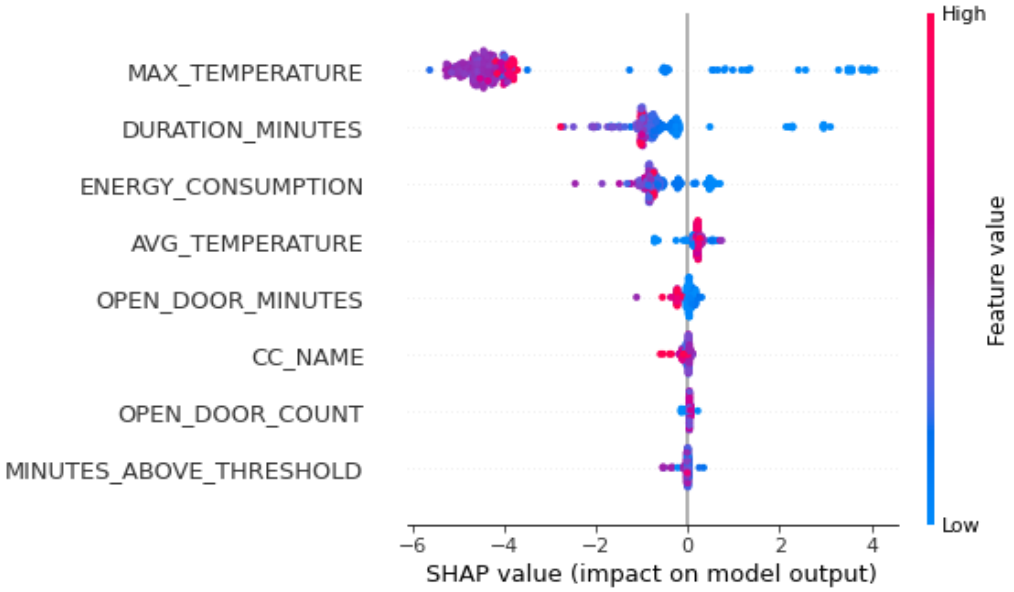


Fig. 31 – SHAP Summary Plot for Class 0 (Inefficient)

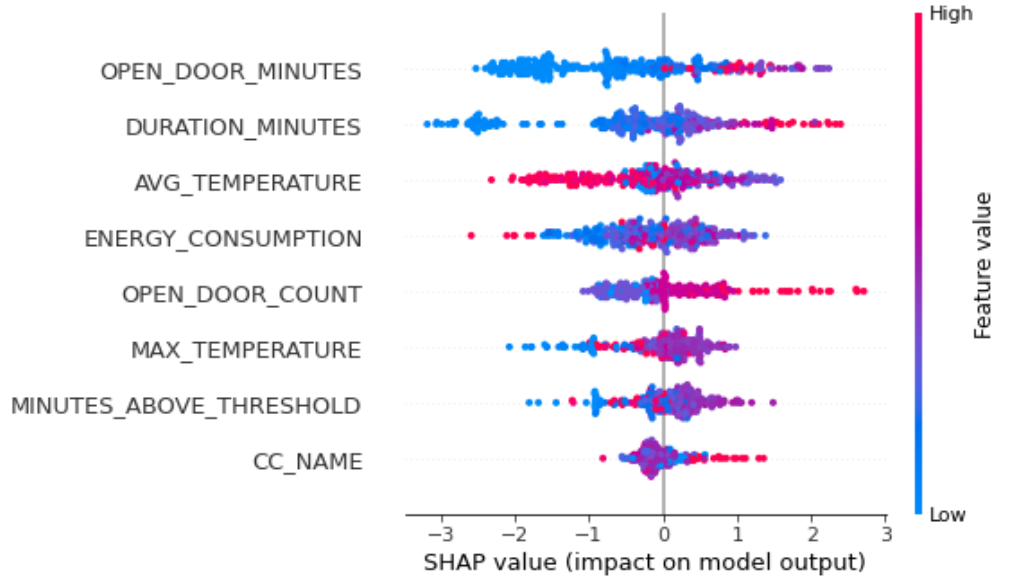


Fig. 32 – SHAP Summary Plot for Class 1 (OK)

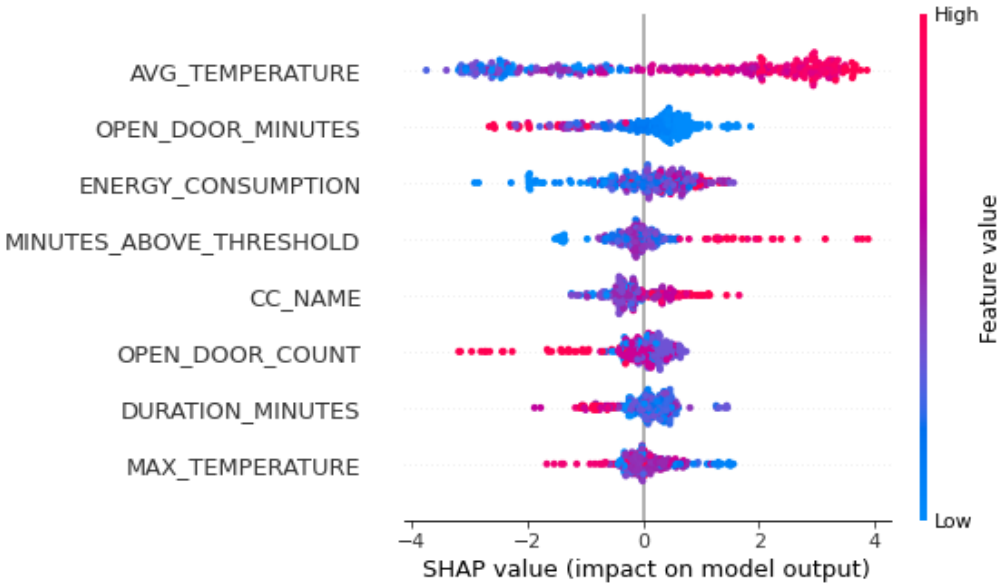


Fig. 33 – SHAP Summary Plot for Class 3 (Wrong)

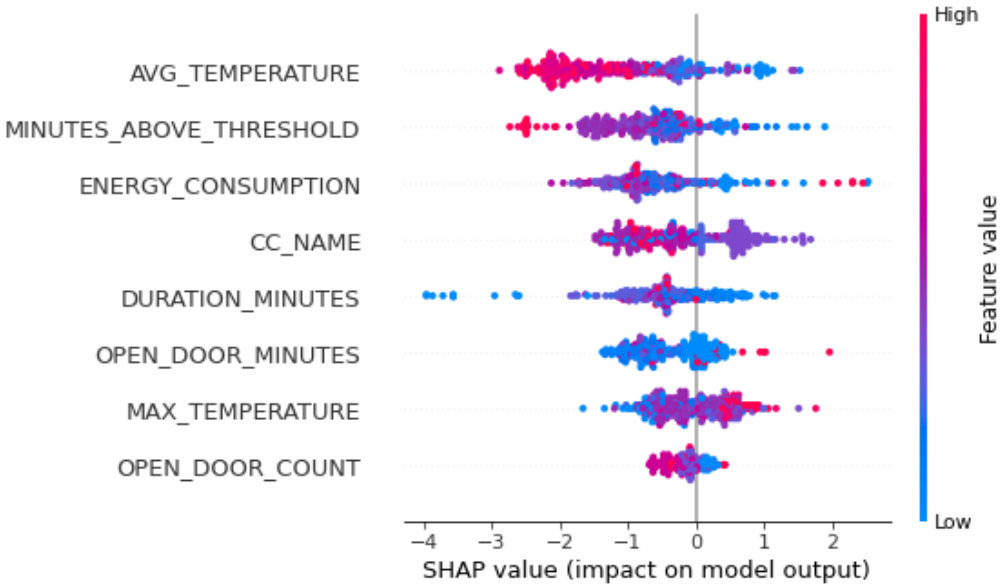


Fig. 34 – Example of an incorrect prediction between Wrong and Inefficient

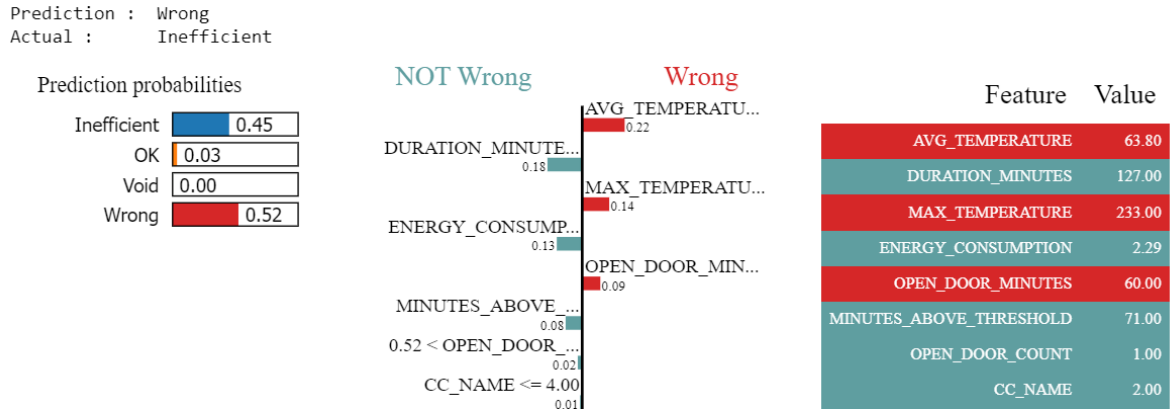


Fig. 35 – Example of an incorrect prediction between Void and Wrong