FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Development of Machine Learning models to predict glass quality of melting furnace

Francisco José Lousada Soares Nogueira Rodrigues

Mestrado em Engenharia Eletrotécnica e de Computadores

Supervisor: Luís Filipe Ribeiro dos Santos Guimarães

July 28, 2022

Resumo

Dada a situação geopolítica e o estado dos mercados de energia europeus no segundo semestre de 2022, existe uma necessidade cada vez mais urgente para uma redução das despesas energéticas, sejam elas em gás natural ou em eletricidade. Com um aumento esperado de cinquenta por cento em média dos preços de energia até ao final do ano, existe uma grande pressão sobre a indústria de produção de recipientes de vidro, devido à natureza contínua e produção em massa em que opera um forno moderno de fabrico de vidro.

A BA Glass tem estado na vanguarda da sustentabilidade, através de projetos redução de emissões, do consumo de água e de um reforço no uso de vidro reciclado. Este último aspeto será abordado neste estudo uma vez que o vidro é quase 100% reciclavél e a sua utilização é bem vista do ponto de vista de sustentabilidade bem como no de redução de custos energéticos, embora implique um aumento expectável de defeitos devido à presença de contaminantes nos resíduos reciclados.

O vidro reciclado, ou casco, como é designada na indústria, é comprado a entidades recicladoras e depois tratado por uma subsidiária da BA Glass, Norcasco. Alguns contaminantes representam uma ameaça substancial à integridade estrutural da embalagem, uma vez que têm pontos de fusão mais elevados do que outros materiais e desenvolverão uma tensão na estrutura da embalagem, comprometendo ainda mais a sua natureza frágil.

O estudo subdividiu-se em dois momentos, sendo que num primeiro momento foi efetuado uma análise detalhada do processo industrial moderno de fabrico de embalagens de vidro, bem como da sua composição através do estudo dos dados das equipas de fusão responsáveis pelo enfornamento das matérias-primas. Foi feita uma distinção entre os tipos de casco utilizado e a sua implicação na qualidade do produto final.

Numa segunda fase, a fim de melhorar a ferramenta desenvolvida pelo departamento de Data Science and Analytics para melhorar o consumo de energia, foi estudada uma possível melhoria nos processos de previsão da qualidade do vidro, utilizando ferramentas estatísticas e de aprendizagem de máquina. Ao obter acesso aos dados das máquinas de inspecção, foi possivél abranger toda a produção, de modo a criar um melhor indicador de qualidade que não dependesse da amostragem. Cruzando estes dados com os dados da equipa de fusão, bem como outros dados do forno já previamente recolhidos pela equipa, foi possível melhorar substancialmente a previsão dos defeitos através de inclusões de material não fundido.

Palavras-chave: Embalagens de vidro, Aprendizagem supervisionada, Vidro reciclado, Qualidade

Abstract

Given the geopolitical situation and the state of the European energy markets in the second half of 2022, there is an increasingly urgent need for a reduction in energy expenditure, be it natural gas or electricity. With energy prices expected to increase on average by fifty percent by the end of the year, this puts a strain on the glass container manufacturing industry due to the continuous state and mass production that modern glass making furnaces operate in.

BA Glass has been at the forefront of sustainability through projects that reduce emissions, water consumption, and glass recycling. This last aspect will be addressed in this study as glass is almost 100% recycled, and the use of recycled glass in the batch composition is regarded as a form of sustainability and energy cost reduction, although it comes with an increase in defects due to the presence of contaminants in the recycled waste.

Recycled glass, or cullet, as it is referred to in the industry, is purchased from recycling entities and then treated by a subsidiary of BA Glass, Norcasco. Some contaminants pose a substantial threat to the container's structural integrity, as they have higher melting points than other materials and will develop a strain in bottle or jar structure, compromising this already brittle form of packaging. The study was divided into two main stages.

The first stage was a detailed analysis of the modern industrial process of glass packaging manufacturing and its composition by studying the data of the batch house and furnace teams responsible for the raw material processing. A distinction between the types of cullet being used and the implication on the quality of the final product was pursued.

In a second step, in order to improve the tool developed by the Data Science and Analytics department to improve energy consumption, a possible improvement in the glass quality prediction processes was studied using statistical and machine learning tools. Access was obtained to the inspection machine's data encompassing the entirety of production to create a better quality indicator that did not rely on sampling. By cross-referencing this data with the data from the batch and furnace team, and furnace operation data previously gathered in the scope of the energy optimizing tool, it was possible to substantially improve the defects' prediction by unmelted batch material inclusions.

Keywords: Glass containers, Supervised learning, Recycled glass, Quality

Agradecimentos

Gostaria de começar esta secção com o meu profundo agradecimento a todos os pessoas que possam ter auxiliado direta ou indiretamente na elaboração desta dissertação.

Deixo o meu agradecimento, em particular, ao meu supervisor **Eng. Paulo Gomes** pelo seu acompanhamento e integração na equipa. Gostaria também de deixar o meu sincero obrigado ao **Daniel Baptista**, ao **Gustavo Rosa**, ao **João Sampaio**, ao **João Nuno Sousa** e á **Mariana Barrias**, por me terem acompanhado nesta primeira experiência num contexto profissional. À **BA Glass** deixo o sincero agradecimento por me ter possibilitado esta oportunidade, fornecendo todas as condições necessárias para a execução deste estudo.

Agradeço também ao **Prof. Luís Guimarães** pelo apoio prestado nestes ultimos quatro meses. Deixo também um obrigado a todos os docentes da Faculdade de Engenharia da Universidade Porto, em especial à **Célia Couto**, à **Fátima Magalhães** e à **Susana Dias**.

À minha família reconheço o apoio dado ao longo dos anos, pois sem o seu apoio incondicional não teria chegado onde estou. Aos meus amigos um obrigado especial por moldarem a pessoa que sou hoje, através dos bons e dos menos bons momentos, pois com eles certamente nunca existirão maus.

Francisco Rodrigues

"Because in a split second, it's gone."

Ayrton Senna

Contents

List of Figures

List of Tables

Abreviaturas e Símbolos

AGV	Automatic Guided Vehicle
AUC	Area Under the Curve
EDA	Exploratory Data Analysis
ETL	Extract, Transform, Load
IIOT	Industrial Internet of Things
IPO	Inital Public Offer
kWh	Kilowatt hour
LNG	Liquefied Natural Gas
ML	Machine Learning
NNPB	Narrow Neck Press and Blow
PLC	Programmable Logic Controller
QR	Quick Response
ROC	Receiver Operating Characteristic
UI	User Interface

Chapter 1

Introduction

This study is part of a curricular dissertation project in a business environment. The work developed in this study aims to tackle some operational challenges that a glass container company faces in the current geopolitical and environmental context. This first chapter outlines key points such as the company's history, the motivation for this study, the objectives we pretend to achieve, and the methodology followed in its development.

1.1 Motivation

Glass containers demand is growing every year as a consequence of the drive to increase the use of sustainable packaging. The glass industry is particularly dependent on energy costs, primarily associated with the glass melting furnace that is key to obtain our final product. At the end of 2021, energy price fluctuations consumed a considerable chunk of the potential profits of BA Glass, and following the crisis in Western Europe, with increasing volatility in European energy markets, the problem is only expected to worsen.

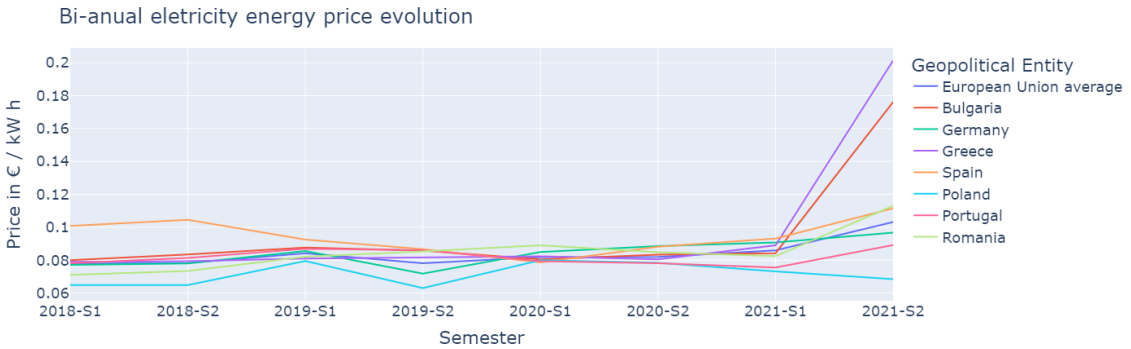


Figure 1.1: Evolution of electricity prices for non-household consumers in €/kWh

Source: NRG - EUROSTAT

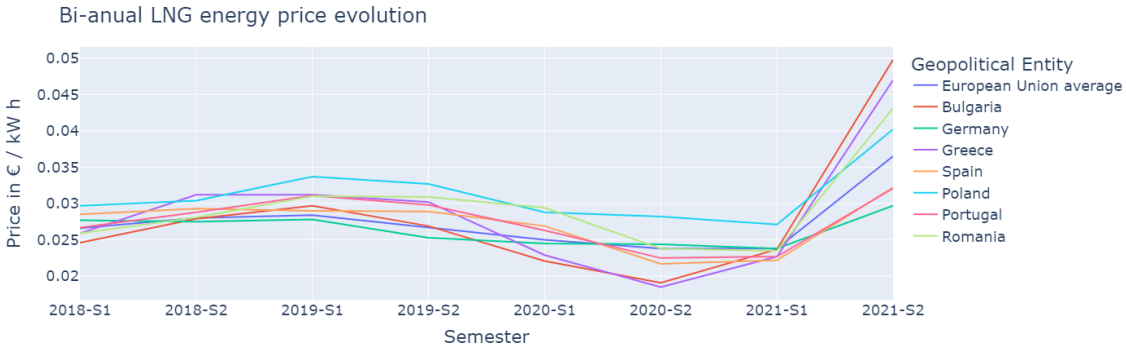


Figure 1.2: Evolution of LNG prices for non-household consumers in €/kWh

Source: NRG - EUROSTAT

In order two circumvent the price increases evidenced by 1.1 and 1.2, a specific focus is being put on an increase in the use of cullet and energy saving in each furnace. These measures are expected to lead to an increase in quality problems, and our aim is to improve current quality predicting methods in order to comprehend this problem better and improve current energy-saving techniques.

1.2 BA Glass

Barbosa e Almeida was founded in 1912 by Raul da Silva Barbosa and Domingos de Almeida and has been one of the main Portuguese glass container company ever since. In 1947, automated technology was firstly introduced, a breakthrough in the then Portugal's manufacturing paradigm. Its current headquarters in Avintes, Vila Nova de Gaia, was constructed in 1969 and has been since the center of the company's operations.

During periods of increasingly high innovation and economic decisions in the '80s and '90s, the acquisition of 94,5 % of CIVE - Companhia Industrial Vidreira, SA' in Marinha Grande, an IPO on VILESA - Vidriera Leonesa, SA, progressive ownership that accounted for 99% of shares and the construction of a factory in Villafranca de los Barros lead to an increase in the furnace count, from the two furnaces in Avintes to eight spread over the Iberian Peninsula.

In the 2000s, the company took on a new name, BA Vidro, following a management buyout and ceased to be listed on the stock exchange. This buyout marks a new era in the company's history because, under successful investor and FEUP alumni Carlos Moreira da Silva, BA is able to acquire eight more plants in Portugal(1), Poland(2), Germany(1), Bulgaria(2), Greece(1) and Romania(1). Since the last acquisition in 2017, the group has been comprised of 12 plants spread across seven countries 1.3.

In 2021 BA Glass employed over 3900 employees and distributed glass containers to more than 70 countries all over the world.[2]

Figure 1.3: Geographic location of BA Glass plants

Source: BA Glass 2020 Anual Report

1.3 Objectives

This study will focus on studying ways to improve how quality is depicted in the current models that are being used in-house. Two obstacles came up when discussing the major shortcomings of the quality models. The first is that, currently, the only raw material being monitored is the cullet as a whole, and although this is one of the main sources of non-conformities in the glass containers, its evaluation only grasps a part of the problem. Secondly, since the quality indicators are obtained via sampling, they are not be able to reflect the entirety of the production and will also hinder the training of the models because other variables which are not always captured, may affect the sampling rate and procedure.

In order to address both of these issues, we will pursue two distinct paths. On one hand, by looking at the raw materials that enter the melting furnace, we aim to correlate certain defects with the use of certain materials. On the other hand, by obtaining the data from the quality inspection machines, we will start to rely on data that encapsulates 100% of the production and hopefully will produce better predictive models for quality purposes. This data will also be helpful for everyday process control at the factory level.

1.4 Data Science Methodology

In this work, we will follow the Cross-industry standard process for data mining methodology proposed in [3]. This standard model has become the de-facto standard for data mining, gaining widespread use by the emerging data mining community.

This approach breaks down a machine learning problem in six key steps, as seen in the figure 1.4.

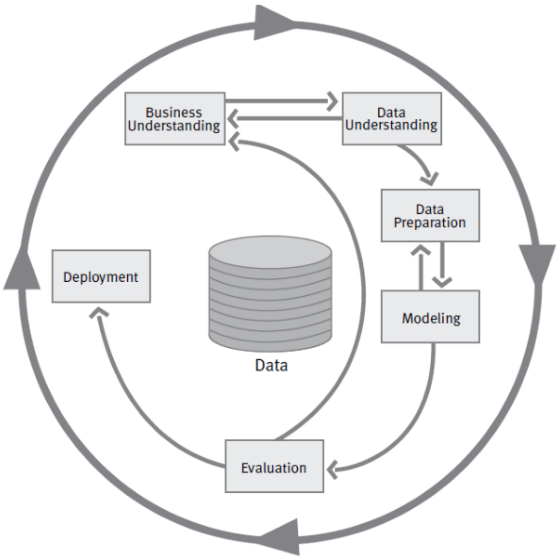


Figure 1.4: Phases of the CRISP-DM reference model

Source: CRISP-DM 1.0 [3]

1. **Business Understanding** - in this stage, the objectives are defined through a thorough understanding of the current business environment and how the data science project will build added value to the organization. When describing these goals, external risk factors and costs must be considered;

2. **Data Understanding** - During this phase, the data is initially collected and interpreted, and its condition is evaluated. A key consideration in this level is to question the data as much as possible with the aim of validating it. When backed with graphical visualizations, this validation can unearth questions such as: is the data manually or automatically collected, is the database in its 'purest' form, meaning if the data has gone through a previous treatment process, and other queries that may arise;

3. **Data Preparation** - this stage is perhaps the most vital step because of the well-known expression in the Data Science community, 'Garbage In, Garbage Out'. This quintessential saying alludes to the fact that if the data quality is bad, even if the most intricate machine learning model is used, the outcome will also be unsatisfactory. Data will be selected, treated, combined, derived, and even formatted in this phase;

4. **Modeling** - at this level, everything regarding the model development and evaluation will occur. It begins by choosing the best models that may suit our data. After the models' selection is completed, the following step must deal with concepts like training, tuning, and testing the model. The choice of algorithm will come down to their performance faced with some previously chosen metrics;

5. **Evaluation** - in this stage, the developer has a functioning machine model with satisfactory performance. In order to validate it for deployment, the business goals are evaluated, and

the current process is assessed for possible changes that may come from the project. The business use of our solution must be defined by the end of this step;

6. **Deployment** - in the final moment of this process for data mining, the project solution will be deployed and may require some training for the end-user. Three different landmarks mark this step the deployment plan, the maintenance plan, and the final report.

Although there is a sequence associated with these steps, moving back and forth between stages is customary and even encouraged. Often one may gain some insights along the way that require changes to what was previously done.

1.5 Dissertation Structure

This dissertation was structured in seven chapters to provide an in-depth description of the project's methods and overall context.

Chapter 1, **Introduction**, gives an overview of the company history, the motivation, and the objectives for the development work, as well as the approach methodology;

Chapter 2, **State of the Art**, offers an overview of the techniques, methodologies, and algorithms used throughout the work, both in the statistical and machine learning domain;

Chapter 3, **Business Understanding**, outlines the entire process taking place in the Avintes plant, presenting detailed explanations of the several stages and the scientific reasoning behind them.

Chapter 4, **Data Understanding**, introduces the data that will be addressed in our study. The description and acquisition of the real-life counterparts are described, and an exploratory data analysis is performed to understand the distributions and values of our data set attributes and possibly identify issues regarding its quality.

Chapter 5, **Data Preparation**, encapsulates the treatment of the data and its quality issues, the creation of new attributes, the deletion of specific records, time shift analysis, and aggregation, among other techniques used to prepare data.

Chapter 6, **Modeling and Evaluation**, describes the model election as well as its training, tuning, and testing. Following this, it offers a detailed view into assessing if the business goals for the project have been achieved and their use for the organization.

Chapter 7, **Conclusion and Future Work**, concludes the dissertation by offering a quick summary of the established work and the subsequent measures to pursue this work's development further.

Chapter 2

State of the Art

Data Science has witnessed an exponential rise in the last decade and positioned itself as a must-have in a competitive business market in order to enable data-driven decision-making. This field study is ever-evolving, and this chapter will look in detail at some of the fundamental data manipulation techniques and machine learning algorithms that have been used throughout the work.

2.1 Data Mining

Data mining was defined in [4] as "the application of specific algorithms for extracting patterns from data. It involves fitting models to, or determining patterns from, observed data".
Due to increased computational power and the growing data collection, this type of approach has gained widespread use in every sector, be it Farming, Manufacturing, Retail, or even Financial services.

This work was developed in an industrial environment, so primarily sensory data will mainly be used, with both automatical and manual acquisition methods. Data will be either in numerical, categorical, or DateTime form.

Data mining encapsulates five types of tasks :

- Association Rule Learning
- **Clustering**
- Classification
- **Regression**
- Summarization

We will pursue the tasks highlighted in bold in our work.

2.2 Bias vs Variance Trade-off

The bias-variance trade-off is a machine learning fundamental that characterizes model prediction. While trying to minimize both these aspects, one may find they are negatively correlated, so a trade-off must be found where both are acceptable. Firstly, both concepts are defined to explain this fundamental principle better.

Bias is the discrepancy between the model's average forecast and the actual value it is attempting to predict. High bias models oversimplify the model and pay very little attention to the training data. When applied to training and test data invariably results in substantial error. This is called underfitting the model.

Variance is the variability of a model's forecast for a particular data point or value, which indicates how widely distributed our data is. A model with a high variance pays close attention to the training data and does not generalize to new data. As a result, these models have significant error rates on test data yet perform exceptionally well on training data. This situation is commonly referred to as overfitting the model.

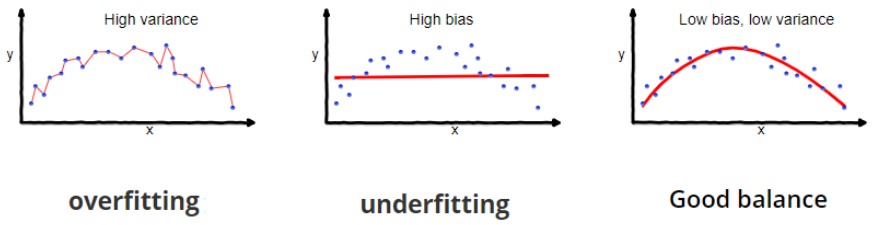


Figure 2.1: Examples of model fitting

Source: in Towards Data Science - Seema Singh

From the figure 2.1, we can graphically see both concepts that were introduced earlier and a third example that illustrates the goal when tunning the models. A model that accurately depicts the real-life relationship between the predictors and the target variable and can reliably generate useful predictions. Most algorithms have their way of tackling this issue and allow the user to tune their parameters in order to achieve a satisfactory model. This tuning principle will be later addressed in this work.

2.3 Black vs Glass Models

This dichotomy is typically used to distinguish models easily understood by humans and others that may be too complex even for a knowledgeable user to grasp their output completely. Glass box or Transparent models earn this name due to their intrinsic simplicity and relatability to human experience. On the other hand, Black box models may have a large number of coefficients and

increased complexity so that other techniques have to be used to interpret the output of these models.

2.3.1 Shapley Values

This mathematical metric was developed by Lloyd Shapley, from which he was awarded the Nobel Prize in Economics in 2012, in the scope of game theory[1], and it aims to answer the following premise :

Given that we have a coalition **C** that collaborates to produce a value **V**, how much did each individual member contribute to the final value?

To explain this idea take the following example:

Imagine a Coalition $C_{1,2,3,4}$ of four players that produces a Value $V_{1,2,3,4}$;
we want to know how much **Player 1** contributes to the final value.

To find a fair answer to this question, we first look at a sample of a coalition that contains Player 1, then look at another without it. The difference in the values produced by both coalitions gives us the marginal contribution of Player 1 to the given coalition that does not include him. We then calculate all the marginal contributions for every possible scenario where our player is not present.

The mean of all the marginal contributions gives us the Shapley Value for Player 1. In other words, it represents the average amount of contribution a player gives to the coalition value.

2.3.2 SHAP

SHAP stands for SHapley Additive exPlanations presented in [6], it brings the previous idea of Shapley Values to the machine learning domain. The authors based it in around three principles :

- **Local Accuracy:** When approximating the original model f for a specific input x, local accuracy requires the explanation model to at least match the output of f for the simplified input x' (which corresponds to the original input x).
- **Missingness:** If the simplified inputs represent feature presence, then missingness requires features missing in the original input to have no impact.
- **Consistency:** it states that if a model changes so that some simplified input's contribution increases or stays the same regardless of the other inputs, that input's attribution should not decrease.

This explanatory model is model agnostic, meaning it works with any model, even the black box kind.

[1] Game theory [5] is the mathematical field of study defined by the following question:

n players $S_1, S_2, ..., S_n$ are playing a given game of strategy A.

How must one of the participants, S_m, play in order to achieve the most advantageous result?

2.4 Supervised Learning

Supervised learning is an ML subgroup of problems where the label of the target variable is known. It predicts the target variable by computing a function from the selected variables from the data set. This type of problem splits itself into two categories Classification and Regression.

2.4.1 Data Split

A supervised model splits itself into two main parts, the training of the model itself and then the testing of the previously trained model. In order to reduce bias, firstly, the data is split into train and test sets. In that way, we can accurately access our model performance with some data that the model has not previously seen, therefore mimicking the future use that our model will have and assessing its performance.

2.4.2 Regression

In a Regression problem, the aim is to predict a numeric variable from the other variables which serve have as inputs in this situation. The models implemented in the Furnace Optimizer and later defined in 3.6, fall into this category.

To access the performance of a regression model, typically, the following metrics are considered: Coefficient of determination 2.1, Mean Squared Error 2.2, Root Mean Squared Error 2.3 and Mean Absolute Error 2.4

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2} \tag{2.1}$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2 \tag{2.2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2}{n}} \tag{2.3}$$

$$MAE = \frac{\sum_{i=1}^{n} |y_i - f(x_i)|}{n} \tag{2.4}$$

2.4.3 Classification

Contrary to the previous case, in this type of problem, the variable that is being predicted may have two (binary) or more (multi-class) levels. One must look at the so-called confusion matrix to assess the performance in this type of case.

		Predicted	
		Yes	No
Actual	Yes	*TP*	*FN*
	No	*FP*	*TN*

Table 2.1: Confusion Matrix for a Binary Classification problem

- *TP* - True Positives
- *FP* - False Positives
- *TN* - True Negatives
- *FN* - False Negatives

From this matrix, the following common metrics can be computed: Accuracy in 2.5, Precison in 2.6 and Recall in 2.7

$$\frac{TP+TN}{TP+FP+FN+TN} \tag{2.5}$$

$$\frac{TP}{TP+FP} \tag{2.6}$$

$$\frac{TP}{TP+FN} \tag{2.7}$$

And from the previously computed metrics a popular metric, the *F1 Score*, can be obtained in 2.8, by using some of the previously calculated metrics.

$$2 * \frac{precision * recall}{precision + recall} \tag{2.8}$$

Another popular metric is derived from the receiver operating characteristic (ROC), and it represents the area under this curve (AUC). This curve plots $(1 - specificity)$ against *sensitivity*, which can be calculated using 2.9 and 2.10, respectively.

$$\begin{aligned}(1 - Specificity) &= FPR(FalsePositiveRate) \\ &= \frac{FP}{TN+FP}\end{aligned} \tag{2.9}$$

$$\begin{aligned}Sensitivity = TPR(TruePositiveRate) &= Recall \\ &= \frac{TP}{TP+FN}\end{aligned} \tag{2.10}$$

2.4.4 Ordinary Least Squares

The ordinary least squares regression (OLS) is perhaps the most famous approach to a regression problem. This method of statistical analysis estimates the relationship between one or more independent variables and a dependent variable. It achieves this result by minimizing the sum of squares on the difference between the observed and predicted values of the dependent variable, which is configured as a straight line. The equation 2.11 characterizes this method.

$$f(x) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{2.11}$$

β explains the initially unknown coefficients. Multiple linear regression models with $p > 1$ and more than one input variable are defined as linear models.

$$RSS = \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 + \sum_{j=1}^{p} X_j \beta_j\right)^2 \tag{2.12}$$

By minimizing the equation 2.12 referred to as the the residual sum of squares, our algorithm will tune its beta values. Here, $y_i - f(x_i)$ describes the residuals, β_0 the estimate of the intercept term, and β_j the estimate of the slope parameter.[7]. This method is sometimes preferred over other more complicated models since it's of the glass box kind, making it easy to understand and extrapolate tangible conclusions.

2.4.5 Decision Tree Methods

A decision tree follows a tree-like structure in which it recursively splits the data in each node into branches that better suit our data characteristics. Leaf nodes will be the final nodes of the trees, and it's where our prediction will occur.

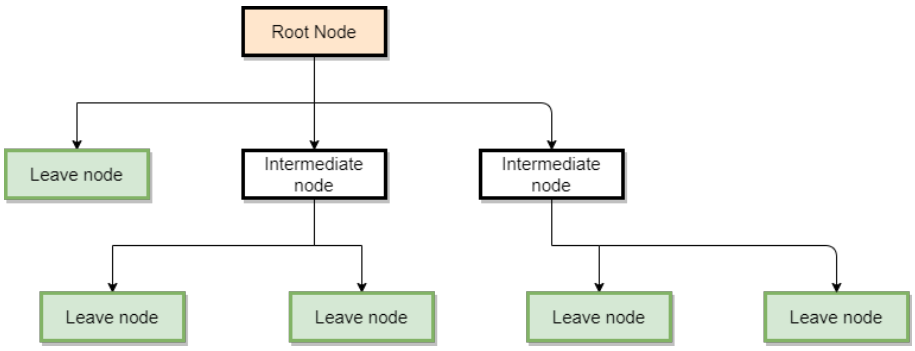


Figure 2.2: Example of a basic decision tree

The term Classification and Regression trees was coined in [8] and defines the two types of outcomes we tend to expect from our regression tree. If the target value is binary, we will use a Classification tree; on the other hand, if the value we want to predict is a numeric one, we will use a Regression tree. As previously stated in 2.1, we aim to try and predict continuous values, so we'll use latter type.

Several decision tree learning methods have been proposed through the years. Early implementation such as ID3 [9] and its evolution C4.3 [10] and the CART [8] have remained relevant throughout the years, suffering minor optimizations and being paired with ensemble methods that have increased their performance and robustness. Ensemble methods combine the predictions of numerous base estimators constructed using a specific learning process to increase generalizability/robustness over a single estimator. These approaches are distinguished into two categories:

The first, Boosting Methods combine several weak learners to create a strong learner. In an analog manner to everyday interactions, it can be described as the wisdom of the crowds outweighing that of an individual expert. With this, we tackle the inherent problem with every machine learning algorithm that has been evidenced in 2.2. Several approaches to two this problem have been developed being, some of the most notable applications: ADABoost [11], a breakthrough application that earned the prestigious Gödel Prize, XGBoost [12], which gained popularity for continuously achieving stellar results in Kaggle competitions and LightGBM [13], which with it's much faster runtime, and comparable accuracy to the XGBoost is gaining widespread adoption among the Data Science Community [14].

The second one, Bagging Methods (also denoted Bootstrap aggregation) creates several subsets of training data with some randomly placed replacements. Each of these subsets is then used to train its decision tree, and then the final tree is composed of an average prediction of all the generated trees. This aims to improve robustness and reduce the variance of our decision tree.A widely used implementation of this method is the Random Forest algorithms proposed in [15].

2.4.6 SVM

Support vector machines are a type of supervised machine learning algorithm proposed in [16] that aims to tackle classification problems by calculating a hyperplane[2] that will create the split between the category levels being classified. This algorithm can be subdivided into two different categories.

The first type is characterized by using a linear classifier to generate the hyperplane. This type of classifier can then be divided depending on the type of margin that we use to define the hyperplane. The margin is a key concept in the SVM algorithm, and it represents the shortest distance between the observation and the decision threshold. The choice of classifier can be simplified by answering the following question - *'Can our data be linearly separatable ?'* If it can be, a **Hard Margin** that doesn't allow for misclassification will be used. If not, a **Soft Margin** will be chosen to allow for misclassification. This error will need to be minimized through the pairing with a loss function, typically the hinge loss.

The second category takes advantage of a concept proposed in [18] commonly referred to as the kernel trick. This concept is the main idea behind the modern implementation of SVM, and it

[2] A hyperplane is a concept in geometry. It is a generalization of the plane into a different number of dimensions. A hyperplane of an n-dimensional space is a flat subset with dimension. By its nature, it separates the space into two half spaces.[17]

uses a kernel function to find support vector classifiers in higher dimensions than the original data.

Current implementations of this algorithm, such as [19], support the following kernel functions:

- Radial Basis Function - similar to the gaussian distribution, but with a radial basis method to improve the transformation, and can be mathematically represented by 2.13
- Polynomial Kernel - transform the original variables via the use of polynomials, and it's defined by 2.14
- Sigmoid Kernel - similar to the sigmoid function used in logistic regression, ilustrated in 2.15

$$K(x,y) = exp\Big(-\frac{||x-y||^2}{2\sigma^2}\Big) \tag{2.13}$$

$$K(x,y) = \tanh(\gamma \cdot x^T y + r)^d, \gamma > 0 \tag{2.14}$$

$$K(x,y) = \tanh(\gamma \cdot x^T y + r) \tag{2.15}$$

2.4.6.1 SVR

When presented with a regression problem, a derivation of the original algorithm described above-named Support Vector Regression, will be used. This method introduces an error parameter ε that represents the acceptable error allowed in the current implementation. This error will be minimized using the l2-norm(also known as euclidean norm) of the coefficient vector.

Another parameter will need to be established for the cases where some of our data points don't fall in the previously defined error margin. The deviation from the margin, represented by the letter ξ, will be minimized via the creation of a new hyper parameter C. Both of these hyper-parameters will be tuned using the methods defined in 2.5.

2.5 Hyperparameter Optimization

Machine learning models typically depend on a varying number of inputs that may span from specific ranges. The problem faced when using this type of model is searching for the optimal combination of parameters for the data set. These parameters which define the model architecture are referred to as hyperparameters, and the discovery of an optimal set is denominated as Hyperparameter optimization, also commonly referred to as tuning. It is defined by the following expression:

> A learning algorithm produces f through the optimization of a training criterion concerning a set of parameters θ, a way to choose λ to minimize generalization error $E_{x\ G_x}[L(x;A_\lambda(X^{(train)}))]$.[20]

Manual tuning, as the name implies, is testing specific hyperparameters by manually experimenting with different values. It relies heavily on user expertise and deep knowledge of the tuned model, usually becoming a cumbersome and time-consuming task.

Several algorithms have been proposed to address this shortcoming, and in this section, some of the most popular methods will be introduced.

One of the most popular approaches is called Grid-Search. It is an exhaustive method that computes a grid of all possible combinations defined by the user. This process can be very computationally costly as it calculates the fitness function for each combination to find the best possible one, and it also acts under the premise that the parameters are equally distributed.

In 2013, another method was proposed to address the latter problem regarding grid-search in [20], called Random Search. All the name implies it randomly generates value from the user-defined ranges and evaluates the fitness for each set. The main advantage of this method versus Grid-Search is that it does not assume an equal parameter distribution.

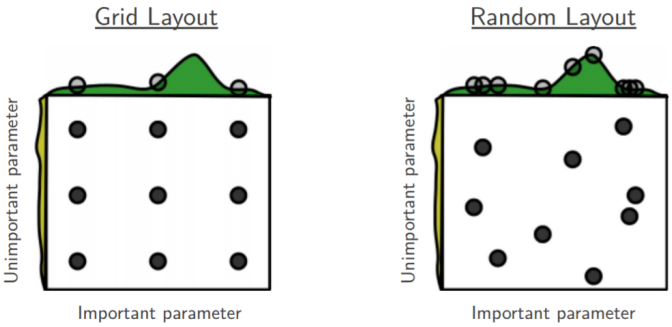


Figure 2.3: Difference between Grid Search and Random Search

Source: in Towards Data Science - Peter Worcester

As can be observed in 2.3, this method randomly samples the solution space for possible parameters, and by doing so, it may find better hyperparameters than grid search or not.

Sequential model-based optimization (SMBO) is an algorithm that evaluates the fitness function using a computationally cheaper surrogate. It then optimizes this surrogate until a point x* becomes the proposal to be assessed in the complete fitness function. These types of algorithms typically utilize the Expected Improvement [21] optimizaton principal.

Tree-structured Parzen Estimator or TPE, for short, is an algorithm based on Bayesian hyperparameter optimization that initially works just like the Random Search sampling method. The main difference is that it records the previously tested parameters, and the following suggestion will take them into account to improve the trials gradually.

2.6 Unsupervised Models

Unsupervised learning is a machine learning approach that learns patterns from unlabeled data. This type of learning can be subdivided into Cluster Analysis and Principal Component. Cluster Analysis itself encapsulates two families of algorithms: clustering and anomaly detection.

2.6.1 Clustering

By using the principle of mimicking present in biology, this family of algorithms tries to split data into different clusters through a similarity metric. This approach results in unlabeled datapoint groups that share similiar characteristics. The interpretation of these clusters is one of the main characteristics of an unsupervised technique as it may unveil novel knowledge from our data set following the cluster analysis or provide the user with unexplainable results.

Two principles can be used to describe clusters:

- **Low inter-class similarity** - points in separate clusters are less similar to one another.
- **High-intra-class similarity** - points in the same clusters are more similar to one another.

2.6.2 K-means

K-means is perhaps the most famous approach to an unsupervised clustering method. It creates a number k of clusters, defined by the user, by interactively computing a fictitious mean point for each cluster denominated centroid. During this process, it computes several different centroids for varying groups of data points, attempting to minimize the Euclidean distance of the points to the centroids. The process continues until a certain stop criterion is met, which can vary from an error measurement such as the sum of the squared error, simply a computational threshold such as a maximum number of iterations, or ideally, until no change occurs in the clusters meaning that the optimal clusters have been defined.

2.6.3 DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)[22] is a density-based algorithm that identifies its clusters following the principle that in a data space, a cluster is a region with a high point density that is isolated from other clusters by regions with a low point density. This specific algorithm has two hyperparameters :

- **epsilon** - the distance that will determine if a point belongs or not in a neighborhood
- **minPTS** - the minimum amount of points in a region to be considered a with high point density.

This algorithm randomly selects a point in the data set and checks if there are minPTS in a radius of epsilon to check if it will create a cluster in this region. Following this, it applies the

same logic to all the points in the now-created cluster. If the sequence gets broken, the group is complete and will move on to other non-previously selected points.

This method has some advantages versus the k-means methods we saw before, such as the definition of the number of clusters and the fact that as k-means focus on the mean value, it may cluster very different observations together.

2.6.4 Anomaly detection

Anomaly detection, is a form of cluster analysis that generates two clusters. The first considered regular observations and the second considered outliers or observations that deviate from the other data points. This method can be **univariate** if our target is just one or **multivariate** if we aim to detect outliers among several variables at once. Two principal assumptions are present in the underlying logic behind these algorithms. The first is that anomalies rarely occur in the data, and the second assumes that these anomalies differ significantly from regular observations.

2.6.5 Isolation Forest

Isolation Forest, proposed in [23], is a tree-shaped method that splits the data across its amplitude and identifies outliers based on the fact that these values are few and different. It makes this determination based on an anomaly score metric, which measures the number of splits required to isolate a specific point. It intrinsically reflects the principle that outlier points occur in low-probability areas.

2.6.6 CBLOF

The Cluster-based Local Outlier Factor (CBLOF) calculates the outlier score based on the cluster-based local outlier factor. The distance of each instance computes an anomaly score to its cluster center multiplied by the observations belonging to its cluster. This method relies on the user identifying the percentage of outliers in the data, being a parameter that can be fine-tuned but sometimes becomes a slow process when dealing with big data sets.

2.6.7 Principal component

Principal component analysis is a method used to reduce the dimensionality of our data and is often used in pair with supervised learning in order to treat large data sets that may include not-so-relevant variables. This method has the advantage of maintaining much of the variance of the original data set even when the size of the same has reduced significantly. Of all of the algorithms exposed in this section, this is probably the oldest, tracing its root back to the 1903 paper [24] and has seen application in almost every field of study, finding a new application in the ever-growing data science ecosystem, particularly big data applications.

2.7 Recursive Feature Elimination

Recursive Feature Elimination (RFE), initially proposed in [25], is a method used in supervised machine learning to identify the attributes that better influence our target variable. It removes features until a specified value is reached and attributes feature importance to each selected variable. Typically this method is paired with some cross-validation methods in order to find the optimal set of features to keep.

2.8 Python

Python is a high-level interpreted programming language and has become famous for being beginner-friendly and applicability to serious projects. Guido Van Rossum created it in 1991, and it is currently in the third version. It is the language of choice for the Machine Learning and Big Data communities having a wide array of third-party libraries. Pip is the standard package manager used to install and manage libraries. In this section, some of the libraries that were used throughout the work will be introduced.

2.8.1 Pandas

Python Data Analysis Library (*pandas*) was created in 2008, and it has become one of the most popular python libraries currently in distribution. The addition of data storing objects like Series and Dataframes has helped Python triumph over **R** in the battle for the most used language in the Data Science community. Another factor is the implementation of *NumPy*, another popular library for mathematical operations, to perform quick and efficient calculations. Throughout this work, when referring to Dataframes, it is referring to the panda object that resembles a spreadsheet, having rows that can also be defined as entries and columns that can be denominated as attributes.

2.8.2 Plotly

Plotly is an open-source plotting library that supports interactive visualizations. Based on the Plotly **JavaScript** library *plotly.js* it features over 40 unique graphs and web-based visualization that natively support Jupiter Notebooks. Managed by the same entity as the Dash library, which enables the creation of Reactive Web Apps in python, the combination of these tools can be instrumental when creating an open-source visualization platform. Although pandas natively supports the widely used *matplotlib* library, I chose to use Plotly since its interactivity and charismatic plots and the dash integration triumph over a more reliable community of *matplotlib*.

2.8.3 Scikit-learn

The scikit-learn library is one of the most valuable assets in an ML Python toolkit. It collects some of the most popular classification, regression, and clustering models, as well as preprocessing,

dimensionality reduction, and evaluation tools. With its first release back in 2012[26], it has rapidly evolved into one of the most accessed ML libraries on Github.

2.8.4 Optuna

Optuna[27] is a next-generation tuning framework that combines a defined by run API with efficient searching and pruning strategies allowing users to construct the search space dynamically. It has an implementation of GridSearch, Random Search, and Tree-structured Parzen Estimator. Being the latter, the one who achieves the best results, it takes advantage of the intrinsic memory characteristic of the algorithm to allow the user to store and resume the tuning trials.

2.9 Summary

Following this extensive review of some of the data scientist's key concepts and tools, special consideration should be given to some of the previously exposed methods since they will be critical in the following work. This is the case with the Boosting methods for decision trees, Support Vector regression, the Optuna tuning framework, and the SHAP library.

Chapter 3

Business Understanding

This chapter gives a brief introduction to why glass was adopted as a container and how it has seen a resurgence in recent times. Following this, a deep dive will be taken into the modern glass-making process and its key steps. In the end, some of the previously developed in-house tools that fit this dissertation scope are presented.

3.1 Glass

3.1.1 Glass definition

The first question we have to ask ourselves is, what is glass as we know it. From [28] we can obtain the definition that "Glass is an inorganic solid material that is usually transparent or translucent as well as hard, brittle, and impervious to the natural elements."

All glasses found to date share two common characteristics. First, no glass has a long range, periodic atomic arrangement. And even more importantly, every glass exhibits time-dependent glass transformation behavior. This behavior occurs over a temperature range known as the glass transformation region. A glass can thus be defined as “an amorphous solid completely lacking in long range, periodic atomic structure, and exhibiting a region of glass transformation behavior.” Any material, inorganic, organic, or metallic, formed by any technique, which exhibits glass transformation behavior is a glass.[29]

3.1.2 Glass packaging

Glass is one of the oldest packaging types and one of the most sustainable, based on the fact that this material is almost 100 % recyclable. Besides this important quality in an increasingly eco-conscious world, there are also other properties of glass that reaffirm it as an excellent type of packaging, such as :

- Substantial chemical inertness - a good inert capacity concerning food and liquids, although reactivity with sodium and other ions must be taken into account for applications in the pharmaceutical industry.

- Impermeability - this quality makes it especially useful for storing *liquids and foods* susceptible to oxidation when exposed to oxygen or moisture.

- Transparency - this characteristic is beneficial when it is necessary to expose the contents of the packaging.

The rigidity of glass can be considered one of its shortcomings. When first formed, glass has a much higher rigidity than steel, but as it cools, microstructures emerge, contributing to its weakening. The appearance of these sub-microscopic structures is controlled by annealing, but the glass becomes fragile and brittle even with this process. On the contrary, it retains some of the characteristics such as resistance to pressure and vertical forces that allow it to be used to store carbonated beverages and the ability of bottles to be stacked vertically, respectively. Thanks to this fragility associated with our material, bottles are typically 100% inspected for minor defects that can compromise the quality of the glass and make it even more brittle. These defects will be addressed later in 3.5.

Its high density (2.5g/cc) is also one of its primary defects because, compared to plastic, it has a crippling weight, associated with increases in the cost of transport. This transportation limitation may be seen as an environmental drawback, but the fact that this material can be almost 100% recycled, as evidenced in the following sections, outweighs this negative aspect.

In an evermore environmental conscious world that strides to eliminate single-use plastics, glass containers consumption has been increasing in recent years

3.1.3 Glass composition

Much of modern glass production focuses on a specific type of glass called **soda-lime-silica**. This type of glass accounts for about 90 % of manufactured glass[30]. Its formula arises from a compromise between the good vitreous properties of pure Silica - SiO_2 and its high melting point. In order to lower this melting point, it is necessary to add Na_2CO_3 - Sodium Carbonate (colloquially known as soda ash). The addition of this ingredient just by itself will create the so-called water glass derived from the creation of sodium silicate, which leads to the deterioration of the properties of glass necessary for its use as a container, namely its characteristic of insolubility. To circumvent this effect, the third main ingredient of this type of glass is added, Lime, which will add Calcium to the composition of the glass (through the presence of CaO), thus making the glass insoluble again.

Table 3.1: Typical composition and properties of soda lime glass [1]

Composition / properties	
Chemical composition (% by weight) :	
Silica (SiO_2)	70-74
Sodium oxide (Na_2O)	12-16
Calcium oxide (CaO)	5-11
Aluminum oxide (Al_2O_3)	1-4
Magnesium oxide (MgO)	1-3
Potassium oxide (K_2O)	≈ 0.3
Sulphur trioxide (SO_3)	≈ 0.2
Ferric oxide (Fe_2O_3)	≈ 0.04
Titanium oxide (TiO_2)	≈ 0.01
Properties :	
Glass transition temperature, T_g(ºC)	573
Coefficient of thermal expansion (ppm/K, 100-300ºC)	9
Density at 20ºC (g cm^{-3})	2.52
Heat capacity at 20ºC (kJ kg^{-1} K^{-1})	9

3.2 Cullet

A high percentage of recycled glass is used in the mixture, which in the industry is referred to as cullet. The cullet rate measures the amount of cullet used versus the amount of molten glass pulled from the oven, reffered to as pull. Apart from being ecologically sustainable and less expensive than the raw material, this recycled glass will have a better yield since the temperature required for its fusion will be lower, and it will present lower melting losses when compared to a furnace of pure raw material. These factors contribute to the fact that adding 10% cullet to the melting mix leads to a 2-3 % reduction in energy consumption.

This energy reduction factor in the current situation described in 1.1, as lead to an increase in this material consumption which varies across the plants depending its availability and quality. In some plants, as is the case of the Gerdelengen plant, there are productions where cullet rates of 90% are achieved, while plants in Bulgaria typically have lower cullet rates of about 35%. The company average sits between 40% to 60% of cullet rates.

3.3 Modern Glass-making Process

Nowadays, glass production is subdivided into three main stages. This section will detail each of these stages with a specific focus on questions regarding possible defects and cullet usage. Although every production line at the Avintes Plant, follows this distribution, the focus will be on the system currently in place at AV5 in greater detail because the following work was developed there.

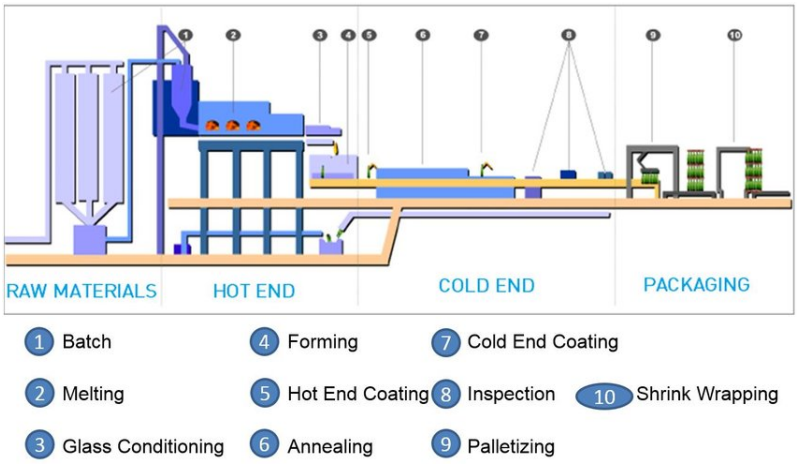


Figure 3.1: The main stages of the modern glass-making process

Source: [31]

3.3.1 Batch House

All the raw materials will be stored, weighed, mixed, and then fed into the oven in the batch house. First, raw materials arrive via trucks and are stored in their respective silos. Cullet either comes from the factory or from Norcasco, a subsidiary of BA, with a cullet treatment station adjacent to the factory or from the factory itself in several different forms. The materials are split into eight main sections relating to the eight different scales and their difference in ranges. There are scales that weigh up to 2 tons and as little as 5kg, this can be seen in A.1. This amplitude comes from the fact that, for example, in one day, we may use 100 tons of sand but only 100kg of coal. There are two types of scales:

- **Hopper style** - The hopper is held up in the air with four sensors that weigh the raw materials. Once the desired weight has been achieved, the shute opens, and it unloads to a conveyor belt that leads to the mixer.
- **Belt style** - This type of scale is only currently being used with the cullet. Materials are unloaded to a conveyor belt that has built-in weight sensors.

Then all materials are mixed in one of two mixers and then stored in one of the two silos that exist right by the furnace. In the case of the AV5, we have silo numbers 51 and 52, corresponding to the right and left silos. Ideally, the mix does not stay too long in the furnace silos, but it is managed daily by the factory workers, and it's subjected to a plethora of factors such as machine maintenance, shift hours, among others. This batching process is controlled by an information system called Glass Soft, which features several SCADA interfaces and allows for preset recipes to be applied. Corrections are constantly being done to the recipe, taking into account several parameters, such as the color deviations, defects, and other factors.

3.3.2 *Hot End*

This section's main components are the Furnace, Forehearth, IS machine, and Lehr.

3.3.2.1 *Furnace*

The AV5 furnace is a Regenerative gas furnace with electrical boosting, meaning that it uses natural gas as it is the primary source of heat and relies on electric boosting as its secondary source of heat. It is regenerative, meaning that it works on a cycle, typically half an hour, in which a change occurs between using the regenerative stacks as air outputs or inputs. This property leads to substantial energy saving and a decrease in exhaust degradation since when an inversion happens, the air being fed into the furnace will be heated, leading to less energy consumption and consequently cooling the regenerative stack.

The mix is continually fed into the furnace, where it is melted into glass, using typical temperatures of 1500º-1600º C. The furnace itself can be subdivided into various sections, highlighted in the figures 3.2 and 3.3:

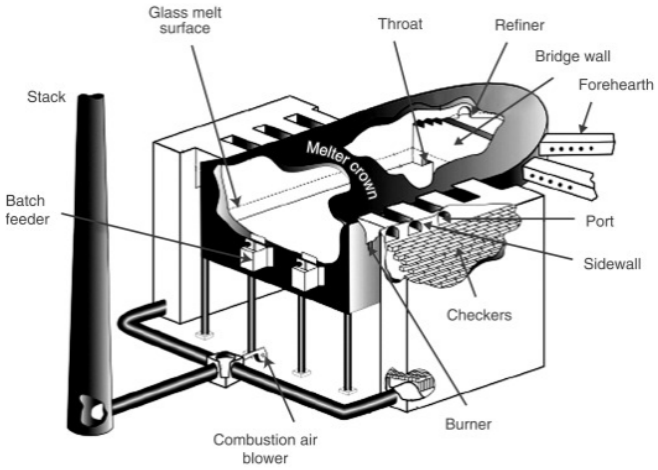


Figure 3.2: Top-view of glass melting furnace

Source: U.S. Department of Energy (2002). "Energy and Environmental Profile of the U.S. Glass Industry."

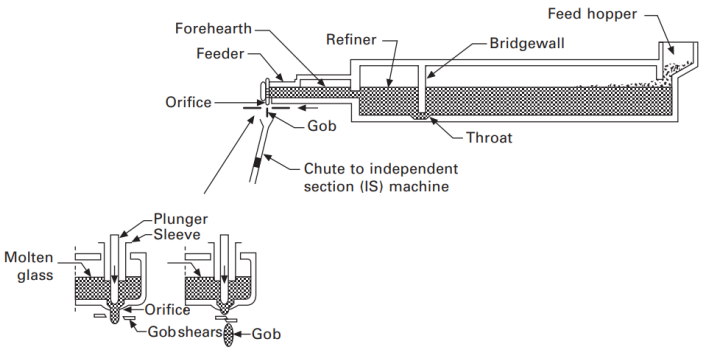


Figure 3.3: Side-view of glass melting furnace

3.3.2.2 *Forehearth*

The forehearth lowers the temperature and conditions the glass, producing a stable glass to be delivered to the forming process. At the end of the forehearth the feeder uses gravity to feed molten glass to the IS Machine. Molten glass flows into orifices with specific dimensions that depend on how much glass needed for the type of bottle being produced. This specific quantity of molten glass will be pushed and cut by water-cooled mechanical shears becoming what is call a *gob*. Each one of these *gobs* will later become a bottle or a jar, or any other container. **Higher production rates** are achieved by using **double**, **triple**, or even **quadruple** gobs to feed the forming machine.

3.3.2.3 *IS Machine*

The Individual Section Machine 3.4 is where the container forming occurs, and it is a high throughput machine with several sections where the gobs from the forehearth are redirected. It can be hydraulic or electrical.

Depending on the type of bottle we are trying to produce, there are different types of forming methods:

- Blow and blow 3.5;
- Press and blow 3.6;
- Narrow neck press and blow.

All of these processes require two moulds a blank mould and a blow mould. The existence of these two moulds will split the IS machine into blank and blow sides.

Figure 3.4: Triple *gob* IS Machine at the line 53 in the Avintes plant

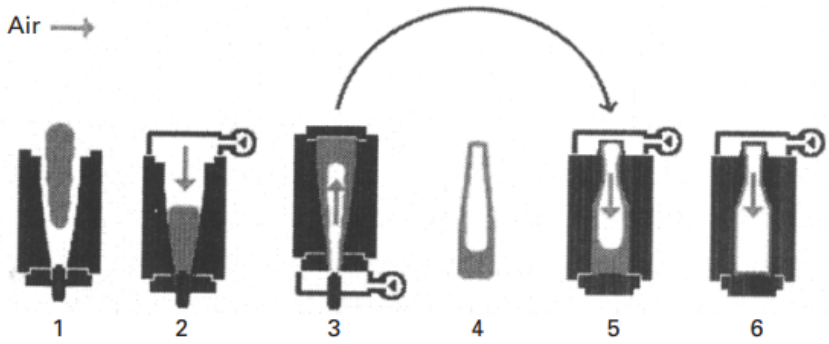


Figure 3.5: Container forming by the **blow and blow process.** 1, gob enters parison mould; 2, settle blow to form finish; 3, counter-blow to complete parison; 4, blank formed; 5, blank transferred to blow mould; 6, final shape blown.

Source: [1]

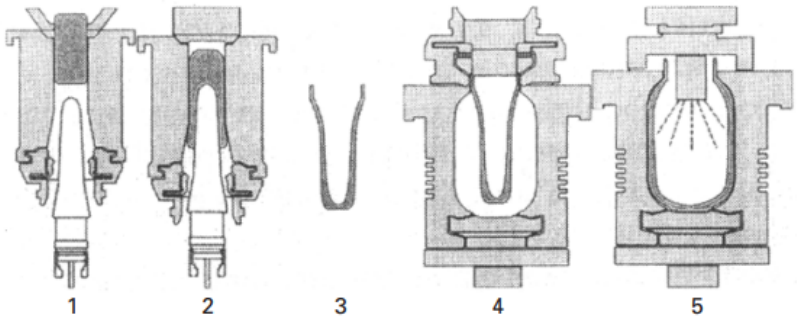


Figure 3.6: Container forming by the **press-and-blow process.** 1, gob drops into parison mould; 2, plunger presses parison; 3, parison completed; 4, parison transferred to blow mould; 5, final shape blown.

Source: [1]

Typically blow and blow methods were used for narrow neck bottles, while press and blow are used for wide-mouthed containers, i.e. jars. In the coming years, NNPB has been replacing blow and blow because by using a metal plunger like the press and blow method, but a much smaller one, it obtains a superior glass distribution enabling the production of lighter containers than the conventional blow and blow, reducing cost and environment impact.

3.3.2.4 *Lehr*

In order to guarantee the container resistance, it has to be cooled gradually. This process, referred to as **annealing**, relieves the glass of surface tensions introduced by the quenching the gob suffers when it is formed into a bottle. In the annealing lehr, the bottles are first reheated and then gradually cooled to have a bottle that does not shatter under a slight temperature change or even spontaneously.

3.3.3 *Cold End*

The cold end section begins at the end of the annealing lehr with an application of lubricant to reduce the damage caused by the bottle rubbing on each other. Then the bootles will be going through several checks in the inspection machines, more specifically :

- Rotational Inspection - controls dimensional, vacuum, out of round, and thickness defects;
- Sidewall inspection (SW) - body visual and dimensional defects;
- Bottom and Finish inspection (BS) - bottom and finish visual defects.

(a) BS in series with a SW inspection machines

(b) Two parallel inspection lines for the same IS

Figure 3.7: An inspection line at the Avintes plant

3.3.4 Packaging

Packaging is the final stage in the process before the products can be shipped to the end consumer. The bottles are automatically stacked into pallets by a robot that also places protective layers in between each level of bottles. Depending on its dimensions, the robot is paired with a PLC that knows the optimal combination to stack the bottles in.

After the pallet is complete, an AGV picks it up and takes it to another robot station. In this station, a protective shrink wrap film that covers the entirety of the pallet is added as protection from possible contaminants that could occur until the delivery to the client is made.

The process ends when the pallet is tagged with the appropriate information, in the form of text, bar and QR codes, and it is pushed to a area where it will await an forklift operator to pick it up for storage.

3.4 Types of Cullet

3.4.1 Internal Cullet

This type of cullet results from production losses at the factory itself and is subdivided by colors, taking into account the color of the glass being produced in the furnace. It typically has very high quality and may only include small amounts of contaminants. The sources of this type of material are :

- Glass stream - when the IS Machines require some maintenance, the flow of glass must not be interrupted. The glass flow is diverted and falls into an underground water tank, where it will be cooled and recycled thereafter.

- Hot end losses - these are rejections that occur immediately after forming by the use of thermal cameras in order to identify possible stresses in the bottle or due to an automatical rejection when moulds are changed or lubrificant is applied to the IS cavities.
- Inspection machine rejections - these are rejections that occur in the cold end via several different machines.
- Rejected pallets - if a specific defect is observed in a particular batch, following a risk assessment, a designated number of pallets are crushed and re-melted.

(a) Internal cullet from glass stream (b) Internal cullet from inspection machine rejections

Figure 3.8: Two distinct sources of internal cullet in Avintes

3.4.2 External Cullet

When it comes to the Avintes Plant, the external cullet comes from Norcasco, a subsidiary of BA Glass, which obtains by treating waste from, for example, glass recycling bins, depicted in 3.9. Usable cullet is separated from the waste in the waste treatment facility.

There are three types of products that derive from this:

- Mix cullet - a mix between mostly ambar and uv glass
- Flint cullet - mainly white or blueish glass that is obtained by the optical separation that occurs.
- Glass powder - a by-product of the effort to remove labels and other organic materials that the waste glass may have. The grinding of the glass pieces against a metal tumbler generates a fine glass powder that can later be separated and used back in the furnace.

Figure 3.9: Garbage waiting to be treated in the Norcasco facility

(a) Flint cullet filtered by Norcasco

(b) Mix cullet filtered by Norcasco

Figure 3.10: Two distinct types of external cullet processed by Norcasco

This split between the two types of cullet occurs because although one may add all types of cullet to the furnace while producing color glass, on the contrary, if producing flint glass, one may only use flint cullet due to its high sensibility to color change.

3.5 Defects

3.5.1 Definition

Defects are non-conformities that show up in the finished product and are detected mostly in the cold end. As previously mentioned, in the glass industry, there is an inspection of the totality of the production. According to [32], the defects we can find in the finished product are of the following type :

Defects	Definition
Blisters	Large bubbles in the glass with individual diameters > 0,8mm
Seeds	Small bubbles in the glass with individual diameters < 0,8mm
Stones	Small pieces of refractory, contaminants, or unmelted batch material

Table 3.2: Most common defects according to the BA TS 99

3.5.2 Cause

When discussing defects in a long production line, as explained in 3.3, a problem emerges when attempting to pinpoint the cause of this non-conformity.

3.5.2.1 Seeds

Seeds have their root cause in the air that gets trapped inside the molten glass either by the air that comes with the mixed batch of ingredients or the release of CO2 from the melting of Sodium and Calcium.

In an effort to reduce this effect, the fining stage process will be greatly influenced by the current Redox state of the melt that can be controlled with the addition of Sulfates and Carbons. This fining process can be chemically explained by the following chemical equation:

$$2SO_3 + C - > 2SO_2 + O_2$$

The gas bubbles of SO2 will rise in the melt and, when combined with the O_2 bubbles, will generate SO_3, which contrary to SO_2, is soluble in the melt.

The CO_2 bubbles will rise and combine with the O_2 bubbles increasing in size and rapidly rising to the surface. Some of the common causes for seeds are :

- Sudden shifts in the redox states - cullet contamination or moisture;
- Inadequate operation of the furnace or the forehearth;
- Reboil - the creation of bubbles from exaggerated heating of the melt in a stage when it has already homogenized and is in the colling stage;
- Excess use of an internal cullet with reductive properties, such as amber;

- Improper batch mixing leading to air pockets.

3.5.2.2 Blisters

This type of defect typically has its root cause in the forehearth or in the forming stage. Although Sulfate Saturation may cause them, the melt will only come close to this when producing the Ambar color, which is not currently in production in the AV5 furnace.

3.5.2.3 Stones

Stones have their cause in materials that are not adequately melted. This material can come from the raw materials if big piles form inside the furnace leading to improper heat transfers or, most commonly, by the contaminants included in the cullet.

The primary contaminants that cause this type of defect are Ceramics and Opals. Due to their higher melting point, they will not homogenize in the melt, will lead to quality issues, and can also be responsible for downtimes in the process due to jammings in the IS Machines.

3.6 Furnace Optimizer

Furnace optimizer is an in-house tool previously developed by the Data & Analytics department at BA to tackle the ever-increasing energy prices as mentioned 1.1. It is an ML Web-App that works as a recommendation system for daily glass furnace operations. This recommendation focus on different models, namely :

- **Energy Model** - gas and electric boosting
- **Temperature Model** - regarding crown temperature
- **Quality Model** - regarding seeds and stones

Historical data is retrieved from multiple sources for these different models and fed into a previously trained model. This model suggests furnace operations parameters such as cullet, gas, and electric boosting use. Currently, this tool has been rolled out to five furnaces :

- **MGB** - Marinha Grande, Portugal
- **AV5** - Avintes, Portugal
- **GA1** - Gardelegen, Germany
- **PV4** - Plovdiv, Bulgaria
- **ATA** - Athens, Greece

The main goals this tool sets out to achieve are a decrease in furnace melting costs, an increase in glass quality consistency, and less dependency on the operator's expertise.

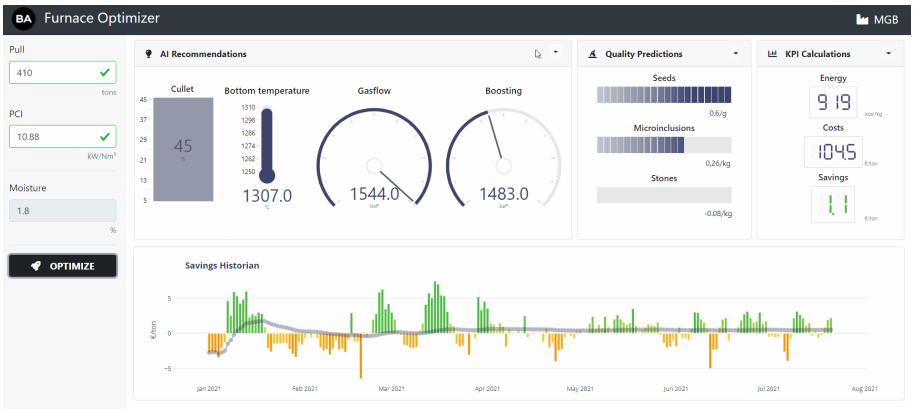


Figure 3.11: Furnace Optimizer UI

3.7 BAMeX

BA Manufacturing eXperience (BAMeX) is a project currently being developed by the Systems & Connectivity department, a sub-branch of the Digitalization and Innovation Teams , in which the Data Science and Analytic department, where this work was developed, is also included.

This platform aims to collect all the data in one single place in order to produce real-time dashboards, visualization, and recommendation tools that aid the day-to-day at the factory. In this effort, a Microsoft Azure Data Lake platform was developed to collect data from the PLCs and Plant Floor Machines databases and store it in the same place. This change will be seen in the 4 by the fact that Glass Soft and Inspection Machine both had two sources of data, one prior to October 2021 and the implementation of this solution and one after that.

3.8 Summary

In this section, a review of the glass composition and everything involved in its fabrication was performed as well as an introduction to some specific details about the Avintes plant and the BA glass company philosophy. An overview of two primary defects that take part in the current quality models and a brief description of the Furnace Optimizer tool and the BAMeX project were carried out.

Chapter 4

Data Understanding

In this section, the attributes from our data sets are explained in order to comprehend what they represent in real life. This includes how they are measured, in what frequency, and how they will be used in the progression of this work.

4.1 Data Presentation

4.1.1 Glass Soft Composition Data

This data is automatically generated from the Glass Soft system, previously mentioned in section 3.3.1. It is comprised of weight measures at 5 minute intervals when the raw materials leave the silo and go into the mixer. After mixing all the desired raw materials, they will be stored in one of the furnace silos connected to the doghouse entrance of the oven. If the silo number ends with one, it is the right furnace silo; on the other hand, if it terminates with the number 2, it is the left furnace silo.

Because of the integration of the BAMeX solution explained in 3.7, the data was retrieved from two distinct sources: Azure DataLake and a PostgresSQL database. A brief description of each attribute, its range or level and type is presented in 4.1

Attribute	Description	Range / Levels	Type
TimeStamp	A timestamp when the measurements occured	1/10/2020 - 31/03/2022	Datetime
Furnace	The destination furnace	AV2, AV4, AV5	Categorical
Furnace Silo	The destination furnace silo	21, 22, 41, 42, 51, 52	Categorical
Raw Material	The type of raw material being weighted	52 different raw materials	Categorical
Raw Material Silo	The origin silo of the raw material	27 different raw material silos	Categorical
Recipe	The current recipe being produced	8 unique recipes	Categorical
Scale	The scale used to weigh the raw material	8 different scales, detailed in \crossref	Categorical
Serie	Identifier of each weighing period		Numeric
Weight	The actual weight	From a few grams up to 2 tons	Numeric
Weight Setpoint	The desired weight		Numeric
Weight Measurement Unit	The weight measurement unit	kg, g	Categorical

Table 4.1: Dataframe containing Glass soft data

4.1.2 Historical Data

This data comes from ETL and Data Scrapping [1] scripts developed previously by members of the Data Science and Analytics team. These scripts generate a *.csv* file that contains all the data that will be used by the Furnace Optimizer tool, previously mentioned in 3.6 In the table 4.2, we will describe these attributes, how they are obtained and what type of attributes they are:

Variable	Unit	Description	Data acquisition	Data origin
Bottom_Temp	^{o}C	The bottom block temperature of the furnace	Thermocouple	Primary : SQL query to PostgreSQL DB Secondary : Fusion file: *Indicadores_Fornos.xslx*
Font_Temp	^{o}C	The temperature in the font region of the temperature	Thermocouple	
Crown	^{o}C	The temperature at the top (Crown) of the furnace	Thermocouple	
Gasflow	m^3/h	The average daily gas flow into the furnace	Gas meter	
Boosting	kWh	The average daily amount of electrical assistance used	Digital Wattmeter	
Air	m^3	The average daily amount of air fed into the furnace	Gas meter	
Pull	*ton*	The amount of glass taken out of the furnace	Digital scale	Fusion file: *Indicadores_Fornos.xslx*
Cullet	*kg*	The amount of cullet going in the furnace	Digial scale	
Stones	*stones/kg*	Stones quality indicator	Manual sampling	
Seeds	*seeds/kg*	Seeds quality indicator	Manual sampling	
Blisters	*blisters/kg*	Blisters quality indicator	Manual sampling	
PCI	kWh/m^3	Lower Heating Value (*Poder calorífico inferior*) of the natural gas being supplied		Data scraping from supplier website
Gas_kw	kWh	The average daily amount of power being provided by gas	*Derived*	$Gasflow * PCI$
Gas_ratio		The ratio of power being supplied by gas	*Derived*	$Gas\_kw/Gas\_kw + Boosting$
Air_gas_ratio		The ratio of air to natural gas being fed into the furnace	*Derived*	$Air/Gasflow$
Energy	$kcal/kg$	The total daily energy being provided to the furnace	*Derived*	$(Gas\_kw + Boosting) * 24 * 860.421/Pull * 1000$ [2]
Color		The color being produced in the furnace	User input	SAP table
Redox		The Redox state of the mixture being fed into furnace		
SO3		The amount of SO3 in the raw materials being fed into the furnace	*Derived*	Fusion file: *Consumo Materias Primas.xslx*
SO2		The amount of SO2 in the raw materials being fed into the furnace		
Humidity		The relative humidity in the raw materials being fed into the furnace		
Ceramicos	*gpt*	The amount of non ferrous metals present in a cullet sample		
Metais não ferrosos	*gpt*	The amount of non ferrous metals present in a cullet sample	Manual Sampling	Norcasco information system
Metais ferrosos	*gpt*	The amount of ferrous metals present in a cullet sample		
Opalas	*gpt*	The amount of opals present in a cullet sample		

Table 4.2: Historical data variables description

Observing the column **Data Origin**, it is possible to conclude that most of the data comes from excel spreadsheets, which, relying on manual entries, will severely affect the data quality.

4.1.3 Inspection Machine Data

The use of this data was influenced by the fact that training our predictive models with indicators obtained via sampling, proved to produce unsatisfactory results due to the small part of the production they encapsulate.

A meeting was arranged with the cold end supervisor in order to discuss what indicators could be found in the cold end that accurately reflect the furnace operation and cullet contamination. The Chili Machines that inspect both the Sidewall and Bottom of every bottle produced stood out as the primary candidates to provide an accurate indicator regarding the existence of Stones in the bottles.

[1] Data scrapping or simply web scraping is the process of extracting data from a website via a computer program and storing it, locally or in a remote location

More specifically, the Bottom Stress and Sidewall Stress Counters, which identify small inclusion that creates stress in the glass during the annealing stage, can be determined when looking at the bottle through a polarized lens.

In the case of Seeds, although there are machines that specifically count this type of defect, the cold end team advised not to use this to constraint the furnace working condition since while this defect can have its root cause in the melting stage, it can also occur during the forming stage, namely on the forehearth or in the IS machine.

The same data split of the previous case was also present in this data set. Still, this time, the extraction and treatment of all the data from the SQL server was pursued since a preliminary analysis demonstrated that the **Bottom Stress** value was always 0. Each row of data obtained via an SQL query contained a timestamp, a LineID, and 180 counters.

Documentation was available on the meaning of these 180 different channels, making it possible to obtain both the Chili Bottom and Sidewall Machines data, formatted as they are currently in BAMeX.The dataframes can be seen in 4.3 and 4.4

Attribute	Description	Right Channel	Left Channel
name	An name composed of machine number and line identifier		
time	Timestamp for each measurement		
Defects	Total number of defects	N6	N43
Dimensional	Number of dimensional defects in the sidewall	N36	N66
Inspected	Total number of inspected bottles	N5	N42
Sidewall	Number of visual defects in the sidewall	N35	N66
Sidewall Stress	Number of visual defects with stress in the sidewall	N121	N68

Table 4.3: Sidewall inspection machine dataframe

Attribute	Description	Right Channel	Left Channel
name	An name composed of machine number and line identifier		
time	Timestamp for each measurement		
Defects	Total number of defects	N9	N46
Inspected	Total number of inspected bottles	N8	N45
Bottom	Number of visual defects in the sidewall	N48	N73
Bottom Stress	Number of visual defects with stress in the sidewall	N119	*Null*
Moulds non read	The mould code present in the bottom section could not be read	N133	N134
Mould rejects	The specific mould read was indicated for rejection	N10	N47

Table 4.4: Bottom and Finish inspection machine dataframe

Unfortunately, the mapping to the **Bottom Stress** seems faulty, so there is no possibility of retrieving this data. This issue was reported to the System&Connectivity department for a later fix. Still, the analysis proceeded with just the **Sidewall Stress** counter, as this type of defect is detected exponentially more than the **Bottom Stress** one, fundamentally because the Sidewall area is much bigger than the Bottom one, making it more probable for inclusion to occur in this area.

4.2 Exploratory Data Analysis

One step of the data understanding stage of the data mining process is the Exploratory data analysis (EDA). Typically this will serve as an exploratory stage to study the distributions of certain numeric variables, comprehend the different levels of categorical values, and potentially identify mistakes or any other problem in the data set, among others.

4.2.1 Historical Data

This sampling issue that affects the Stones indicator, is also present with the variables extracted from the Norcasco information system, particularly since variables known to be the cause of Stones, such as Ceramic and Opals contamination, show little to no correlation with these values.

4.2.2 Glass Soft Data

Following the daily aggregation of the data later discussed in 5.1.2.1, it was possible to obtain some visualizations to observe the evolution of the different quantities of raw material A.2 that enter the furnace and, more specifically, the amount of cullet A.3. There is a period before **September 2021** that no materials entered the furnace since it was under reconstruction, and the production was considered stable after **October 2021**, so we will use this period for our analysis.

A boxplot analysis A.4 for each variable on a time axis was also performed to assess if there were any significant changes in some of the raw materials being used throughout the operation of the AV5 furnace.

4.2.3 Inspection Machine Data

Several visualizations were performed for this data set to assess if the limits set by the [32] standard were actually being met. This was verified in B.2 and B.1, but there were periods where line 54 was responsible for a lot of stones compared to the others.

This can be explained by the fact that it's a **quadruple *gob*** IS machine with high cadency, evidenced by B.4, and the shortcoming of our new quality indicator, that it only counts one stone per container. Therefore having a higher cadency, differentiates this line from the others that produce bigger containers, shown in B.5 and B.3 at a slower cadency that could have more inclusion in just one container.

4.2.4 Unsupervised Analysis

In order to explore possible novel features that may be used to predict stones, some unsupervised approaches were pursued. As referenced in 2.6, the difficulty with clustering or anomaly detection is the interpretability of the clusters or anomalies.

A promising result was found concerning the Redox attribute using the DBSCAN algorithm, discussed in 2.6.3.

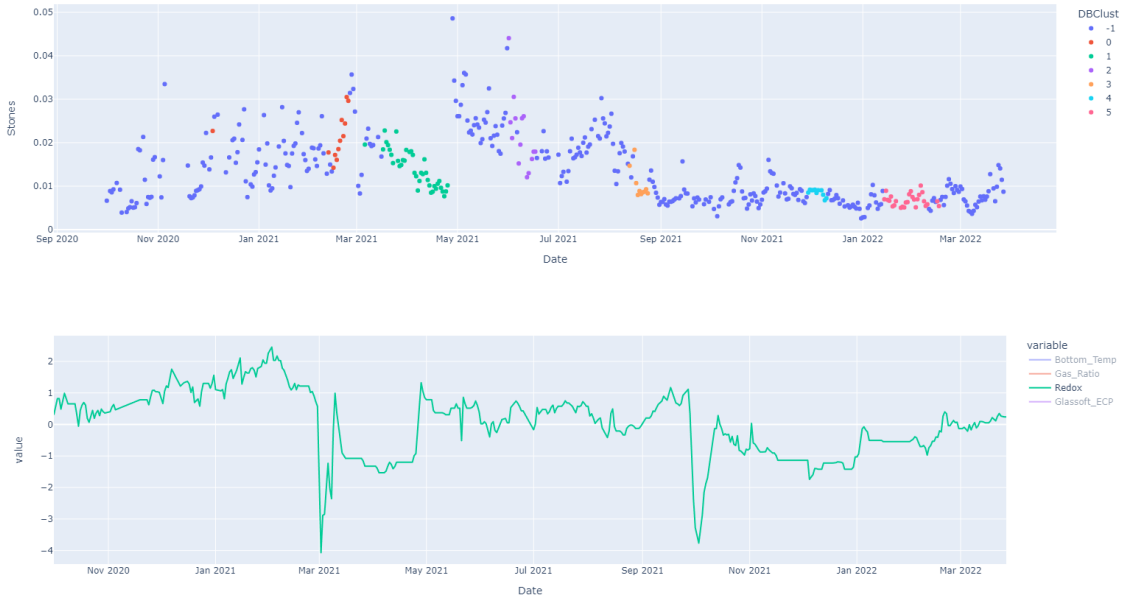


Figure 4.1: DBSCAN clustering algorithm

Source: [31]

The figure 4.1 has seven different clusters, the predominant one labeled as **-1**. The other clusters seem to have a stone count distribution significantly lower than the one present in the cluster labeled as **-1**. Manually several variables were explored, and the one that stood out was the redox value. Particularly we can see that the clusters labeled **1** and **0** have contrary trends in stone occurrences during an abrupt change in the redox stage, presumably a color change.

Chapter 5

Data Preparation

As stated in section 1.4, in this stage of the methodology, new variables will be generated, abnormal values will be dealt with, and since the data sets in question are time series, the corresponding lags between variables are calculated.

5.1 Feature Engineering

5.1.1 Historical Data

This data has already suffered an ETL process and is practically ready to be used in a machine learning model. The only transformation that was done to this data set was the encoding of **Color** Variables into *Dummy Variables*. This type of encoding creates new columns for every level of the categorical attribute. In the case of the AV5 furnace, only two different colors have been produced in our time frame resulting in two different columns that are mutually exclusive, as can be seen in 5.1.

Variable	Color_VB	**Color_UV**	**Color being produced**
i	1	0	VB Green
i +1	0	1	UV Green

Table 5.1: Dummy variables for the Color attribute

5.1.2 Glass soft data

As mentioned in the previous chapter, this was obtained via two distinct sources. Following preliminary treatment, including renaming the dataframes columns, a merge was completed using the **Series** attribute to avoid duplicate entries.

The composition system is present in all three furnaces from the Avintes Plant. Since the work was developed in the AV5 furnace, only the data regarding it was selected. After this, three distinct recipes could be identified:

- F5 Completa - the standard recipe being manufactured in the AV5 furnace;
- AV5 5 ton UV - an experimental recipe used in the period from **02-11-2021** to **05-11-2021**;
- Feldespato - a single recorded weight, this material is used with a defect called Devitrification[1].

The Raw Material attribute, which indicates the type of raw material being weighted, had several redundant data entries since it allows the user to manually input this information.

For example, '*C.sódio, C.sodio saisa, C. sodio solvay, Carbonato de sódio*' all refer to the same raw material, Sodium Carbonate, but due to the inclusion of the supplier or simply an abbreviation, it produced an increase in this attribute levels. Following the renaming of these values, it was possible to **reduce** the levels from **44** to **21**. The Batch and Furnace Team validated both the elimination of the experimental recipes and the aggregation of materials.

The following step was to aggregate the data on both an **hourly** and **daily** basis. The hourly aggregation was done to allow for possible time shifts in the future due to the fact that we are leading with time series data, and the daily aggregation was used to compare, to the historical data currently in use by the furnace optimizer.

Finally, an aggregation based on the Raw Material levels was done to identify if the material being used was cullet and, if so, what type and the percentage in which these materials were used in hourly or daily time frames.

Cullet type aggregation :

- External - Casco Mistura;
- Internal - Casco UVA, Casco AM, Casco AM Cobre, Verde Escuro, Flint UV, **Casco Branco**, Branco Flint, Branco Azulado;
- Complete - Casco UVA, Casco AM, Casco AM Cobre, Verde Escuro, Flint UV, Casco Branco, Branco Flint, Branco Azulado, Casco Mistura.

Here although **Casco Branco** is technically external cullet since it comes from the Norcasco treatment facility, it was suggested to perform this aggregation in this way, due to the fact that this type of cullet has a much more tight control than the mix cullet one.

Following this aggregation we were able to obtain the respective cullet percentages :

$$Glassoft\ External\ Cullet\ Percentage = \frac{Total\ External\ Cullet(t)}{Total\ Raw\ Materials(t)} \tag{5.1}$$

$$Glassoft\ Internal\ Cullet\ Percentage = \frac{Total\ Internal\ Cullet(t)}{Total\ Raw\ Materials(t)} \tag{5.2}$$

$$Glassoft\ Cullet\ Percentage = \frac{Total\ Cullet(t)}{Total\ Raw\ Materials(t)} \tag{5.3}$$

$$\tag{5.4}$$

[1] Desvitrification is a rarely occurring defect where crystals form on the surface of the glass, making it lose its transparent quality

5.1.2.1 Daily Aggregation

An aggregation based on the timestamp allowed us to obtain the total daily values for each material.

Since the furnace optimizer uses daily values, an immediate comparison was made to the data being used. For this, it was necessary to obtain the pull values to calculate the cullet rate. Since the ETL process may lead to the elimination of certain data entries, the SAP table where the ETL program gets its values were directly accessed.

By getting the pull values it was possible to calculate the cullet rate by the expression 5.5

$$Glassoft\ CulletRate = \frac{Total\ Cullet(t)}{Pull(t-1)} \tag{5.5}$$

Given that the cullet was previously split into External and Internal, it was also possible to calculate this metric for both cases, using the expressions found in 5.6 and 5.7

$$Glassoft\ Internal\ Cullet\ Rate = \frac{Total\ Internal\ Cullet(t)}{Pull(t-1)} \tag{5.6}$$

$$Glassoft\ External\ Cullet\ Rate = \frac{Total\ External\ Cullet(t)}{Pull(t-1)} \tag{5.7}$$

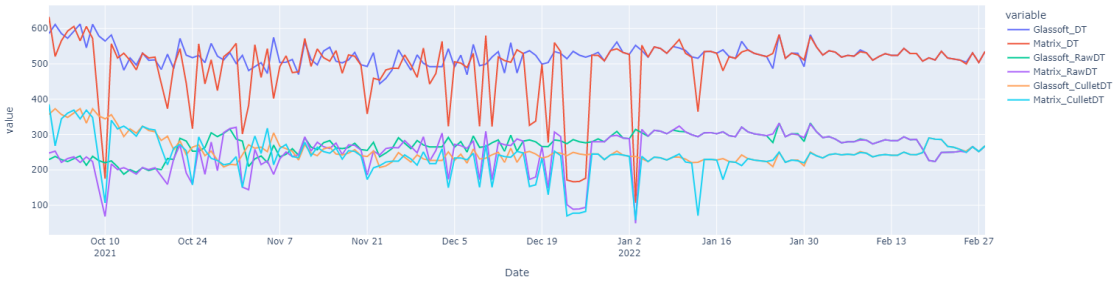


Figure 5.1: Offset between the SAP Matrix and the computed Glass soft Cullet Rate

There is an apparent offset before February, which can be seen in the figure 5.1, having its root cause related to improper data input from the factory operators. **This can be a possible cause of the underperforming Furnace Optimizer in the Avintes Plant.**

5.1.3 Inspection machine data

From section 4.1.3 it was learned that both Sidewall Stress and Bottom Stress could be good indicators for Stones, and they would grasp the entirety of the production. Due to a previously discussed error in the current mapping, the Bottom Stress value was not being captured. Thus only the Sidewall Stress value will be used from now on.

A new attribute called **SWS_Ratio**, Sidewall Stress Ratio was computed to have a sidewall measurement independent of the number of bottles being produced. It was calculated using the expression present in 5.8

$$SWS\_Ratio = \frac{Sidewall\ Stress}{Inspected} \tag{5.8}$$

5.1.4 Time shifts

Considering that the data being dealt with is of the time series type and that there is access to both hourly records of the composition system and the sidewall stress rejections, it is possible to perform a cross-correlation analysis to find the average time shifts between both of these data sets.

The time lag that exists between both our data sets can be decomposed into three parts:

1. From the IS machine until the cold zone inspection area, a bottle will take between 1:30 to 2:30 hours. This amount varies because the throughput of the packaging area is superior to that of the hot end, but it does not operate continuously due to human integration. To circumvent this characteristic, queues have been built into the production line, creating this degree of uncertainty when it comes to the travel times between the hot end and cold end;

2. The residency time in the furnace for normal pulls is about 24 to 30 hours;

3. The only unknown variable is the time the mixed ingredients stay in the furnace silos, as this can vary due to many different factors, such as shifts, and silo cleaning, among others.

When performing this cross-correlation, the top 10 best time shift correlations were identified, and then a mean was taken from the acceptable range, which will assume a minimum of 30minutes until the mixed ingredients are fed into the furnace. The range was defined from 26-48 hours.

```
Top 10 time lags = [23 20 14 35 26 31 7 32 33 36]
Acceptable = [35 26 31 32 33 36]
Mean time shift = 32.1666 hours
```

This time shift will later be used to bring our variables into the same instance prior to applying the final models.

5.1.5 Converting Sidewall Stress to Stones

Since sampled stones were calculated by dividing the number of stones by the total amount sampled, following the same logic, one can easily calculate the stone value by dividing our total sidewall rejection number by the pull, since we sampled the entirety of the glass produced.

A comparison between the stones obtained from the historical data set and the new metric obtained via the sidewall stress data is present in 5.2. It confirms that although the error being made

was not as big as expected when taking into account that only 0.045%[2] of the entire production was being sampled, the new indicator improves significantly on these results.

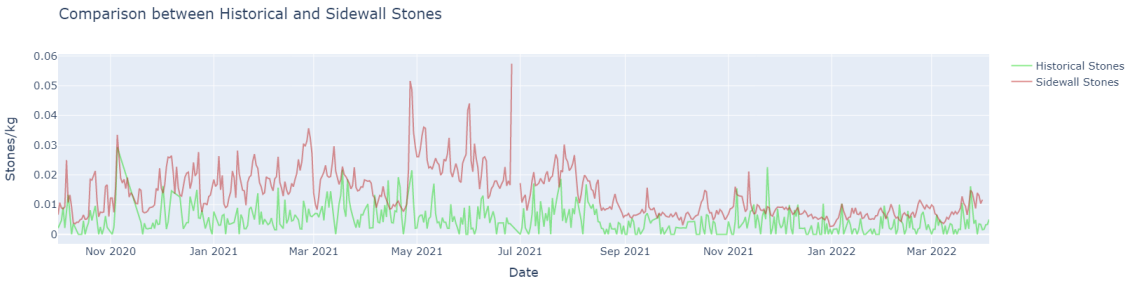


Figure 5.2: Comparison between Historical and Sidewall Stones

5.1.6 Shifted Stones

Since industrial glass-making is a continuous process, a variable was created, taking the stones variables and introducing a 24 hours shift. In simpler terms, yesterday's stones will be used to predict the stones that will occur today.

5.2 Feature Selection

In this section the selection of the appropriate variables for our main regression model will be discussed. Two courses of action were taking regarding this: Manual Selection, and Recursive Feature Elimination.

5.2.1 Manual Selection

It is known from a user experience that Bottom Temperature and cullet use have a great effect on the number of stones the final product will have. Considering this, we selected the following variables :

- Bottom_Temp - Bottom Temperature
- G_CP - Glassoft Cullet Percent
- G_ECP - Glassoft External Cullet Percent
- G_ICP - Glassoft Internal Cullet Percent

The Glass Soft variables have been shifted 32 hours at this stage. A correlation matrix analysis between the historical data variables and the sidewall stress was obtained via the inspection machines. This confirmed the choices of the **Botom_Temp** and **Glassoft_CP** since they both show strong correlations of **0.586** and **0.594**, respectively.

[2]Calculated assuming an average weight of 0,5 kg for the 360 containers for a pull of 400 ton

Following this, based on other users inputs discussion with the co-workers, the following variables were also selected:

- Gas_Ratio
- Redox
- Pull

5.2.2 Recursive Feature Selection

In a second approach to feature selection, a recursive feature elimination algorithm with a Random Forest estimator was used, and the feature importance threshold was set at 2.5%. Seven attributes were selected with an importance over the desired threshold and they can be seen in the 5.2, ordered by their computed feature importance.

Feature	**Importance (%)**
Bottom_Temp	30.00%
Glassoft_CP	16.56%
Glassoft_ECP	15.37%
Gas_Ratio	12.64%
Boosting	12.16%
Redox	10.77%
Gas_kw	2.50%

Table 5.2: Result of the RFE algorithm

Chapter 6

Modeling and Evaluation

This chapter discusses the process of selecting, developing, and assessing a quality predictive model.

6.1 Modeling

To predict our target variables stones, three distinct models were utilized. These models are **Light Gradient Machine Bosting Regression** (LGBM), **eXtreme Gradient Boosting** (XGBoost), and **Support Vector Regression** (SVR).

Four different data sets will be considered for each case, since there will be a first split in the variables between the variables that were manually selected and the ones that have been selected with the Recursive Feature Elimination method. The second split will occur with the addition or omission of the **Stones_Shifted** variable, as the addition of this variable enhances our model performance but changes the nature of the prediction. In this way, the four different data sets will be obtained and are detailed in the table 6.1.

	A	**B**	**C**	**D**
Bottom_Temp	X	X	X	X
Glassoft_CP	X	X	X	X
Glassoft_ECP	X	X	X	X
Glassoft_ICP	X	X		
Gas_Ratio	X	X	X	X
Redox	X	X	X	X
Pull	X	X		
Boosting			X	X
Gas_Kw			X	X
Stones_Shifted		X		X

Table 6.1: Variables contained in each data set

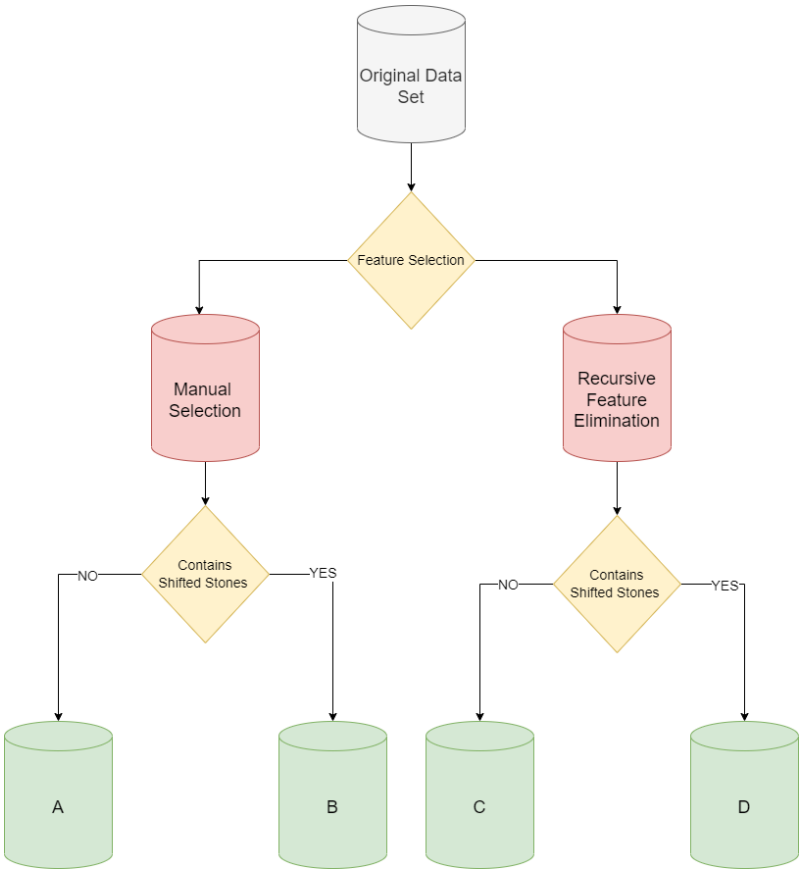


Figure 6.1: A model explaining how the 4 data sets were formed

Since we have a target value, we will need to proceed with Data Split into train and test partitions, as was explained in 2.4.1. The train-test split was performed on the 1st of December of 2021, resulting in **76,2%** (353) of **train** data and **23,8%** (110) of **test** data.

To tune our models, we used the **Optuna** framework that was presented in 2.8.4 with attention to the tuning times. This is mainly due to the fact that complaints have been made from the factory workers concerning longer computation times that more intricate models may have. The maximum tuning times were set around 5 minutes, but with the implementation of Pruning Callbacks, an Optuna functionality that automatically gives up on a study if it is not evolving in the desired way, it was able to reduce this tuning time. These Pruning methods are only available to Decision Tree-based methods, as is the case with XGBoost and LGBM.

The SVR method takes much less time to train, and therefore it was subjected to a lot more training instances than the other two methods. Typically the SVR method is the fastest, with 200 iterations taking about 30 seconds, while the LGBM takes 1 minute and 20 seconds for 20 trials, and XGBoost is the slowest, with 10 trials taking 2 minutes and 40 seconds[1]. Boosting methods tend to overfit the data and only need a reduced amount of iterations to achieve satisfactory results.

[1] for reference, the tuning of the models was performed using an Intel Xeon E3 1240 v2 with 16GB of RAM

The metric for the studies was obtained using **Kfold cross-validation** with five folds for each trial, with it being an average of the mean absolute error.

In order to access our model performance, we need to determine what metrics will be used initially. As introduced in 2.4.2 the performance metrics that will be used are the **Coefficient of Determination**, **Mean of the Squared Error**, **Root Mean Squared Error**, and the **Mean Absolute Error**. In order to verify if the model is being overfitted, a split was done between the evaluation metrics for the train and test set.

The models' outputs were interpreted using the **SHAP** library introduced in 2.3.2. It was used with both the LGBM and XGBoost models. Although there is built-in support for Kernel Explains in SHAP, it is a very computationally demanding task taking about 50 minutes to complete and therefore was not pursued.

6.2 Evaluation

Following the methodology set out in previously in this section, we will look at the performance metrics and graphical outputs of our models to compare them and evaluate their predictive capabilities.

From the results presented in C.1, one can conclude that the best model was XGBoost for all data sets but showed better results in the ones that included the Stones_Shifted attribute. Slightly behind comes the LGBM Regressor, which achieves it's best results on the B data set. In third place comes the fastest method, but the one that performed worst, the SVR.

It is particularly evident from the difference in train and test R^2 value, the propensity to closely fit the training data that boosting methods naturally have. Nevertheless, both of our boosting methods performed remarkably well when faced with unseen data.

Let us compare our methods as described in 6.2.1 and 6.2.2 in order to evaluate how each of the models dealt with the different data sets.

6.2.1 A and C data set

From the visual analysis of the annex C.1 and C.5 , we can see that both the XGBoost and LGBM methods anticipated the increase in stones that occurred in late February. Another common aspect is that they both often overestimate the stones. On the contrary, the SVR method tends to make predictions below the actual number.

The main difference between LGBM and XGBoost for the A and C data set seems to be the ability for the XGBoost in C.1 and C.5 and LGBM in C.5 to predict a sudden increase in stones by sometimes exaggerating or sometimes, as is the case at the end of March, to predict this increase with remarkable accuracy.

From interpreting the SHAP values in figures C.2 and C.6, we can conclude that both low Bottom Temperatures and high Gas Ratios have high feature values. It can also be noted that in the LGBM cases, low amounts of external cullet will have high feature values, which confirms established knowledge base.

In C.2 the Pull and Glassoft_ICP are centered and don't have a particular distinction in color, meaning that although they are not bad, they are also not great predictors for the stone variable.

In C.6 the same can be said about the boosting values as they are centered and do not have a great color distinction. Another thing we can interpret from these graphs is that the XGBoost can give less importance to these average predictors than the LGBM.

6.2.2 B and D data set

Based on the analysis of figures present in C.3 and C.7 we can see that there is a small increase in predictability with the addition of the Stones_Shifted variable, particularly in the January Period, except for the LGBM with the D data set method.

There also seems to exist an increase in the ability to predict the increase in stones before it happens. Regarding the SVR method, it particularly underperforms with the D data set, overestimating the seeds in the entirety of the prediction.

Looking at the SHAP values plots in figures C.4 and C.8, we can see that the added variable has become the top predictor in all 4 cases and overpowered the other features in the LGBM case.

In XGBoost, Bottom Temperature and Gas Ratio come either in 2nd or in 3rd, cementing their spot as one of the best predictors for stones.

Chapter 7

Conclusion and Future Work

7.1 Conclusions

In the glass-making business, the process is extremely complex and hard to grasp. In these past months, I have gained significant knowledge of the glass-making process and the various factors that influence it.

The goal we set out to achieve in the objectives section 1.3 has been fulfilled with the creation of a new quality indicator, which encapsulates the entirety of the production without relying on sampling. Besides only looking into a small subset of the data, this inspection method is human-dependent, and it may include several unknown variables that are not being captured. This new indicator, paired with the data obtained from the Glass Soft composition system and the other previously acquired furnace data, is showing promising predictive results.

7.2 Future Work

Within the scope of the work that was developed in the Avintes plant, several projects could be pursued. The first is rethinking the current Furnace Optimizer tool to deal with hourly data, to allow for several variables from the furnace to be brought to the same instance as per example the inspection machines variables. This approach would surely increase robustness, a problem that has plagued this tool, more evident when there is a significant change in the process.

A new recommendation system for furnace operation could also be created by combining interactive graphs and real-time sampling with some of the knowledge gained from this study and bringing Statistical Process Control aspects into the mix. This combination could help both Cold-End and Hot-End teams in their everyday life on the factory floor.

Also, one of the shortcomings of the new quality indicator that was created is the fact that it relies on a sidewall stress counter that does not count the amount of inclusion present in each container. Since the inspection machines function with the visual assessment done via a photograph, this indicator could be improved with a computer vision to obtain the number of inclusions per container, therefore obtaining the sampling characteristics for the entire production.

Regarding the entirety of the BA group, due to the increase in cullet consumption, other factories have started to face serious issues in their production lines. A future project in Villafranca de Los Barros is currently being planned, and the work that was developed in this dissertation will be directly applied there and completed with other sources of data such as the cullet treatment factory, truck sampling and registration.

This plant has been suffering from jamming in the IS Machines due to big contaminants, leading to a stoppage of the affected sections and increased production losses. The aim for this project will be to possibly identify the sources that have a higher probability of containing these specific contaminants and based on the other sources of data regarding the cullet quality, build a recommendation tool for the Batch and Furnace team, in order to adjust the furnace conditions appropriately

Appendix A

Glass soft data visualizations

A.1 Comparison between Sand and Coal quantities

Figure A.1: Comparison of daily values [*ton*] obtained from the Glass Soft information system

A.2 Glass soft daily

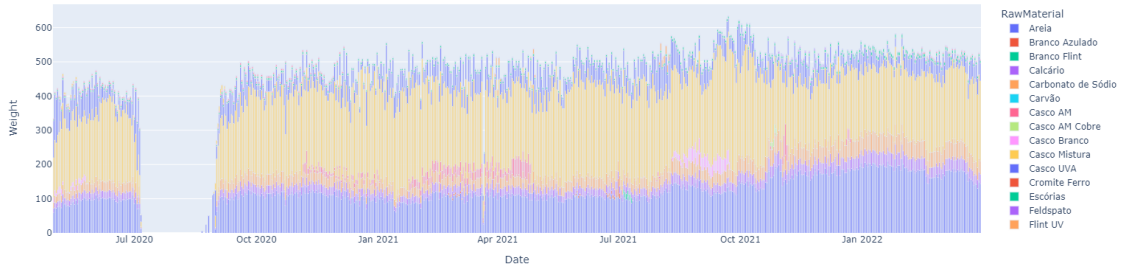


Figure A.2: A bar-plot with the daily weights for the AV5 furnace

A.3 Glass soft daily cullet

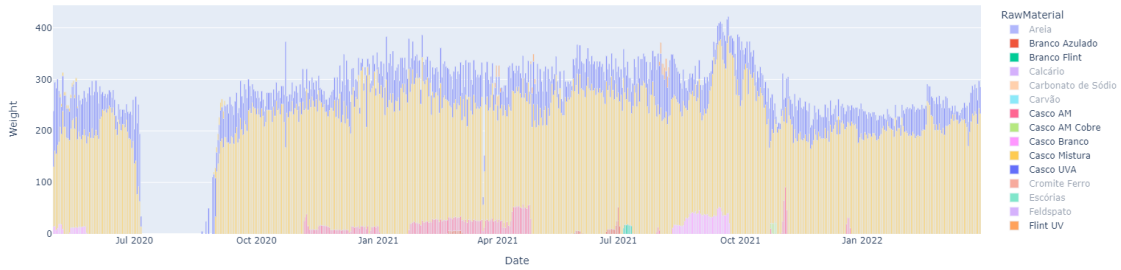


Figure A.3: A bar-plot with the daily weights for the AV5 furnace

A.4 Boxplots

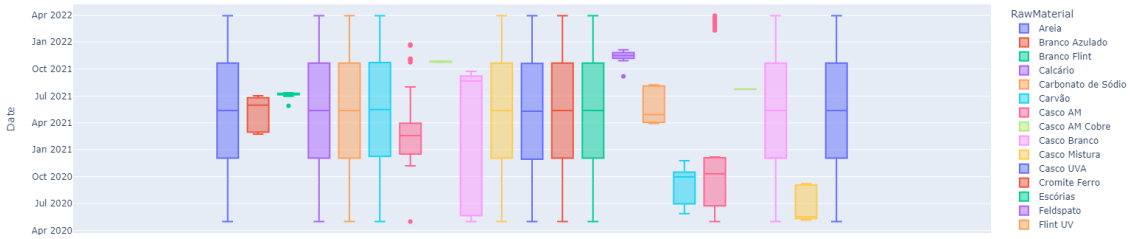


Figure A.4: Box plot distribution of the materials over time

Appendix B

Inspection machines data visualizations

B.1 Daily stones by line

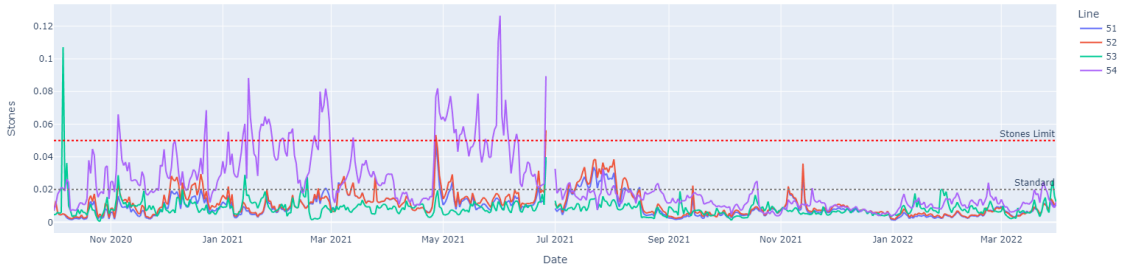


Figure B.1: A line chart comparing the stones values between the different lines

B.2 Daily furnace stones

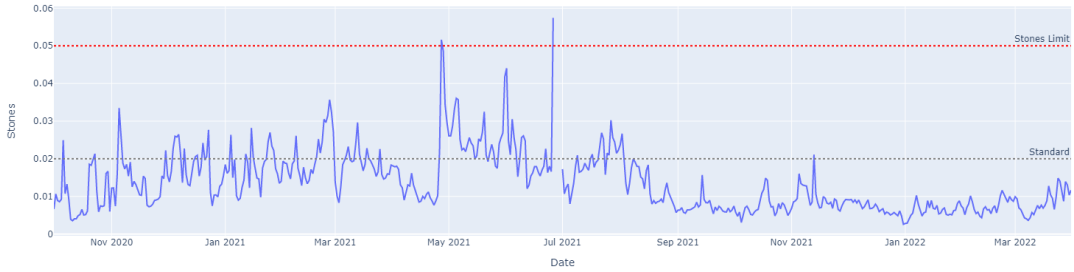


Figure B.2: A line chart for the evolution of the AV5 furnace stones

B.3 Daily pull by line

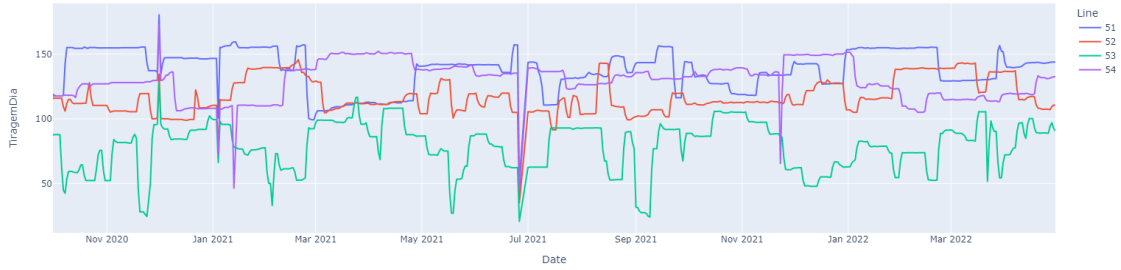


Figure B.3: A line chart for the evolution of the AV5 furnace pull by line

B.4 Daily inspections by line

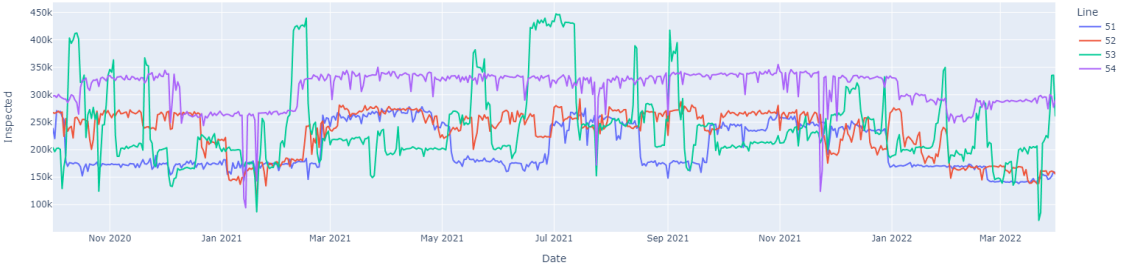


Figure B.4: A line chart for the evolution of inspections by line

B.5 Bottle weight box plot

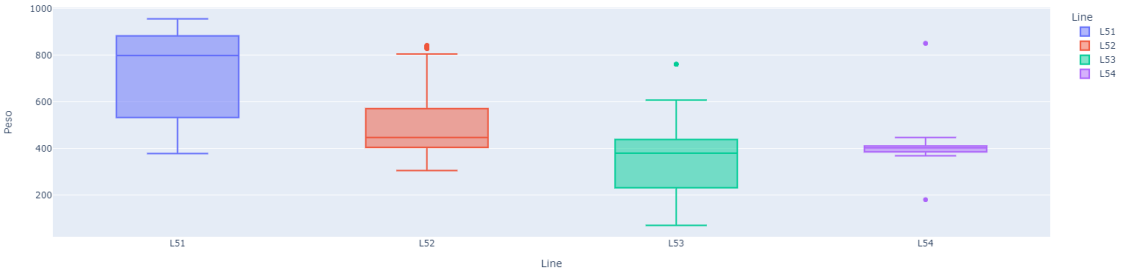


Figure B.5: A box plot for the weight of the bottle being produced in each line

Appendix C

Regression Results

C.1 Regression Performance Metrics

Model	Dataset	R2_test	MAE_test	MSE_test	RMSE_test	R2_train	MAE_train	MSE_train	RMSE_train
SVR	**A**	-0.436336	0.001937	6.73E-06	0.002595	0.369192	0.004532	3.72E-05	0.006096
	B	-0.394380	0.001946	6.54E-06	0.002556	0.373655	0.004473	3.69E-05	0.006074
	C	-7.674505	0.005748	4.07E-05	0.006376	0.445074	0.003810	3.27E-05	0.005718
	D	-7.857347	0.005805	4.15E-05	0.006443	0.455286	0.003802	3.21E-05	0.005665
LGBM	**A**	-0.641078	0.002197	7.69E-06	0.002773	0.935599	0.001407	3.79E-06	0.001948
	B	-0.123491	0.001747	5.27E-06	0.002295	0.946113	0.001212	3.17E-06	0.001782
	C	-0.596342	0.002055	7.48E-06	0.002735	0.941896	0.001356	3.42E-06	0.001850
	D	-1.693974	0.003166	1.26E-05	0.003553	0.612352	0.003522	2.28E-05	0.004779
XGBoost	**A**	-1.079990	0.002447	9.75E-06	0.003122	0.984583	0.000709	9.08E-07	0.000953
	B	0.247779	0.001466	3.53E-06	0.001878	0.988167	0.000634	6.97E-07	0.000835
	C	-0.633378	0.002196	7.66E-06	0.002767	0.984066	0.000713	9.39E-07	0.000969
	D	0.247779	0.001466	3.53E-06	0.001878	0.988167	0.000634	6.97E-07	0.000835

Table C.1: The complete result for predictive models for each data set

C.2 Predictions and SHAP values

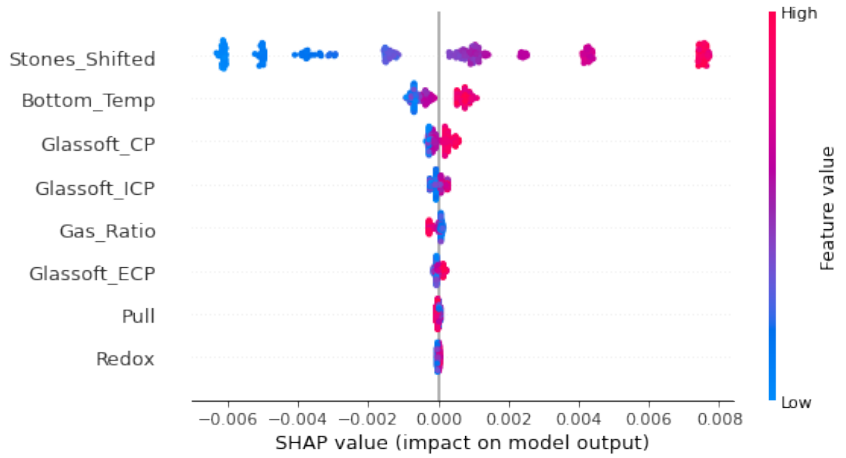
Figure C.1: The predictions of our three models for the A dataset

(a) Shap values for the LGBM Regressor

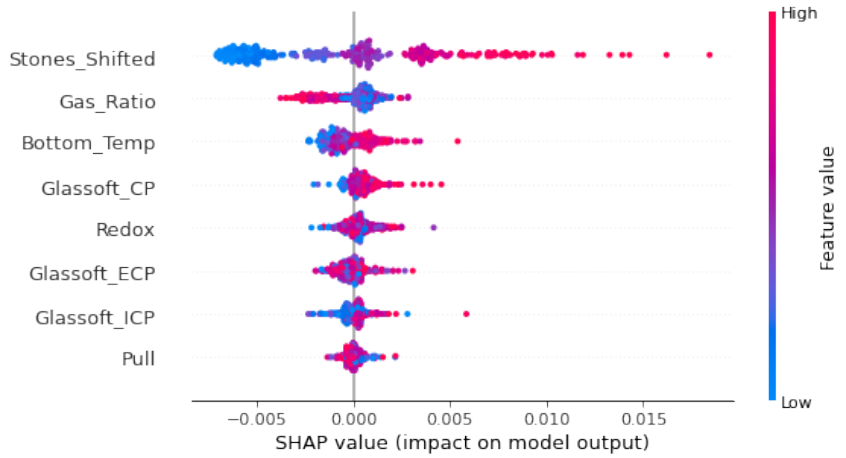
(b) SHAP values for XGBoost Regressor

Figure C.2: SHAP values for the decison-tree based methods response to the A datset

Figure C.3: The predictions of our three models for the B dataset



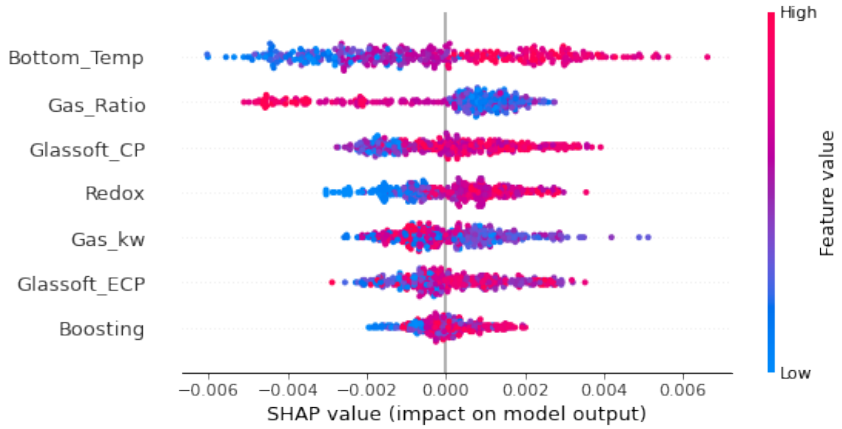
(a) Shap values for the LGBM Regressor



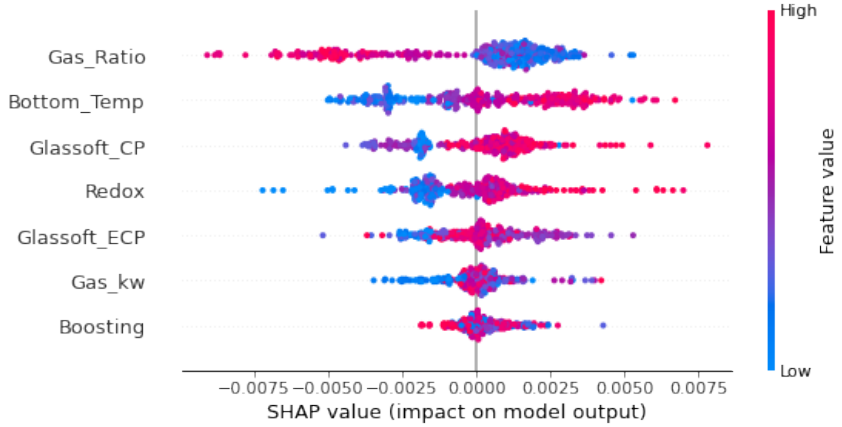
(b) SHAP values for XGBoost Regressor

Figure C.4: SHAP values for the decison-tree based methods response to the B datset

Figure C.5: The predictions of our three models for the C dataset



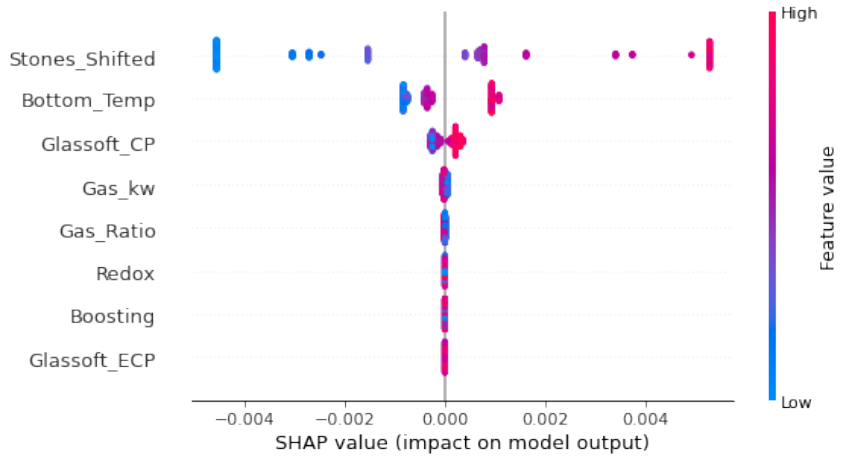
(a) Shap values for the LGBM Regressor



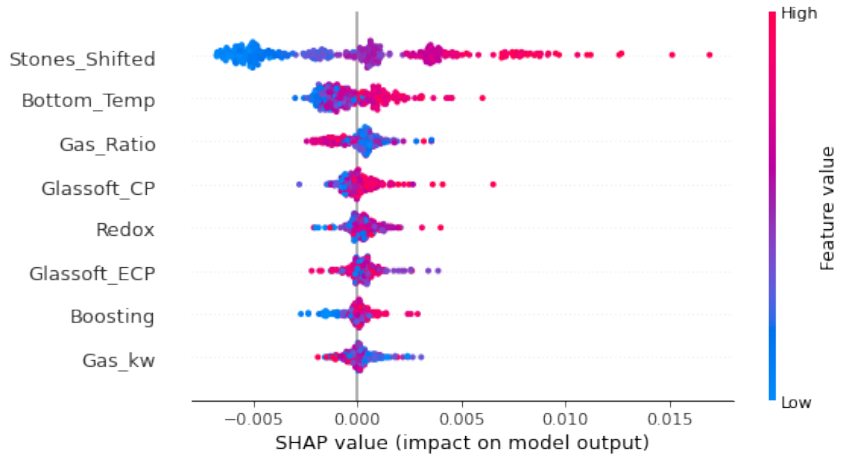
(b) SHAP values for XGBoost Regressor

Figure C.6: SHAP values for the decison-tree based methods response to the C datset

Figure C.7: The predictions of our three models for the D dataset



(a) Shap values for the LGBM Regressor



(b) SHAP values for XGBoost Regressor

Figure C.8: SHAP values for the decison-tree based methods response to the D datset

References

[1] A. Emblem. *Packaging Technology: Fundamentals, Materials and Processes*. Woodhead Publishing in Materials. Elsevier Science, 2012.

[2] BA Glass. BA 2021 Anual Report, Looking for Balance, 2021.

[3] Peter Chapman, Janet Clinton, Randy Kerber, Tom Khabaza, Thomas P. Reinartz, Colin Shearer, and Richard Wirth. Crisp-dm 1.0: Step-by-step data mining guide. 2000.

[4] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

[5] J v Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

[6] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

[7] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learnin. *Cited on*.

[8] L Breiman, JH Friedman, R Olshen, and CJ Stone. Classification and regression trees. 1984.

[9] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[10] J Ross Quinlan. C4. 5: Programs for machine learning, 1993.

[11] Yoav Freund, Robert Schapire, and Naoki Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

[12] Tianqi Chen and Carlos Guestrin. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016.

[13] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[14] Bex T. You are missing out on lightgbm. it crushes xgboost in every aspect, Oct 2021. Available at https://rb.gy/umayaj.

[15] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[16] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[17] Bhupesh Kumar Singh. Evaluation of genetic algorithm as learning system in rigid space interpretation. In *Nature-Inspired Computing: Concepts, Methodologies, Tools, and Applications*, pages 1184–1228. IGI Global, 2017.

[18] Mark A Aizerman. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and remote control*, 25:821–837, 1964.

[19] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

[20] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

[21] Donald R Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of global optimization*, 21(4):345–383, 2001.

[22] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[23] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

[24] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

[25] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422, 2002.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[27] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[28] T. Editors of Encyclopaedia Britannica. Glass, August 2021.

[29] J.E. Shelby. *Introduction to Glass Science and Technology*. EngineeringPro collection. Royal Society of Chemistry, 2005.

[30] New World Encyclopedia. Glass - new world encyclopedia, 2021.

[31] Empakglass. Glass container technology training, 2013.

[32] BA Glass. Glass quality - technical specification 99, March 2021.