applied sciences

MDPI

Article

A Machine Learning Approach for Predicting and Mitigating Pallet Collapse during Transport: The Case of the Glass Industry

Francisco Carvalho [1], João Manuel R. S. Tavares [2] and Marta Campos Ferreira [3,*]

[1] Faculdade de Engenharia, Universidade do Porto, 4200-465 Porto, Portugal; up201806858@edu.fe.up.pt
[2] Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial, Departamento de Engenharia Mecânica, Faculdade de Engenharia, Universidade do Porto, 4200-465 Porto, Portugal; tavares@fe.up.pt
[3] Institute for Systems and Computer Engineering, Technology and Science Faculdade de Engenharia, Universidade do Porto, 4200-465 Porto, Portugal
* Correspondence: mferreira@fe.up.pt

Abstract: This study explores the prediction and mitigation of pallet collapse during transportation within the glass packaging industry, employing a machine learning approach to reduce cargo loss and enhance logistics efficiency. Using the CRoss-Industry Standard Process for Data Mining (CRISP-DM) framework, data were systematically collected from a leading glass manufacturer and analysed. A comparative analysis between the Decision Tree and Random Forest machine learning algorithms, evaluated using performance metrics such as F1-score, revealed that the latter is more effective at predicting pallet collapse. This study is pioneering in identifying new critical predictive variables, particularly geometry-related and temperature-related features, which significantly influence the stability of pallets. Based on these findings, several strategies to prevent pallet collapse are proposed, including optimizing pallet stacking patterns, enhancing packaging materials, implementing temperature control measures, and developing more robust handling protocols. These insights demonstrate the utility of machine learning in generating actionable recommendations to optimize supply chain operations and offer a foundation for further academic and practical advancements in cargo handling within the glass industry.

Keywords: cargo loss; predictive tool; data mining; CRISP-DM framework; glass manufacturer

Citation: Carvalho, F.; Tavares, J.M.R.S.; Campos Ferreira, M. A Machine Learning Approach for Predicting and Mitigating Pallet Collapse during Transport: The Case of the Glass Industry. *Appl. Sci.* **2024**, *14*, 8256. https://doi.org/10.3390/app14188256

Academic Editor: Arkadiusz Gola

Received: 31 July 2024
Revised: 10 September 2024
Accepted: 11 September 2024
Published: 13 September 2024

1. Introduction

Cargo loss within logistics systems poses significant financial challenges for businesses, incurring not only the costs associated with replacing damaged or lost goods but also additional expenses such as incident management, increased insurance premiums, missed business opportunities, and potential reputation damage [1]. The urgency of addressing these losses is stressed by their widespread impact across various sectors, particularly in industries where goods are susceptible to damage during transport, such as in the ceramic [2] and the glass industries [3].

However, previous research on logistics has been predominantly focused on broad solutions, such as route optimization and general cargo securing techniques, without sufficiently addressing the specific issue of pallet stability during transportation. This is a critical gap, as the collapse of pallets can lead to significant product loss and safety hazards, especially in industries with fragile goods like the glass industry. For instance, the work presented in Huixia Cui [4] is focused on optimizing transportation routes, while the work Aleksander Nieoczym [5] is about general cargo securing methods. These works, although valuable, fail to account for the particular risks associated with unstable pallets, particularly in the glass industry where product fragility and unique packaging requirements are major factors.

Moreover, many previous studies lack the granularity required to effectively predict and prevent the collapse of pallets because of insufficient data or overly generalized approaches. For example, the studies presented in Turbaningsih [6] and Praveen Nath S [7] explore ways to improve cargo handling but do not dive deep into the specific variables that lead to pallet instability in glass product transport. Thus, the complexity of the problem is often underestimated and, as a result, previous solutions fall short when it comes to mitigating pallet collapse incidents.

The instability and collapse of pallets during transport can lead to significant product damage and loss, triggering the cascading financial effects previously mentioned, which may be difficult to estimate due to the lack of data or access to them [1]. However, these losses extend beyond property damage and can result in severe human consequences. Inadequate cargo securing for road transport has led to significant fatalities and injuries. For example, in 2014, approximately 1200 individuals in Europe lost their lives in accidents directly linked to improperly secured cargo [8]. These examples highlight the severe consequences of pallet instability and collapse, and stress the necessity for more research focused on overcoming this problem.

In response to the current literature gaps, this study aimed to specifically address the problem of pallet collapse, mainly in the glass industry due to its particular vulnerability to such a problem. Thus, this study fills the research gap by leveraging data analytics and machine learning techniques to predict pallet collapse, thereby offering a more tailored solution to this overlooked problem in the logistics of fragile goods.

Data analytics and machine learning, whose studied techniques are detailed in the following section, have been widely used to solve various problems, such as detecting and classifying vehicle-deck collisions on railway bridges [9], optimizing gate assignments in airports [10], identifying vehicles to reduce traffic infractions [11], and predicting ice resistance to ensure safe ship navigation in icy regions [12]. Here, it is used to predict pallet stability during transportation in the glass industry, and to identify and mitigate factors contributing to pallet collapse.

Therefore, the main objective of this research is twofold: first, to proactively predict potential pallet collapse incidents by analysing data from past incidents and the logistics chain; and, second, to provide practical recommendations that can prevent such collapses, thereby minimizing financial and human risks. The glass industry was chosen as the application case because of its fragile products, but the findings apply to other industries that face similar risks. Thus, this study differentiates itself by addressing the specific contributing factors to pallet instability, which previous research has neglected.

To ensure a rigorous and systematic approach, this research employed the CRoss-Industry Standard Process for Data Mining (CRISP-DM) framework to structure the investigation and analysis. The use of this framework aided in ensuring that the study is both comprehensive and methodologically sound.

In summary, this study makes significant contributions to the literature by focusing on the often-neglected issue of pallet stability, particularly within the glass industry. It provides data-driven predictions and strategies specifically aimed at mitigating the risks of pallet collapse, thereby addressing gaps left by previous generalized approaches in the field. Additionally, the study offers a methodological framework that is not limited to the glass industry but can be applied to other sectors facing similar challenges, broadening the overall impact of the research.

The remainder of the article is organized as follows: Section 2 presents a comprehensive literature review to situate this study within the existing body of knowledge. Section 3 outlines the developed methodology, detailing the data collection process, the used analysis techniques, and the employed machine learning models. Section 4 discusses the findings and interprets their implications for both theory and practice. Finally, Section 5 concludes the article, summarizing the key outcomes and suggesting avenues for future research.

2. State of the Art

This section presents an overview of the current state of the art concerning the prediction of problems with the cargo during transport, namely problems related to the collapse of pallets. To tackle this goal, a Systematic Literature Review (SLR) approach was followed, with the procedure and key findings being described in the following.

2.1. Systematic Literature Review

The collection of articles relevant to this study was systematically conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [13], which are a set of evidence-based recommendations aimed at improving the transparency and reproducibility of systematic reviews. The PRISMA framework is widely adopted across various disciplines to ensure that literature reviews are conducted rigorously and comprehensively. A detailed step-by-step procedure was followed, which is visually represented by the flowchart in Figure 1. This process is further detailed in the subsequent sections. The literature search was carried out using the Scopus and Web of Science databases, ensuring a comprehensive review of pertinent studies.

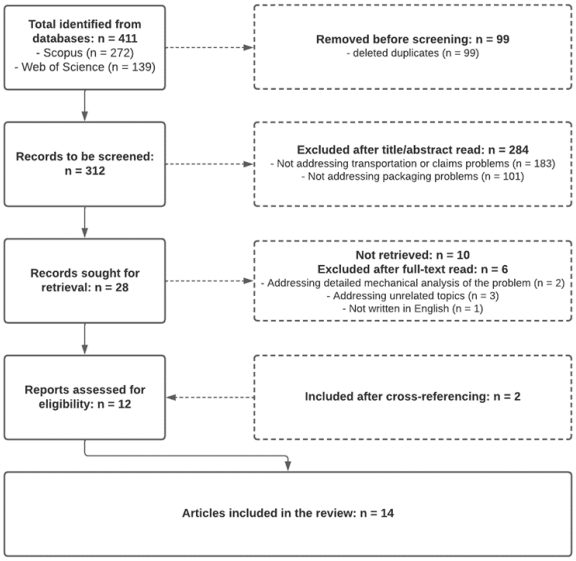

Figure 1. PRISMA diagram of the selection process of the articles reviewed in the current study.

Regarding the articles gathering, the following terms were searched: "damag* good*", "damag* in transit", "freight transport* damag*", "damag* freight", "transit damag*", "freight disrupt*", "deliver* disrupt*", "cargo damag*", "cargo loss", "packag* damag*", "glass damag*", "bottle damag*", "truck transport* damag*", "pallet damag*", "predict* model", "prevent*", "forecast*", "anticipat*", and "damag* claim predict*".

Firstly, a search on both databases was performed combining the previous terms into the following query: ("damag* good*" or "damag* in transit" or "freight transport* damag*" or "damag* freight" or "transit damag*" or "freight disrupt*" or "deliver* disrupt*" or "cargo damag*" or "cargo loss" or "packag* damag*" or "glass damag*" or "bottle damag*" or "truck transport* damag*" or "pallet damag*") AND ("predict* model" or "prevent*" or "forecast*" or "anticipat*"); OR "damag* in transit"; OR "packag* damag*"; OR "transport* damag*"; OR "damag* claim predict*". This query was formulated following a series of preliminary experiments that tested various combinations of topic-related terms and Boolean operators, tailored to meet the objectives of this review.

The query yielded a total of 411 results. Figure 1 illustrates the distribution of the results across the two databases and details the process used to determine the final selection of articles for the literature review. Initially, duplicate entries were removed. Subsequently, the titles and abstracts of the remaining articles were examined for relevance, and those deemed irrelevant were excluded. Articles that were inaccessible in full-text form were

then eliminated. Finally, after a thorough review, the articles deemed unsuitable for the study were also excluded, resulting in a total of 14 articles to be reviewed.

The selected articles were systematically categorized into two distinct groups to facilitate a detailed analysis of the various methodologies employed. The first category, Cargo Damage Classification Techniques, encompasses studies that focus on the methods and technologies used to classify cargo damage. This grouping allows for a concentrated examination of the techniques and tools applied in understanding and categorizing types of cargo damage. The second category, Prediction of Product Behaviour During Transport and Handling, includes research that investigates predictive models and assessment methods for monitoring product behaviour throughout the transportation and handling stages. This classification highlights studies aimed at forecasting potential issues and enhancing the resilience of products during transit, and are presented next.

2.2. Cargo Damage Classification Techniques

Few studies have been published in the literature that specifically address the prediction and further avoidance of the collapse of pallets during transportation; however, many do focus on the classification of cargo damage. There exist different approaches to tackling this topic, with some of them based on questionnaires, while others endorse a data science-based methodology.

In [14], the author proposes a classification of damage to palletized loads based on his experience in the field, which was further verified and validated through an analysis of questionnaire results sent to five selected groups potentially affected by load damage. The author concluded that damage to loads can be categorized into two distinct groups, based on whether the damage reduces the value of the transported products. The study in Tkaczyk [14] also emphasizes the importance of anticipating damage that diminishes cargo value due to the resultant externalities, such as the increased production of goods and packaging, and heightened transportation demands, which contribute to greater environmental pollution.

In the study described in Wu et al. [15], a machine learning methodology was employed to predict the severity of cargo loss during transport. To achieve this, interpretable models such as Decision Tree and Logistic Regression models were used. The analysis identified key factors influencing cargo loss, including transit types, product categories, and shipping destinations. Similar machine learning methods were employed in the study presented in Hashemi et al. [16], albeit with a different objective. The research was focused on predicting and categorizing various types of accidents likely to occur on vessels in the Mississippi River. The study found that the constructed Neural Networks model yielded the most satisfactory results, outperforming those obtained from Multiple Discriminant Analysis- and Logistic Regression-based approaches.

The significance of predicting insurance claims within maritime port operations is highlighted in Panchapakesan et al. [17]. This study was pioneering in its application of machine learning techniques to assess the impact of various attributes on shipping container damage. Among several algorithms evaluated, including Neural Network and Decision Tree algorithms, the Random Forest algorithm demonstrated the most satisfactory performance. Notably, the duration a container spends in the storage yard emerged as the most predictive feature for assessing the likelihood of container damage claims.

The study described in Panchapakesan et al. [18] investigated the use of machine learning techniques, such as Random Forest, in conjunction with traditional methods like Decision Tree, to mitigate the operational challenges of processing insurance claims for customers with damaged shipping containers at a maritime port in Canada. The research findings revealed that the Random Forest algorithm yielded the most favourable results in assessing the severity of damage to shipping containers.

2.3. Prediction of Product Behaviour during Transport and Handling

Numerous studies have explored the challenges of freight damage during transit, focusing on how external factors during transportation and product handling could affect the integrity of goods.

In Jarimopas et al. [19], the authors measured the vibration levels experienced by commercial truck shipments in Thailand and assessed their impact on tightly packed tangerines. The experiment evaluated how the combination of different sets of vehicles, vehicle speeds, and road types influenced the vibrations in the vehicles. The results indicated that higher truck speeds led to more significant damage to the transported fruit. Moreover, it was discovered that tangerines transported on laterite roads suffered the most damage and that fruit transported in 2 ton trucks incurred less damage compared to those in 6 ton trucks. An experiment with a similar goal, examining the impact of transport vibrations on apple bruising and evaluating the effectiveness of different packaging options for protecting the fruit, was presented in Fadiji et al. [20]. Unlike the method used in Jarimopas et al. [19], this study employed an electro-dynamic shaker to simulate the vibrations. The research demonstrated that both the design of the packaging and the levels of vibration to which the vehicles were exposed had a significant influence on the extent of bruise damage suffered by the apples.

Additionally, in Schlimme et al. [21], the potential damage to frozen French-fried potatoes during transport and handling was investigated, examining shipments at both conventional and cryogenic temperatures. The study simulated in-transit vibrations and the effects of dropping or mishandling the product. The findings revealed that dropping the transported product was the primary cause of breakage.

Studies exploring the interaction between road profiles, vehicle size, speed, suspension characteristics, and the dynamic behaviour of the load were presented in Schoorl and Holt [22] and Schoorl and Holt [23], where similar methodologies were employed. These studies found that, for loads of fruits and vegetables, there is a significant interaction between the energy absorbed by the cargo and the movement of the vehicle body.

A few studies have adopted methodologies that integrate deep learning models to classify damage levels inflicted on products. The study in Ding et al. [24] investigated the effects of rough handling activities, such as dropping, kicking, and throwing, on the internal acceleration of cargo. Based on the collected acceleration data, the authors developed a Convolutional Neural Network (CNN) to characterize the damage imposed on products. A similar methodology was proposed in Todisco and Mao [25] focused on classifying damage levels in electronic assemblies subjected to high-acceleration mechanical shocks. The test product underwent six impacts using a drop tower; then, a Convolutional Variational Auto-Encoder (CVAE) model was developed, which enabled the detection of three distinct damage levels on the test product.

The studies presented in Emenike et al. [26] and Yu et al. [27] followed a unique approach by utilizing real-time sensor measurements to gather critical data on the condition of freight throughout its transportation journey.

The study proposed in Emenike et al. [26] explored the use of temperature sensors in the transportation of perishable goods by developing Neural Network models. These models were designed to predict in-container temperatures based on measurements taken at the periphery of the containers, thereby providing a potential variable of interest in addressing cargo loss issues. While the work in Yu et al. [27] also utilized real-time sensors, the suggested implementation pursued a different objective from that of the work in Emenike et al. [26]. In this study, acceleration sensors were embedded within the freight to facilitate data collection. These data were then transmitted to an external device where they were analysed using a pre-developed C4.5 Decision Tree model combined with a clustering model to classify the status of the freight.

This systematic review highlights a significant research gap: the lack of specific studies focusing on preventing pallet collapse in the glass industry during transportation. Therefore, the current study aimed to address this gap by applying machine learning techniques

to predict and mitigate such incidents, thereby enhancing the safety and efficiency of glass product transportation.

3. Methodology

As one can conclude from the previous section, the literature reveals several gaps in addressing specific challenges related to pallet stability in logistics, particularly in the glass industry. While general research has been focused on broad aspects of cargo handling, such as route optimization and securing techniques, detailed studies on pallet stability remain sparse. Existing works often lack granularity in predicting pallet collapse and fail to account for the unique challenges posed by fragile goods. Additionally, previous research has not sufficiently integrated data-driven approaches or machine learning techniques to address these specific issues effectively. Therefore, this study aims to fill these gaps by providing targeted predictions and strategies to mitigate pallet instability mainly focused on the glass industry.

The primary objective of this research was twofold: firstly, to develop and apply a predictive model capable of identifying potential incidents of pallet collapse using advanced data analytics and machine learning techniques; secondly, to devise and recommend targeted preventative strategies based on the insights derived from these predictive models to mitigate the risk of cargo loss. This approach involved analysing a range of variables to identify the key factors that significantly contribute to pallet collapse. By pinpointing these critical variables, the study aimed to enable proactive measures and preventive actions that effectively reduce the likelihood of such incidents. To achieve these objectives, this study employed a systematic approach guided by the CRISP-DM methodology, as outlined by Wirth and Hipp [28], which facilitated a structured and rigorous execution of the data mining process through its six phases, each crucial to the development and implementation of the predictive models and strategies:

- Business Understanding: The study began by clearly defining the specific problem of pallet collapse within the glass packaging industry, understanding its implications for logistics and business operations. The key business goals were established, focusing on reducing cargo loss and improving supply chain efficiency through predictive analytics.
- Data Understanding: Relevant data from a leading glass manufacturer were collected, including variables related to pallet geometry, environmental conditions, and transportation details. The data were thoroughly explored to identify patterns and assess their quality, ensuring that the collected dataset was suitable for the modelling phase.
- Data Preparation: This phase involved cleaning and preprocessing the data to address any inconsistencies or missing values. Features relevant to predicting pallet collapse were carefully selected and engineered to enhance the models' predictive capabilities. The data were also split into training and testing sets to facilitate unbiased model evaluation.
- Modelling: In this phase, machine learning algorithms, specifically Decision Tree and Random Forest models, were applied to the enhanced data. The models were trained to predict the likelihood of pallet collapse based on the selected features. Hyperparameter tuning was conducted to optimize the performance of these models, focusing on metrics such as Precision, Recall, and F1-score.
- Evaluation: The models' performance was rigorously evaluated against the test data to ensure their effectiveness in predicting pallet collapse. The evaluation criteria were aligned with the study's objectives, emphasizing the models' ability to generalize and accurately predict unseen cases.
- Deployment: Finally, the insights gained from the predictive models were translated into actionable recommendations. These preventative strategies were designed to address the key risk factors identified in the analysis, such as optimizing pallet stacking patterns and enhancing packaging materials.

Subsequently, the following sections provide a detailed and comprehensive description of the work completed across the various sequential stages of the CRISP-DM methodology.

3.1. Business Understanding

In the Business Understanding stage, the focus was on gaining a deep understanding of the study's scope, objectives, and constraints within the business context. Targeted research and meetings with relevant stakeholders in the field were conducted to gather valuable insights and establish clear lines of communication. Through these interactions, factors that could potentially contribute to pallet collapse during transportation were precisely identified. These factors were then selected as key variables for further investigation in subsequent stages of the project. The insights obtained from these discussions provided clear direction for the collection of necessary data and informed the understanding of the variables critical to addressing the identified problem effectively.

3.2. Data Understanding

Given that the problem at hand entailed a binary classification task, distinguishing between instances with a high probability of experiencing issues related to collapsed pallets and those without, it was imperative to gather transportation information about both scenarios. This involved collecting data on situations that resulted in pallet collapse incidents as well as instances where no problems occurred during transportation.

To ensure comprehensive data coverage, information on transport, logistics, and product characteristics was gathered from a multinational company specializing in the production and distribution of glass containers for the food and beverage industries, Table 1. Additionally, the values corresponding to the average daily temperature at the delivery point were obtained from Meteostat (see https://meteostat.net/pt/, accessed on 21 March 2024), a reputable meteorological data provider.

Table 1. Collected variables and respective descriptions.

Variable	Description
Material_Ref	Code of the shipped material
Destination_Country	Transport destination country
Distance	Distance between the origin and destination points, in kilometres
Temperature_Delivery_Date	Average daily temperature at the delivery point, in Celsius degrees
Delivery Date - Picking Date	Duration, in days, that the material remains inside the vehicle before the commencement of transport
Warehousing_Time	Number of days of the material inside the warehouse
External_Warehouse(Y/N)	Binary variable taking value 1 (one) if the material was shipped from an external warehouse, and 0 (zero) otherwise
Carrier_ID	Internal code assigned to the carrier responsible for the transport
SAP_Customer_number	Internal code designated for the customer which is to be filled out by the transport
WOOD_PALLET	Specific type of wood pallet being used
LAYER_SEPARATOR	Specific type of layer separator being used in the pallet
Foil_thickness	Thickness, in millimetres, of the foil wrapping the pallets
Larger_Side	Measurement, in millimetres, from the leftmost container to the rightmost container along the largest side of the pallet
Smaller_side	Measurement, in millimetres, from the leftmost container to the rightmost container along the smallest side of the pallet
Container_Diameter	Maximum diameter, in millimetres, of the material
Container_Bearing_Diameter	Diameter, in millimetres, of the bottom surface of the material
Container_Weight	Weight, in grams, of the material
Container_Height	Height, in millimetres, of the material
Pallet_Height	Height, in millimetres, of the pallet
Truck_Weight	Weight, in kilograms, of the vehicle
Complaint (Y/N)	Target binary variable, taking a value of 1 (one) if the transportation resulted in a problem related to collapsed pallets, and a value of 0 (zero) if no such problem occurred

The data collection process spanned from 2017 to 2023 and led to a highly imbalanced dataset. Specifically, the dataset included over 1 (one) million instances from the majority class, representing situations where no problems were encountered. In contrast, the minority class, which denoted instances of collapsed pallets, comprised only 526 examples.

In the context of this specific problem, the imbalance observed in the dataset is inherent to the nature of the problem itself, rather than being a result of any shortcomings in the data collection process. Specifically, the frequency of transportation instances without any collapsed pallets far exceeds those where such incidents were observed.

3.3. Data Preparation

The third step of the CRISP-DM methodology, data preparation, is crucial in data mining studies. This phase involves a series of techniques to transform raw data into a format that is consistent, well organized, and suitable for analysis and modelling.

Raw data frequently contains various issues, such as inconsistencies, errors, missing values, and outliers, which can negatively impact the reliability of machine learning models' outputs. Therefore, undertaking data preparation tasks is essential to remove inconsistencies and enhance the accuracy and reliability of the models. This process not only ensures the quality of the data but also enables a more accurate understanding of the original problem being addressed.

3.3.1. Handling Missing Values

Missing values, defined by the absence or unavailability of data for certain variables or observations within a dataset, can arise from a variety of factors such as errors during data collection or data that are inherently unobtainable. The presence of missing values significantly impacts the trustworthiness and accuracy of data analysis and modelling. Therefore, effectively addressing missing values is of paramount importance in the field of data mining.

According to [29], the occurrence of missing values can be classified into three distinct types:

- Missing Completely at Random (MCAR) refers to situations where the occurrence of missing values for a particular attribute is completely unrelated to any other data, whether observed or missing. In other words, the probability of an instance having a missing value is independent of the values of any variables, whether they are recorded or missing.
- Missing at Random (MAR) indicates that the occurrence of missing values is related to the observed data but not to the values that are missing. The probability of a value being missing is dependent on the values of other variables within the dataset. MAR suggests that the missingness can be systematically estimated or predicted based on the available information from other observed variables, thus allowing for a structured approach to handle these missing values in analyses.
- Not Missing at Random (MAR) indicates a scenario where the likelihood of an instance having a missing value for a particular attribute is influenced by the value of that attribute itself. In other words, the occurrence of missing values is dependent on the specific values that are missing. This dependency suggests that the missing data are related to inherent qualities of the data, making the handling of these missing values particularly challenging, as the reasons behind the missingness need to be explicitly modelled or accounted for in the analysis.

The investigation conducted to identify missing values in the collected dataset revealed that these were predominantly due to gaps in the company's records. Notably, missing values were consistently linked to specific materials, with the same "Material_Ref" showing identical missing attributes across all related entries. This pattern strongly suggests that the missing values are directly related to the inherent characteristics of the materials, stemming from an absence of essential information in the company's records.

This consistent occurrence underscores the need to address these gaps to improve data completeness and reliability.

In the scenario where missing values are found exclusively in the "Foil_thickness" column for certain materials, it becomes apparent that the occurrence of missing values is dependent on the characteristics of the material attribute itself. This observation indicates that the missing data in the dataset are not randomly distributed but are instead directly related to the attributes of the materials. Consequently, this pattern of missingness is formally classified as NMAR, where the probability of missing data is influenced by the values that are missing. This classification is crucial for informing the appropriate strategies for handling missing data in the analysis.

The only column in the dataset that exhibited missing values was the "Foil_thickness" column, which stores the micron measurement of the foil wrapping for each pallet. As shown in Table 2, the proportion of missing values within the "Foil_thickness" column, compared to the total number of observations in the dataset, was found to be low. This indicates that, while missing data are present, they affect only a small fraction of the overall dataset.

Table 2. Missing value incidence in "Foil_thickness" attribute.

Number of Missing Values	Number of Observations	Incidence
45	6081	0.74%

Based on the findings in Strike et al. [30] and Ren et al. [31], it can be assumed that, in datasets with a minimal occurrence of missing data, typically, when the missing rate is below 10% or 15%, excluding missing data does not significantly impact the outcomes of data mining or analysis. In this specific case, where the proportion of missing values was only 0.74%, the decision was made to eliminate observations with missing values from the dataset. As a result, the total number of observations decreased from 6081 to 6036. This approach ensured the integrity of the analysis while maintaining the robustness of the dataset.

3.3.2. Handling Outliers

Since outliers can significantly deviate from the majority of the dataset, they often exert an undue influence on the interpretation of relationships among variables. Therefore, identifying and properly addressing outliers is crucial in data analysis and modelling.

Box plots were used to analyse the numerical variables in the dataset. According to Aguinis et al. [32], a box plot is a straightforward graphical technique for identifying outliers. It provides a summary of a variable's distribution, showing the minimum value (excluding outliers), the lower quartile (Q1), the median (Q2), the upper quartile (Q3), and the maximum value (excluding outliers). In a box plot, outliers are identified as data points that lie beyond the whiskers, which extend from the minimum to the maximum value within the range of typical non-outlier data.

Box plot analysis of the numerical variables revealed outliers in the "Foil_thickness" variable, as depicted in Figure 2. This visualization clearly illustrates points that fall outside the typical range, offering a comprehensive view of the distribution and the presence of outliers for this variable.

The presence of outliers in the dataset can be definitively attributed to data entry errors. This conclusion is based on the recorded values for foil thickness associated with the outlier observations (0.023 micron), which do not align with the company's specified foil material specifications. Therefore, these outliers are considered error outliers, reflecting significant deviations from expected values due to inaccuracies in the data collection process [32].

The appropriate approach for addressing error outliers is to either correct the data points to their accurate values or remove the corresponding observations from the dataset [33]. Since the "Foil_thickness" variable had only 32 outlier observations, accounting for 0.53% of

the dataset, it was decided to exclude these observations. As a result, the total number of observations decreased from 6036 to 6004.

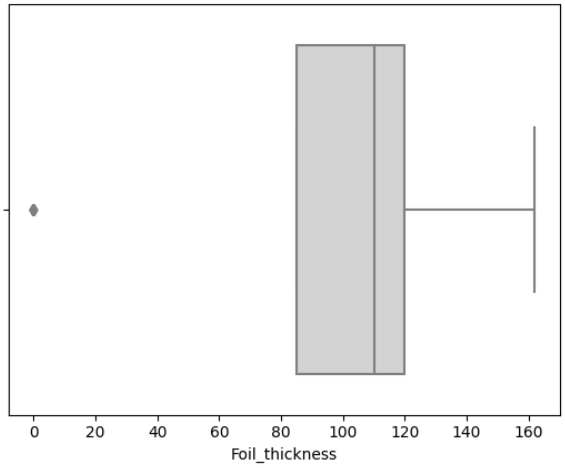

Figure 2. Box plot depicting the distribution of the "Foil_thickness" variable.

3.3.3. Feature Selection

Feature selection involves carefully choosing a subset of relevant features while discarding those that are irrelevant or redundant for the model built. This process has been shown to significantly improve the efficiency of prediction tasks and enhance the used machine learning model's performance. By eliminating extraneous features, feature selection not only enhances the clarity of the results but also streamlines the prediction process [34]. Prioritizing the most influential features facilitates a clearer understanding of the underlying relationships between the predictors and the target variable, aiding in the extraction of meaningful insights and enabling more accurate decision making based on the model's output.

Feature selection offers numerous benefits that significantly enhance data modelling. Firstly, it improves machine learning models' performance by reducing the risk of overfitting and minimizing the impact of irrelevant or noisy features. This ensures that the model captures the most relevant patterns and relationships within the data, leading to improved predictive accuracy. In addition to enhancing the model's performance, feature selection offers practical benefits. By excluding irrelevant features, it reduces the computational burden, leading to faster model training and improved overall computational efficiency. This streamlines data mining processes, enabling quicker analysis and more efficient decision making.

After considering the available options for feature selection, the decision was made to employ a filter method. In the realm of feature selection, there are two primary approaches: filter methods and wrapper methods. Filter methods rely solely on the intrinsic characteristics of the data to identify relevant features, independent of any specific predictor. In contrast, wrapper methods incorporate predictor optimization as part of the feature selection process [35]. In this particular case, the filter method was chosen due to its computational efficiency, making it the preferred approach for this study.

To evaluate the relationship between the independent variables and the dependent target variable, correlation tests were conducted. Two separate tests were employed, depending on the nature of the independent variable, whether it was numerical or categorical.

A chi-square independence test was performed to evaluate the correlation between the categorical independent variables and the binary target variable. This statistical test aims to determine whether there is a relationship between two categorical variables by examining whether the observed frequency distribution significantly differs from the expected distribution under the assumption of independence. The calculation of the chi-square (χ^2) statistic is given by the following:

$$\sum \chi^2_{i-j} = \frac{(O-E)^2}{E}, \quad (1)$$

where O is the observed value, E the expected value, χ^2_i the i cell's chi-square value, and $\sum \chi^2_{i-j}$ the sum of all the cell chi-square values with $i-j$ representing all the cells, i.e., from the first cell (i) to the last (j) cell.

Table 3 provides a summary of the χ^2 value and its corresponding p-value that together assess the relationship between each independent categorical variable and the correspondent target variable.

Table 3. Chi-square independence test summary: independent variables vs. "Complaint (Y/N)".

Variable	χ^2	p-Value
Material_Ref	2603.59	3.05×10^{-78}
Destination_Country	249.03	1.92×10^{-46}
External_Warehouse(Y/N)	0.18	0.67
Carrier_ID	767.33	2.74×10^{-41}
SAP_Customer_number	1679.51	5.49×10^{-48}
WOOD_PALLET	274.67	4.38×10^{-46}
LAYER_SEPARATOR	163.29	6.52×10^{-32}

Based on the results presented in Table 3, there is insufficient statistical evidence to conclude that the variable "External_Warehouse(Y/N)" is not independent of the target variable "Complaint (Y/N)". As a result, the variable "External_Warehouse(Y/N)" was excluded from further analysis.

A similar test was conducted to examine potential correlations between the remaining independent categorical variables that were not excluded following the initial analysis with the target variable "Complaint (Y/N)". The results, presented in Table 4, revealed a statistically significant correlation between "Material_Ref", "Carrier_ID", "SAP_Customer_number", and the variable "Destination_Country". As a result, the three aforementioned variables were excluded from further analysis, while "Destination_Country" was retained for subsequent analysis.

Table 4. Chi-square independence test summary between the remainder categorical variables and "Destination_Country".

Variable	χ^2	p-Value
Material_Ref	165,937.62	0.00
Carrier_ID	54,922.74	0.00
SAP_Customer_number	251,841.51	0.00
WOOD_PALLET	0.83	0.061
LAYER_SEPARATOR	0.77	0.082

To evaluate the correlation between a continuous numeric variable and a binary target variable, the point-biserial correlation coefficient was computed. This coefficient represents Pearson's product-moment correlation specifically designed for scenarios where one variable is dichotomous (binary) while the other is continuous [36].

The point-biserial correlation coefficient, denoted as r_{pb}, ranges between -1 and 1 (one). The sign of the coefficient indicates the direction of the relationship between the variables. A positive value signifies a positive association, implying that higher values of the continuous variable tend to be associated with the presence of the binary variable. Conversely, a negative value suggests a negative association, where higher values of the continuous variable are more likely to be linked with the absence of the binary variable.

The magnitude, i.e., the absolute value, of r_{pb} indicates the strength of the relationship, with values closer to 1 (one) signifying a stronger association, while values closer to 0 (zero) indicate a weaker relationship, according to the following:

$$r_{pb} = \frac{\bar{Y_1} - \bar{Y_0}}{\bar{s_y}} \times \sqrt{\frac{N_1 \times N_0}{N \times (N-1)}}, \tag{2}$$

where $\bar{Y_0}$ and $\bar{Y_1}$ are the mean of the continuous observations coded 0 (zero) and 1 (one), respectively, N_0 and N_1 the number of observations coded 0 (zero) and 1 (one), respectively, N the total number of observations ($N_0 + N_1$), and $\bar{s_y}$ the standard deviation of all the continuous observations, given as follows:

$$\bar{s_y} = \frac{\sum Y^2 - \frac{(\sum Y)^2}{N}}{N-1}. \tag{3}$$

Table 5 summarizes the point-biserial correlation coefficient values (r_{pb}) and their corresponding p-values, which assess the relationship between the independent continuous variables and the target variable. These values offer insights into the strength and significance of the associations. Additionally, Figure 3 displays a point-biserial scatter plot illustrating the relationship between the continuous variable "Temperature_Delivery_Date" and the target variable "Complaint (Y/N)", providing a visual representation of their relationship.

Table 5. Point-biserial correlation coefficient test summary: independent variables vs. "Complaint (Y/N)".

Variable	r_{pb}	p-Value
Distance	0.14	5.80×10^{-28}
Temperature_Delivery_Date	0.12	1.77×10^{-19}
Delivery Date_Picking Date	0.02	0.15
Warehousing_Time	−0.002	0.88
Foil_thickness	0.14	3.36×10^{-27}
Larger_Side	−0.03	0.04
Smaller_side	−0.05	1.41×10^{-4}
Container_Diameter	−0.11	4.65×10^{-19}
Container_Bearing_Diameter	−0.08	1.85×10^{-9}
Container_Weight	−0.01	0.42
Container_Height	0.06	4.89×10^{-7}
Pallet_Height	−0.02	0.23
Truck_Weight	0.01	0.64

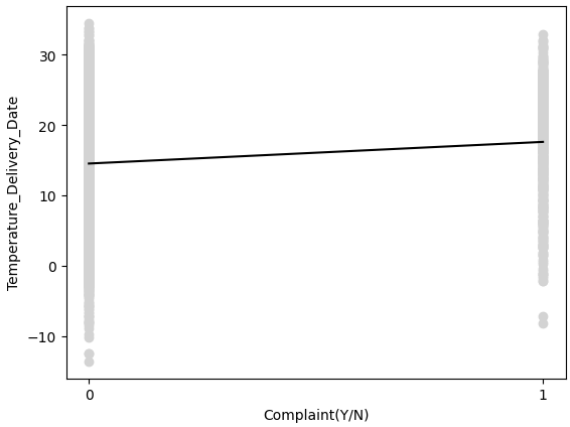

Figure 3. Point-biserial scatter plot "Temperature_Delivery_Date" vs. "Complaint (Y/N)".

Based on the findings, one can conclude that there is insufficient statistical evidence to establish a significant correlation between the variables "Delivery Date - Picking Date", "Warehousing_Time", "Container_Weight", "Pallet_Height", and "Truck_Weight", and the target variable. Consequently, these variables were excluded from further analysis due to the lack of significant evidence.

3.3.4. Data Discretization

Data discretization is a fundamental data preprocessing technique that involves converting continuous data into discrete or categorical forms by partitioning a continuous variable into a limited number of intervals or categories.

The research described in Peker and Kubat [37] emphasized the importance of data discretization in reducing the dimensionality and complexity of continuous variables. This reduction can lead to improved computational efficiency, enhanced model generalization, and a better understanding of underlying patterns and relationships within the data. Considering these benefits, the decision was made to discretize the variable "Distance" into three distinct bins, as shown in Table 6. As a result of this discretization process, a new variable named "Trip" was created.

Table 6. Discretization of the variable "Distance" and consequent creation of the variable "Trip".

Distance	Trip
<150	short
$[150, 2000]$	medium
>2000	long

To assess the correlation between the newly created variable "Trip" and the target variable, a Spearman's rank correlation test was conducted. The Spearman's rank correlation coefficient (r_S) was used to measure the strength and direction of the relationship between two variables, whether continuous or ordinal. Given that "Trip" is an ordinal variable, the Spearman's rank correlation test was deemed the appropriate method for this analysis.

The Spearman's rank correlation focuses on the rank order of the two datasets under study. Its coefficient can take values between -1 and 1 (one), where a value of 1 (one) indicates a perfect monotonic increasing relationship, while a value of -1 indicates a perfect monotonic decreasing relationship. The closer the absolute value of r_S is to 1 (one), the stronger the monotonic relationship between the datasets [38].

The calculation of the Spearman's rank correlation coefficient (r_S) is given by the following:

$$r_S = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \times (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \times \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \quad (4)$$

where x_i is the rank of the measurement for the i-th sample of the X ("Trip") set, $\bar{x}$ the sample mean of X, y_i the rank of the measurement for the i-th sample of the Y ("Complaint (Y/N)") set, $\bar{y}$ the sample mean of Y, and n the datasets' size.

Based on the findings presented in Table 7, one can conclude that there is a statistically significant correlation between the variables "Trip" and "Complaint (Y/N)". This result supports the effectiveness of the discretization applied to the initial continuous values, as it successfully captured the underlying relationship between the variables.

Table 7. Spearman's rank correlation coefficient test: "Trip" vs. "Complaint (Y/N)".

Variable	r_s	p-Value
Trip	0.07	4.14×10^{-7}

In Table 8 the variables to be modelled after the data preparation phase are presented.

Table 8. Final variables to be included in the machine learning models after data preparation.

Variable
Destination_Country
Trip
Temperature_Delivery_Date
WOOD_PALLET
LAYER_SEPARATOR
Foil_thickness
Larger_Side
Smaller_side
Container_Diameter
Container_Bearing_Diameter
Container_Height
Complaint (Y/N)

3.4. Modelling

To tackle the task of predicting pallet collapse resulting from specific transportation activities, a thorough exploration of different modelling techniques was undertaken. In particular, Decision Tree and Random Forest models were considered and evaluated. These models were chosen for their strong interpretability, effectiveness in handling imbalanced data, ability to highlight feature importance, and overall performance.

Decision Tree allows for a clear visualization of decision-making processes, which is crucial for understanding how specific variables influence the prediction of pallet collapses during transportation. On the other hand, Random Forest is particularly well suited for predicting rare events, such as pallet collapse in an imbalanced dataset. By averaging multiple decision trees, they effectively reduce overfitting and enhance generalization to unseen data, making them highly effective in this context. Also, both Decision Tree and Random Forest provide mechanisms for assessing feature importance, allowing one to identify and rank the factors that contribute to pallet collapse and guide the development of effective preventive strategies. Finally, prior studies, such as Panchapakesan et al. [17] and Wu et al. [15], have demonstrated the effectiveness of Decision Tree and Random Forest models in similar classification tasks.

Although there are many other machine learning methods, such as Neural Networks, Multiple Discriminant Analysis, and Logistic Regression, which could also be applied, Decision Tree and Random Forest were chosen for their balance of performance, interpretability, and ease of use. In particular, Neural Networks, while powerful, often require larger datasets and more computational resources and can be challenging to interpret, which was not ideal for the goals of this study.

For all approaches addressing the classification problem, Grid Search Cross-Validation (GridSearchCV) was utilized. As described by Yasin et al. [39], GridSearchCV partitions the parameter space into a grid and systematically evaluates every possible combination of hyperparameters. It then selects the combination that achieves the best performance according to the specified evaluation metrics.

The dataset was initially divided into two distinct subsets: a training set and a test set. In this particular scenario, the training set was created by randomly selecting 80% of the original data, while the remaining 20% was allocated to the test set.

As previously mentioned, the hyperparameters to be tuned were specified, and a list of hyperparameter combinations to be evaluated was generated.

The training data underwent k-fold cross-validation, a method where the dataset is divided into k equally sized folds. In this case, k was set to 10, resulting in the creation of 10 folds. During each iteration of the cross-validation process, one fold is used as the validation set to assess the performance of different hyperparameter combinations, while the remaining $k - 1$ folds are combined to form the training set. This iterative process is repeated k times, providing a thorough evaluation of the model's performance. By using

multiple train–test splits, the risk of bias from a single split is reduced, thus improving the reliability of the model evaluation.

Given the imbalanced distribution of the target variable, as illustrated in Figure 4, where the target variable "Complaint (Y/N)" distribution is shown, it was essential to address this imbalance to prevent the model from biasing towards predicting only the majority category. To address this issue, various data balancing techniques were implemented to ensure a more equitable representation of both minority and majority categories within the dataset. These balancing techniques were applied within each iteration of the cross-validation process, specifically to the training data.

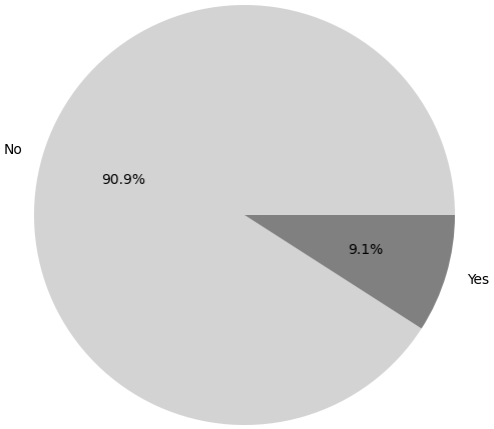

Figure 4. Target variable "Complaint (Y/N)" distribution in the studied dataset.

Random oversampling was selected to address the class imbalance in the dataset, specifically targeting the minority class within the training data. This technique involves randomly selecting minority examples, denoted as S_{min}, from the original training set, S, and adding a set E of replicated examples from the minority class to augment the training set. By applying this approach, the total number of examples in the minority class, denoted as S_{min}, is augmented by $|E|$, thereby increasing the representation of the minority class in the training set. Consequently, the class distribution of the overall training set is adjusted to account for the added replicating examples [40]. Four distinct variations of the random oversampling method were utilized, each addressing different class distributions of the target variable. The explored four variations were as follows:

- $\frac{|S_{min}|}{|S|} = 0.25$;
- $\frac{|S_{min}|}{|S|} = 0.3$;
- $\frac{|S_{min}|}{|S|} = 0.4$;
- $\frac{|S_{min}|}{|S|} = 0.5$.

A comprehensive analysis of the developed machine learning models is provided in the following sections, discussing their characteristics and capabilities in detail to identify the most suitable approach for addressing the problem under study.

3.4.1. Decision Tree

The Decision Tree algorithm begins by using a collection of examples to construct a tree-like structure designed to classify new cases. Each example is defined by a set of attributes, which may be numeric or categorical, and is assigned a label indicating its class.

In the Decision Tree model, each internal node performs a test on an attribute, and the result of this test determines the path to follow. If the test outcome is true, the case moves to the left branch; if false, it moves to the right branch. The leaf nodes of the tree represent class labels rather than further tests [41].

In this study, the implementation of the Decision Tree model involved careful tuning of hyperparameters and selecting appropriate values for exploration during the tuning process. The model was implemented using the R programming language, specifically through the *rpart* package, which provides tools for fitting and analysing classification and regression problems [42]. This package offers robust functionality for handling various types of variables, including categorical and continuous data, and includes mechanisms to control the complexity of the resulting tree model through parameters that facilitate the pruning of unnecessary branches.

One of the tuned hyperparameters was the splitting criterion, denoted as *split* in the package. The options for the splitting criteria were limited to Information Gain and the Gini index. The Decision Tree model developed here using Information Gain as the splitting criterion can be seen as a hybrid of the ID3 and C4.5 algorithms. This is because it combines the Information Gain criterion from ID3 with the capability to handle various variable types and the built-in pruning mechanism of C4.5.

The complexity parameter (*cp*) was another hyperparameter tuned in the model. This parameter acts as a threshold to determine whether a split improves the overall accuracy sufficiently to justify its inclusion. To identify the optimal value for *cp*, a range from 0.001 to 0.01 was tested, with increments of 0.001. Pruning, guided by the complexity parameter, is essential for preventing overfitting by reducing the complexity of the tree and ensuring better generalization to unseen data.

The final hyperparameter tuned in the Decision Tree model was *minbucket*, which specifies the minimum number of observations required in any terminal leaf node. This parameter determines when to cease splitting a node based on the number of observations it contains. Together with the *cp* parameter, *minbucket* helps control the complexity of the tree and mitigate overfitting. During the tuning process, *minbucket* was tested across a range of values from 2 to 100, with increments of 1 (one). Additionally, the *minsplit* parameter, which defines the minimum number of observations needed in a node for a split to be attempted, was set to three times the value of *minbucket*.

In Table 9, the hyperparameters that were tuned in the Decision Tree model are presented, along with the range of values that were tested for each parameter.

Table 9. Decision Tree model's tuned hyperparameters and correspondent values.

Hyperparameter	Range of values
split	Information Gain or Gini index
cp	0.001 to 0.01, with increments of 0.001
minbucket	2 to 100, with increments of 1 (one)

3.4.2. Random Forest

In Breiman [43], the concept of Random Forest was introduced as an ensemble of tree predictors, where each tree's outcome depends on a randomly sampled vector that is independently and identically distributed across all trees in the ensemble. Random Forest is an ensemble method used for both classification and regression tasks. It combines multiple Decision Trees to make predictions, and the final prediction is obtained by aggregating the predictions from all individual trees, employing a majority voting scheme for classification tasks.

Similarly to the Decision Tree algorithm, the Random Forest model was implemented using the R programming language. The *randomForest* package was employed for this purpose, providing comprehensive tools for building Random Forest models to tackle both classification and regression problems [44].

Consistent with the Decision Tree model, the splitting criterion, referred to as *splitrule* package, was one of the hyperparameters optimized in the Random Forest implementation. The values considered for the splitting criterion were Information Gain and the Gini index, as the Gain Ratio option was not available in this context, similar to the *rpart* package.

The number of trees (*ntree*) was another hyperparameter tuned in the model. This parameter determines the size of the ensemble and significantly impacts the model's predictive performance and generalization ability. However, increasing the number of trees excessively can lead to diminishing returns or higher computational costs. In this study, the tested values for *ntree* ranged from 100 to 120, with increments of 1 (one).

The feature selection process in this study utilized the Forest-RI approach, which involves random input selection for splitting at each node. Accordingly, hyperparameter tuning was performed for the variable *mtry*, which denotes the number of variables randomly sampled as candidates for each split. The values tested for *mtry* ranged from 1 to 4, with increments of 1 (one).

The minimum size of terminal nodes, represented as *nodesize* in the *randomForest* package, was the final hyperparameter tuned in the study. This parameter determines when to stop splitting a node based on the minimum number of examples required in any terminal leaf. It helps regulate the complexity of the trees and reduces the risk of overfitting the training data. During the tuning process, *nodesize* values were tested from 2 to 20, with increments of 1 (one).

Table 10 details the hyperparameters tuned in the Random Forest model, including the range of values tested for each parameter during the tuning process.

Table 10. Random Forest model's tuned hyperparameters and correspondent values.

Hyperparameter	Range of Values
splitrule	Information Gain or Gini index
ntree	100 to 120, with increments of 1 (one)
mtry	1 to 4, with increments of 1 (one)
nodesize	2 to 20, with increments of 1 (one)

4. Results and Discussion

The final two stages of the CRISP-DM methodology are presented in this section. First, the models built in the previous phase are assessed. Then, a set of recommendations for preventative strategies based on predictive modelling results are discussed.

4.1. Evaluation Results

Within the CRISP-DM methodology, the evaluation phase holds significant importance as it allows for a comprehensive assessment of the quality and effectiveness of the developed machine learning models. As already mentioned, to address the challenge of predicting pallet collapse, this study employed Decision Tree and Random Forest due to their proven effectiveness in handling classification tasks, particularly in scenarios with class imbalances. Decision Tree was chosen for its simplicity and interpretability, making it easier to understand the decision-making process. Random Forest, an ensemble method, was selected for its ability to reduce overfitting and improve predictive accuracy by averaging the results of multiple trees.

During this phase, the primary objective was to assess the models' generalization capabilities, specifically their accuracy in predicting unseen data. To achieve this, 20% of the dataset was reserved during the initial split as a dedicated test set. This reserved set was used to evaluate the models' performance on data that was not seen during training. By employing this separate test dataset, the evaluation phase ensured an unbiased assessment of each model's predictive performance and its ability to generalize beyond the training data.

The evaluation of the developed models, which addressed a binary classification problem, involved constructing a confusion matrix for each model. The confusion matrix provides a detailed view of a model's accuracy, allowing for the calculation of various performance metrics. Given the objective of identifying pallets at risk of collapsing during transportation, minimizing false negatives is critical. At the same time, it is important to

avoid excessive false positives to prevent unnecessary alerts. The *F*1-*score* was chosen as the primary metric for optimization during the hyperparameter tuning process. The *F*1-*score* metric, which represents the harmonic mean of *Precision* and *Recall*, offers a balanced evaluation of the model's performance:

$$F1\text{-}score = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{5}$$

where *Precision* and *Recall* are calculated as follows:

$$Precision = \frac{TP}{TP + FP}, \tag{6}$$

$$Recall = \frac{TP}{TP + FN}, \tag{7}$$

and TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives.

To facilitate a comparison among the developed models, Table 11 was built, presenting a comprehensive overview of each model's performance. This table allows for an analysis of each model's strengths and weaknesses based on the *Precision*, *Recall*, and *F*1-*score* values.

To aid in interpreting Table 11, it is important to note the naming convention used: for example, "DecisionTrees7525" refers to the Decision Tree model with a class imbalance, where 75% indicates the prevalence of the majority class ("No") and 25% represents the minority class ("Yes"). This convention helps identify each model's specific configuration and clarifies the class distribution used during training.

Table 11. Models' performance taking into account the *Precision*, *Recall*, and *F*1-*score* metrics.

Model	*Precision*	*Recall*	*F*1-*score*
DecisionTrees7525	0.350	0.533	0.423
DecisionTrees7030	0.333	0.543	0.413
DecisionTrees6040	0.283	0.648	0.394
DecisionTrees5050	0.299	0.695	0.418
RandomForests7525	0.442	0.455	0.448
RandomForests7030	0.415	0.509	0.457
RandomForests6040	0.405	0.582	0.478
RandomForests5050	0.392	0.591	0.471

The confusion matrix for the model with the best performance, RandomForests6040, based on the *F*1-*score* results, is presented in Table 12. This matrix shows the results for the 1206 examples in the test set, facilitating the calculation of the metrics previously discussed.

Table 12. Confusion matrix for RandomForests6040, the best-performing model built.

	Actual Positive (1)	Actual Negative (0)
Predicted Positive (1)	64	94
Predicted Negative (0)	46	1002

Tables 13 and 14 present the optimal values of the hyperparameters that resulted in the maximum *F*1-*score* values during the tuning process for the Decision Tree and Random Forest models, respectively. Hyperparameter tuning involved an exhaustive grid search, where key parameters such as *cp*, *minbucket*, *splitrule*, *ntree*, and *mtry* were adjusted to optimize the *F*1-*score*. This tuning process ensured that the models were fine-tuned to strike the right balance between sensitivity (*Recall*) and precision (*Precision*).

Table 13. Optimal hyperparameters for the Decision Tree model to maximize *F*1-*score*.

Model	*Split*	*cp*	*Minbucket*
DecisionTrees7525	Information Gain	0.004	13
DecisionTrees7030	Information Gain	0.004	10
DecisionTrees6040	Information Gain	0.002	7
DecisionTrees5050	Information Gain	0.001	7

Table 14. Optimal hyperparameters for the Random Forest model to maximize *F*1-*score*.

Model	*Splitrule*	*Ntree*	*Mtry*	*Nodesize*
RandomForests7525	Information Gain	111	2	7
RandomForests7030	Gini index	106	2	16
RandomForests6040	Information Gain	103	2	19
RandomForests5050	Gini index	111	2	19

The analysis of the results clearly shows that the Random Forest model outperformed the Decision Tree model based on the *F*1-*score* metric, which was the primary metric of interest. The Random Forest model proved to be more effective at capturing the complexities of the classification problem, resulting in more accurate predictions than the Decision Tree model. This is in line with previous studies, such as Panchapakesan et al. [17], that demonstrate the superiority of Random Forest over other models such as the Decision Tree and Neural Network algorithms.

Several steps were taken to refine the dataset, including normalization of continuous variables, encoding of categorical variables, and handling missing data. These preprocessing steps significantly improved the models' performance, particularly the Random Forest model, which is sensitive to feature scaling and encoding. Also, the superior performance of the Random Forest model can be attributed to its ensemble nature, which aggregates the predictions of multiple decision trees to reduce variance and improve generalization. Unlike the single Decision Tree, the Random Forest model was able to capture complex interactions between features, leading to higher accuracy and a better balance between *Precision* and *Recall*.

Identifying the most informative features, which are critical for solving a classification problem, is essential in addressing the underlying challenge. In this context, the concept of Impurity Decrease was used to evaluate each feature's contribution to reducing data impurity during the decision-making splits in the Decision Tree model. This approach was applied within the RandomForests6040 model, which demonstrated superior performance.

The relative importance of features was assessed by calculating the average Impurity Decrease across all 103 trees in the model. The analysis identified "Destination_Country" as the most influential variable in the prediction task, followed closely by the "Container_Height", "Container_Diameter", and "Temperature_Delivery_Date" features. These findings are visually represented in Figure 5. The high importance of features like "Destination_Country" and "Container_Height" can be linked to the inherent logistics and handling challenges in different regions and the physical dimensions of the cargo, respectively. This insight is critical for logistics companies to understand where to focus their efforts in securing cargo during transit.

Comparing the results of a study with the ones of the existing literature is crucial in research, as it helps validate and contextualize the findings within a broader knowledge framework. Evaluating the alignment or divergence of a new study's results with the ones of prior studies allows for an assessment of their consistency, reliability, and generalizability.

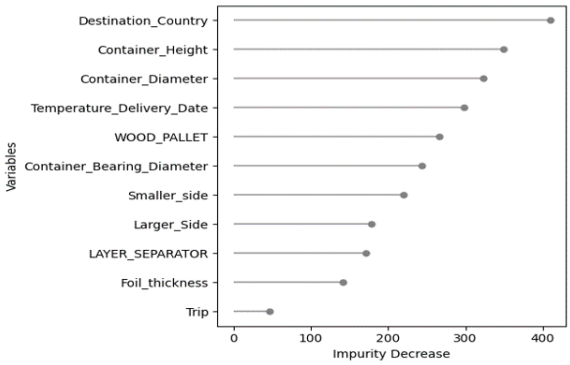

Figure 5. Feature importance, within the RandomForests6040 model, as measured by Impurity Decrease.

The findings of this study align with those of Panchapakesan et al. [17], as both identify the Random Forest model as the most effective approach for addressing the problem under study. However, a discrepancy arises in identifying the most influential feature for predicting complaints about shipping containers. While Panchapakesan et al. [17] identified the duration of container storage as the most informative variable, the current study found that the corresponding variable, "Warehousing_Time", did not show a statistically significant correlation with the target variable, "Complaint (Y/N)". This divergence in findings may be attributed to differences in the contexts of the two studies and the varying influence of business type, which could lead to different outcomes.

The research by Wu et al. [15] offered valuable insights into predicting cargo loss severity during transportation. Their study highlighted Transit Types, Product Categories, and Shipping Destinations as key features. While the dataset used in the current study lacked sufficient information to directly verify the findings related to Transit Types and Product Categories, the identification of Shipping Destinations as a significant feature aligns with their results. Consistent with Wu et al. [15], the current study found "Destination_Country" to be the most critical factor in predicting pallet collapses during transport. This agreement on the importance of Shipping Destinations strengthens the validity of the current study's conclusions.

This study makes a novel contribution to the existing literature by identifying new variables related to the geometry of the shipped product, specifically "Container_Height" and "Container_Diameter", as well as the average temperature recorded at the delivery location on the delivery date, denoted as "Temperature_Delivery_Date". These variables are found to be highly informative for predicting pallet collapse during transportation. This is the first study to recognize the significance of these variables in addressing the issue. Moreover, this research breaks new ground as the first to specifically focus on predicting pallet collapse during transportation. It also addresses the unique product-related challenges encountered in the glass industry, adding to its innovative approach.

4.2. Preventive and Mitigation Strategies

The predictive analysis conducted in this study reveals key insights into the factors contributing to pallet collapses during transportation. Therefore, based on the identified influential features, several preventative strategies can be implemented to reduce cargo loss.

Firstly, optimizing packaging specifications is crucial. The study identified both "Container_Height" and "Container_Diameter" as significant predictors of pallet collapse. To address this, companies should standardize packaging dimensions to ensure stability. By adhering to optimal container sizes and avoiding excessive height or diameter, the likelihood of imbalances that could lead to collapses is reduced. Additionally, reinforcing packaging through the use of double-walled or specially designed containers for high-risk shipments can further enhance stability. Testing packaging designs for their ability to withstand various transport conditions is also recommended.

Temperature control is another critical factor. The "Temperature_Delivery_Date" feature was found to impact the risk of pallet collapse significantly. To mitigate this risk, it is essential to implement measures for maintaining optimal temperature conditions during transport. This could involve utilizing temperature-controlled trucks or containers and ensuring regular monitoring to adhere to specified temperature ranges. Integrating temperature logging systems into the shipping process can provide real-time data, allowing for prompt corrective actions if temperature deviations occur.

Reviewing and adjusting shipping routes based on the analysis of the "Destination_Country" feature can also help reduce risks. It is important to analyse shipping routes and destination-specific factors that may contribute to increased risk, such as exposure to extreme weather conditions or rough terrains. Developing a risk assessment framework based on historical data for different destination countries can guide adjustments in shipping strategies and packaging solutions according to the risk profile of each destination.

Improving handling procedures is another vital strategy. Enhancing training programs for handling and loading procedures will ensure that personnel are well informed about best practices for securing and managing loads. This can significantly reduce the chances of pallet collapses. Additionally, investing in high-quality handling equipment and performing regular maintenance will prevent equipment failures during loading and unloading.

Leveraging predictive analytics for proactive measures is also recommended. Developing early warning systems that utilize predictive models to flag high-risk shipments based on identified features can trigger preventive actions, such as additional packaging or routing adjustments, before shipment dispatch. Continuous monitoring of predictive models as new data become available allows for real-time adjustments to strategies, improving overall risk management.

Finally, collaborating with industry partners can enhance these efforts. Sharing best practices and learning from others' experiences can lead to improved standards and innovative solutions. Engaging with stakeholders and partners can foster collaborative efforts to address common challenges in transportation and logistics.

By implementing the aforementioned strategies, businesses can proactively address the factors contributing to pallet collapses and effectively reduce cargo loss during transportation. The insights derived from predictive modelling will support informed decision making and enhance the efficiency and safety of the supply chain.

5. Conclusions

This study aimed to predict potential pallet collapse incidents using advanced data analytics proactively and to recommend preventative strategies based on these predictions to reduce the incidence of cargo loss. To achieve this goal, the CRISP-DM methodology was employed, which facilitated a systematic progression through six key stages that structured the research process and the development of the predictive tool.

Two primary modelling techniques were employed: the Decision Tree and Random Forest models. These models were selected for their proven effectiveness in predictive analytics within similar contexts. The hyperparameters of both models were meticulously fine-tuned through Grid Search Cross-Validation, and the class imbalance problem was addressed using random oversampling to balance the dataset between the majority and minority classes.

The conducted evaluation focused on $F1\text{-}score$, a metric chosen for its ability to harmonize the trade-off between *Recall* and *Precision*. The results demonstrated that the Random Forest model consistently outperformed the Decision Tree model, aligning with previous studies that highlight the Random Forest model as superior for this type of prediction.

Significantly, this study not only reaffirmed the importance of the destination country as a key predictor but also introduced novel insights into other influential factors. Variables such as container height, container diameter, and delivery date temperature emerged as crucial predictors of pallet collapse, marking an advancement in the literature.

Future work should aim to expand the scope of this research by incorporating additional data sources and variables to refine further the predictive models, as well as exploring alternative machine learning techniques. Additionally, investigating the real-time application of these models in operational settings could provide practical insights and further validate their effectiveness. Addressing these areas will not only advance the current research but also contribute to more robust and adaptive solutions for the glass industry and beyond.

Author Contributions: Conceptualization, J.M.R.S.T. and M.C.F.; methodology, F.C., J.M.R.S.T. and M.C.F.; validation J.M.R.S.T. and M.C.F.; formal analysis, F.C.; investigation, F.C.; writing—original draft preparation, F.C.; writing—review and editing, J.M.R.S.T. and M.C.F.; supervision, M.C.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets presented in this article are not readily available due to the company's confidentiality requirements.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Tkaczyk, S.; Drozd, M.; Kędzierski, Ł; Santarek, K. Study of the stability of palletized cargo by dynamic test method performed on laboratory test bench. *Sensors* **2021**, *21*, 5129.
2. Ferreira, G.F.O.; Sergio Dulcini, V.T. Topological optimization of a pallet motion blocking made of long-fiber reinforced thermoplastic composite. In Proceedings of the 6th Brazilian Conference on Composite Materials, Minas Gerais, Brazil, 14–18 August **2022**.
3. Burges, D. *Cargo Theft, Loss Prevention, and Supply Chain Security*; Butterworth-Heinemann: Oxford, UK, 2012.
4. Cui, H.; Qiu, J.; Cao, J.; Guo, M.; Chen, X.; Gorbachev, S. Route optimization in township logistics distribution considering customer satisfaction based on adaptive genetic algorithm. *Math. Comput. Simul.* **2023**, *204*, 28–42.
5. Nieoczym, A.; Caban, J.; Vrabel, J. The problem of proper cargo securing in road transport—Case study. *Transp. Res. Procedia* **2019**, *40*, 1510–1517.
6. Turbaningsih, O. The study of project cargo logistics operation: A general overview. *J. Shipp. Trade* **2022**, *7*, 24.
7. Nath, P.; Upadhyay, R.K. Reformation and optimization of cargo handling operation at Indian air cargo terminals. *J. Air Transp. Res. Soc.* **2024**, *2*, 100022.
8. Directive2014/47/EU of the European Parliament and of the Council of 3 April 2014 on the Technical Roadside Inspection of the Roadworthiness of Commercial Vehicles Circulating in the Union and Repealing Directive 2000/30/EC. Available online: https://eur-lex.europa.eu/eli/dir/2014/47/oj (accessed on 23 August 2024).
9. Hallak, K.; Abdallah, A. A Supervised Machine Learning Monitoring System for Vehicle-Railway Bridge Collision. In *Artificial Intelligence and Applications*; University of Lorraine: Vandœuvre-lès-Nancy, France, 2024; Volume 2.
10. Deng, W.; Cai, X.; Wu, D.; Song, Y.; Chen, H.; Ran, X.; Zhou, X.; Zhao, H. MOQEA/D: Multi-Objective QEA with Decomposition Mechanism and Excellent Global Search and Its Application. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 12517–12527.
11. Hasanvand, M.; Nooshyar, M.; Moharamkhani, E.; Selyari, A. Machine Learning Methodology for Identifying Vehicles Using Image Processing. In *Artificial Intelligence and Applications*; University of Lorraine: Vandœuvre-lès-Nancy, France, 2024; Volume 2.
12. Sun, Q.; Chen, J.; Zhou, L.; Ding, S.; Han, S. A study on ice resistance prediction based on deep learning data generation method. *Ocean Eng.* **2024**, *301*, 117467.
13. Page, M.J.; Moher, D.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ* **2021**, *372*, n160.
14. Tkaczyk, S. Classification of Damages of Palletized Loads in Road Transport and Its Impact on Environmental Protection. *Rocz. Ochr. Środowiska* **2022**, *24*, 457–471.
15. Wu, P.J.; Chen, M.C.; Tsau, C.K. The data-driven analytics for investigating cargo loss in logistics systems. *Int. J. Phys. Distrib. Logist. Manag.* **2017**, 47, 68–83.
16. Hashemi, R.R.; Le Blanc, L.A.; Rucks, C.T.; Shearry, A. A neural network for transportation safety modeling. *Expert Syst. Appl.* **1995**, *9*, 247–256.

17. Panchapakesan, A.; Abielmona, R.; Falcon, R.; Petriu, E. Prediction of container damage insurance claims for optimized maritime port operations. In Proceedings of the Advances in Artificial Intelligence: 31st Canadian Conference on Artificial Intelligence, Canadian AI 2018, Toronto, ON, Canada, 8–11 May 2018; Proceedings 31; Springer: Berlin/Heidelberg, Germany, 2018; pp. 265–271.
18. Panchapakesan, A.; Abielmona, R.; Petriu, E. Improving shipping container damage claims prediction through level 4 information fusion. *Int. J. Logist. Syst. Manag.* **2021**, *40*, 489–509.
19. Jarimopas, B.; Singh, S.P.; Saengnil, W. Measurement and analysis of truck transport vibration levels and damage to packaged tangerines during transit. *Packag. Technol. Sci. Int. J.* **2005**, *18*, 179–188.
20. Fadiji, T.; Coetzee, C.; Chen, L.; Chukwu, O.; Opara, U.L. Susceptibility of apples to bruising inside ventilated corrugated paperboard packages during simulated transport damage. *Postharvest Biol. Technol.* **2016**, *118*, 111–119.
21. Schlimme, D.; Ashby, B.; Turczyn, M.; Fowke, M.; Vo, Q. Damage to French-Fried Potatoes Caused by Simulated Transport and Handling Tests at Cryogenic Temperatures. *J. Food Sci.* **1984**, *49*, 217–220.
22. Schoorl, D.; Holt, J. Road-vehicle-load interactions for transport of fruit and vegetables. *Agric. Syst.* **1982**, *8*, 143–155.
23. Schoorl, D.; Holt, J. Verification and application of a model for predicting damage to horticultural produce during transport. *Agric. Syst.* **1985**, *16*, 67–83.
24. Ding, A.; Zhang, Y.; Zhu, L.; Du, Y.; Ma, L. Recognition method research on rough handling of express parcels based on acceleration features and CNN. *Measurement* **2020**, *163*, 107942.
25. Todisco, M.; Mao, Z. High-Rate Damage Classification and Lifecycle Prediction viaDeep Learning. In *Data Science in Engineering, Volume 9, Proceedings of the 39th IMAC, A Conference and Exposition on Structural Dynamics 2021*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 225–232.
26. Emenike, C.C.; Van Eyk, N.P.; Hoffman, A.J. Improving cold chain logistics through RFID temperature sensing and predictive modelling. In Proceedings of the 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), Rio de Janeiro, Brazil, 1–4 November 2016; pp. 2331–2338.
27. Yu, W.; Ye, W.Z.; Tateno, S. Real time logistics monitoring system of packages during transportation using decision tree combined with clustering method. In Proceedings of the 2017 International Automatic Control Conference (CACS), Pingtung, Taiwan, 12–15 November 2017; pp. 1–6.
28. Wirth, R.; Hipp, J. CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, Manchester, UK, 11–13 April 2000; Volume 1, pp. 29–39.
29. Little, R.J.; Rubin, D.B. *Statistical Analysis with Missing Data*; John Wiley & Sons: Hoboken, NJ, USA, 2019; Volume 793.
30. Strike, K.; El Emam, K.; Madhavji, N. Software cost estimation with incomplete data. *IEEE Trans. Softw. Eng.* **2001**, *27*, 890–908.
31. Ren, L.; Wang, T.; Sekhari Seklouli, A.; Zhang, H.; Bouras, A. A review on missing values for main challenges and methods. *Inf. Syst.* **2023**, *119*, 102268.
32. Aguinis, H.; Gottfredson, R.K.; Joo, H. Best-practice recommendations for defining, identifying, and handling outliers. *Organ. Res. Methods* **2013**, *16*, 270–301.
33. Dash, C.S.K.; Behera, A.K.; Dehuri, S.; Ghosh, A. An outliers detection and elimination framework in classification task of data mining. *Decis. Anal. J.* **2023**, *6*, 100164.
34. Yu, L.; Liu, H. Efficiently handling feature redundancy in high-dimensional data. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 24–27 August 2003; pp. 685–690.
35. Sahebi, G.; Movahedi, P.; Ebrahimi, M.; Pahikkala, T.; Plosila, J.; Tenhunen, H. GeFeS: A generalized wrapper feature selection approach for optimizing classification performance. *Comput. Biol. Med.* **2020**, *125*, 103974.
36. Kornbrot, D. Point biserial correlation. In *Wiley StatsRef: Statistics Reference Online*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2014.
37. Peker, N.; Kubat, C. Application of Chi-square discretization algorithms to ensemble classification methods. *Expert Syst. Appl.* **2021**, *185*, 115540.
38. Puth, M.T.; Neuhäuser, M.; Ruxton, G.D. Effective use of Spearman's and Kendall's correlation coefficients for association between two measured traits. *Anim. Behav.* **2015**, *102*, 77–84.
39. Yasin, H.; Caraka, R.E.; Hoyyi, A. Prediction of crude oil prices using support vector regression (SVR) with grid search-Cross validation algorithm. *Glob. J. Pure Appl. Math.* **2016**, *12*, 3009–3020.
40. He, H.; Garcia, E.A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284.
41. Ahmed, A.M.; Rizaner, A.; Ulusoy, A.H. A novel decision tree classification based on post-pruning with Bayes minimum risk. *PLoS ONE* **2018**, *13*, e0194168.
42. Therneau, T.; Atkinson, B.; Ripley, B.; Ripley, M.B. Package 'rpart'. 2015. Available online: http://cran.ma.ic.ac.uk/web/packages/rpart/rpart.pdf (accessed on 20 April 2016).
43. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
44. RColorBrewer, S.; Liaw, M.A. *Package 'Randomforest'*; University of California: Berkeley, CA, USA, 2018.