Development of a preventive maintenance plan for Natura's production plant to improve production times and reduce downtime by predicting equipment failures

Jose Gabriel Pina

A Thesis Submitted in Partial Fulfilment

of the requirements for the

Degree of

Master of Science in Data Analytics

Diagram

Description automatically generated with medium confidence

September 2024

Supervisor: Taufique Ahmed

Summary

[Introduction 5](#_Toc177131070)

[Context y Justification: 6](#_Toc177131071)

[Relevance 7](#_Toc177131072)

[Contribution 8](#_Toc177131073)

[Objectives 8](#_Toc177131074)

[General Objective 9](#_Toc177131075)

[Specific objectives 9](#_Toc177131076)

[Develop a failure prediction model: 9](#_Toc177131077)

[Optimize the management of spare parts and resources: 10](#_Toc177131078)

[Optimize the maintenance plan: 10](#_Toc177131079)

[Literature Review 10](#_Toc177131080)

[Machine Learning 13](#_Toc177131081)

[Description of the Machine Learning Algorithms 14](#_Toc177131082)

[Random Forrest (RF) 14](#_Toc177131083)

[Logistic regression 15](#_Toc177131084)

[Decision Tree 15](#_Toc177131085)

[Support Vector Machine (SVM) 16](#_Toc177131086)

[Neuronal Networks 16](#_Toc177131087)

[Artificial Neuronal Networks (ANN) 17](#_Toc177131088)

[Recurrent Neuronal Networks (RNN) 18](#_Toc177131089)

[Convolutional Neuronal Networks (CNN) 18](#_Toc177131090)

[Regression VS Classification 19](#_Toc177131091)

[Validity 20](#_Toc177131092)

[Strategy 21](#_Toc177131093)

[Metrics 21](#_Toc177131094)

[Results 21](#_Toc177131095)

[Results using stop durations 22](#_Toc177131096)

[Results using uptime 22](#_Toc177131097)

[Sampling strategy 22](#_Toc177131098)

[Primary Research and Methodology 23](#_Toc177131099)

[Explanation and procedure for research 24](#_Toc177131100)

[Data preparation and graphic analysis 24](#_Toc177131101)

[Statistical analysis 36](#_Toc177131102)

[Null and alternative hypotheses 36](#_Toc177131103)

[Kendall Tau 36](#_Toc177131104)

[Spearman rho 37](#_Toc177131105)

[Predictive models 37](#_Toc177131106)

[FAILURE PREDICTION (Stop Duration) 38](#_Toc177131107)

[Random Forrest 38](#_Toc177131108)

[Logistic regression 41](#_Toc177131109)

[Decision Tree 41](#_Toc177131110)

[Support Vector Machine 42](#_Toc177131111)

[Neuronal Network 42](#_Toc177131112)

[Results. (Stop Duration) balanced 43](#_Toc177131113)

[FAILURE PREDICTION (Uptime) 46](#_Toc177131114)

[Random Forrest 46](#_Toc177131115)

[Logistic regression 47](#_Toc177131116)

[Decision Tree 48](#_Toc177131117)

[Support Vector Machine 48](#_Toc177131118)

[Neuronal Network 49](#_Toc177131119)

[Results. (Uptime) balanced 49](#_Toc177131120)

[Conclusion 50](#_Toc177131121)

[Recommendations 51](#_Toc177131122)

[Bibliography 52](#_Toc177131123)

List of Figures

[Figure 1 Classification of machine learning 13](#_Toc177062170)

[Figure 2 Classification of the most common machine learning algorithms. 14](#_Toc177062171)

[Figure 3 A sample feedforward deep neural network. 17](#_Toc177062172)

[Figure 4 Diagram of MLP 18](#_Toc177062173)

[Figure 5 Diagram of Recurrent Neural Network 18](#_Toc177062174)

[Figure 6 Diagram of Convolutional Neural Network 19](#_Toc177062175)

[Figure 7 Reading and concatenation of df\_log 25](#_Toc177062176)

[Figure 8 Index of translation columns df 25](#_Toc177062177)

[Figure 9 Data conversion and data types 28](#_Toc177062178)

[Figure 10 Failures per month and line 29](#_Toc177062179)

[Figure 11 Failure count per machine 30](#_Toc177062180)

[Figure 12 Failure count per line 30](#_Toc177062181)

[Figure 13 Number of stops for each cause 31](#_Toc177062182)

[Figure 14 Frequency of details 31](#_Toc177062183)

[Figure 15 Box Plot of stop duration (seconds) 32](#_Toc177062184)

[Figure 16 Mean stop duration by machine 33](#_Toc177062185)

[Figure 17 Impact Ratio of stops per machine (min) 35](#_Toc177062186)

[Figure 18 Cross Validation 39](#_Toc177062187)

[Figure 19 Correlation Matrix 40](#_Toc177062188)

[Figure 20 Decision Tree Results. (Stop Duration) balanced 42](#_Toc177062189)

[Figure 21 Results. (Stop Duration) balanced - Accuracy 43](#_Toc177062190)

[Figure 22 Results. (Stop Duration) balanced – Precision 44](#_Toc177062191)

[Figure 23 Results. (Stop Duration) balanced - Recall 44](#_Toc177062192)

[Figure 24 Results. (Stop Duration) balanced - F1 Score 45](#_Toc177062193)

[Figure 25 Results. (Stop Duration) balanced - AUC-ROC 45](#_Toc177062194)

List of Tables

[Table 1 Failure record per year 7](#_Toc177062195)

[Table 2 Results (Stop Duration) 22](#_Toc177062196)

[Table 3 Results (Uptime) 22](#_Toc177062197)

[Table 4 Meaning and content of the df columns 26](#_Toc177062198)

[Table 5 Translation of the column ‘Machime’. 27](#_Toc177062199)

[Table 6 Translation of the column stop\_cause\_description 28](#_Toc177062200)

[Table 7 Translation of the column detail 28](#_Toc177062201)

[Table 8 Failures per month and line 29](#_Toc177062202)

[Table 9 Mean stops duration and failure count 34](#_Toc177062203)

[Table 10 Random Forrest Results. (Stop Duration) 38](#_Toc177062204)

[Table 11 Random Forrest Results. (Stop Duration) balanced 41](#_Toc177062205)

[Table 12 Logistic Regression Results. (Stop Duration) balanced 41](#_Toc177062206)

[Table 13 Decision Tree Results. (Stop Duration) balanced 41](#_Toc177062207)

[Table 14 SVM Results. (Stop Duration) balanced 42](#_Toc177062208)

[Table 15 NN Results. (Stop Duration) balanced 42](#_Toc177062209)

[Table 16 Results. (Stop Duration) balanced 43](#_Toc177062210)

[Table 17 Random Forrest Results. (Uptime) balanced 46](#_Toc177062211)

[Table 18 Logistic Regression Results. (Uptime) balanced 47](#_Toc177062212)

[Table 19 Decision Tree Results. (Uptime) balanced 48](#_Toc177062213)

[Table 20 SVM Results. (Uptime) balanced 48](#_Toc177062214)

[Table 21 NN Results. (Uptime) balanced 49](#_Toc177062215)

[Table 22 Results. (Uptime) balanced 49](#_Toc177062216)

# Introduction

In this study I will base myself on the maintenance information of the production plant of the Natura group in the city of Buenos Aires, to determine through data analysis if the maintenance plan that is currently being used is adequate for the facilities and covers the needs of the machinery used, or if on the contrary it should be adapted and could through changes improve the productivity of the plant, by reducing downtime and delays due to corrective maintenance that derive from poor planning in preventive maintenance.

For this we will use different models for the prediction of data in order to compare the results and to be able to determine which one is the best adapted to the needs of the factory according to the data that we have and the relevance of these at the time of making the predictions. Within the data analysis I will use 5 prediction models (Random Forrest, Logistic Regression, Decision Tree, support vector machine and Neural Networks) which I chose based on the characteristics of what each one could contribute to the investigation taking into account the advantages and disadvantages of each one, aspect that I will deepen later with the development of the investigation.

It is worth noting that I chose to approach my research from the premise that my problem is a classification problem, in which the duration of the stoppages as well as the working times of each machine helped me to determine and differentiate between failures that generate downtime and maintenance stoppages that although they are found in the data, they may be scheduled stoppages or minor adjustments that should not be counted for the company as a failure or deficiency in maintenance.

The findings of this study could influence the adoption of new strategies for predictive maintenance contributing to more sustainable and efficient production as ineffective maintenance management can result in unplanned and prolonged downtime, which affects a plant's ability to meet production deadlines and maintain necessary inventory levels. In contrast, predictive maintenance allows you to anticipate and prevent failures before they occur, minimizing downtime and ensuring continuous and efficient operation, which means not only protecting your assets, but also ensuring your products reach the market in a timely manner.

## Context y Justification:

Avon's production plant in Moreno, inaugurated in 1977, is fundamental for the company in Argentina and the South American region as it is responsible for covering 70% of the country's domestic demand, and supplies part of the markets in Chile, Uruguay and Paraguay. This plant produces more than 400,000 products a day, so having to deal with downtime represents a setback that a priori must be minimized as much as possible in order to meet the deadlines for both internal and external markets. As we have already pointed out, this plant is part of Avon's global strategy, both in terms of capacity and geographical location. Now, since the beginning of 2020, integrated into the Natura & Co group, new quality standards have been established for products and manufacturing, which have helped to improve in the aspect that concerns us and showed a change to reinforce the commitment to innovation and sustainability.

In the following table, considering the data I have, you can see how the trend of failures over the last 3 years has been, considering that at this point the maintenance staff of the company Natura already had two years of management at the time of the first record:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Year** | **2022** | **2023** | **2024** | **Total** |
| **Stop\_date count** | **7132** | **5879** | **1586** | **14597** |

Table 1 Failure record per year

Taking these data into account, we can roughly say that management has been good since the change of administration, as we can see a decrease in failures over time, although this is something that, as I said, we can visualize with these data, although we will go into it in more depth later to determine if there is a trend in the data, either due to seasonality or machine use.

# Relevance

Nowadays, as industries advance and become more agile, they need to adapt in order to stay relevant and have an impact on the branch they belong to, because of this, most of them are adopting the idea of Industry 4.0. This methodology represents the revolution in manufacturing, integrating technologies such as artificial intelligence, the internet of things (IoT) or data analytics for the optimization of production and maintenance, according to (Lasi, et al., 2014), Industry 4.0 drives automation and data-driven decision making to improve efficiency and reduce operating costs. (Ran, et al., 2019), defines predictive maintenance (PdM) as an advanced strategy that, thanks to emerging technologies such as the Internet of Things (IoT), big data, and deep learning, allows predicting failures and optimizing asset management. Unlike reactive and preventive methods, PdM minimizes unnecessary interventions and reduces costs by performing condition-based maintenance, thus improving reliability and operational efficiency. Furthermore, (Javaid, et al., 2022) point that manufacturers would benefit from machine monitoring solutions, predictive maintenance techniques and other advanced operational technologies that will help them minimize downtime, improve performance and reduce the overall cost of producing quality components.

So, with this understanding we focus this research on the implementation of this predictive maintenance technique in order to try to achieve the benefits explained above taking into account that it could also generate a benefit in terms of logistics, space and inventory control for the ordering and stocking of spare parts for example.

# Contribution

The main contribution that I am looking for with my research is to bring Natura a little bit closer to this Industry 4.0 philosophy because nowadays, beyond the efficiency that their planning system may have, they are a step behind in this aspect because they do not use any type of tool that adapts to this philosophy. Now a days they keep records in Excel documents and plan the maintenance steps in a reactive way, considering the working times of the machines, the reliability of the equipment is calculated and based on the working times and failure trends, preventive maintenance or adjustment of each equipment is planned according to the case and the need.

But working in this way, as I mentioned earlier, does not necessarily exploit the greatest efficiency of the processes and time, as it is possible that equipment is being intervened without being necessary or at intervals that are not correct and generate unnecessary stoppages. For this reason and using predictive models I will seek to apply these new technologies to the failure logs that were provided by them in order to be able to predict failures and take action in the future to improve the performance of the factory if this is possible.

# Objectives

In this study, I address the challenges associated with downtime on production lines due to equipment failures. Predicting failures and optimizing maintenance processes are critical to improving operational efficiency and reducing associated costs in factories. According to (Palmer, 2015), in his book Maintenance Planning and Scheduling Handbook, maintenance costs can range from 5% to 15% of total production costs, and in highly specialized industries, these costs can rise to 30%. This highlights the importance of implementing a preventive and predictive maintenance plan to minimize expenses and improve operational efficiency. The objectives of this research, aimed at developing and implementing a predictive model to maximize equipment availability and optimize resource management, are presented below.

## General Objective

To develop a forecasting model based on historical maintenance and performance data to predict future failures in production equipment to reduce downtime, optimize maintenance times and improve operational efficiency in the production plant by optimizing the maintenance plan and spare parts management where applicable.

## Specific objectives

In order to achieve the primary objective of my research, it will be essential to meet the requirements of the following objectives, as each one of them represents an important part of the improvements that I am proposing. I will explain and detail them below to provide a little more clarity to what has been expressed:

### Develop a failure prediction model:

Description: Create a predictive model using machine learning techniques to anticipate equipment failures based on historical data, thus minimizing interruptions in production.

Justification: This objective is crucial to avoid unplanned downtime, providing greater effectiveness in maintenance windows. On the other hand, it is important to highlight that, as mentioned before, different prediction models will be used in order to compare their results and thus choose the one that best suits the requirements and on the other hand the one that provides the best results when making the predictions.

### Optimize the management of spare parts and resources:

Description: Make recommendations to improve the management of spare parts and resources needed for maintenance by integrating the predictive model into inventory and human resources planning.

Rationale: Optimized management of spare parts and resources can reduce costs and improve operational efficiency by ensuring that the necessary materials and personnel are available when required. Making recommendations for this objective will be much easier after determining the best performing predictive model as having a clearer view of the results, as making decisions or recommendations is much easier once the results of the research have been shown.

### Optimize the maintenance plan:

Description: Develop or improve the existing maintenance plan by making recommendations so that it is aligned with the actual needs of the plant, based on the criticality of the equipment and the predictions of the model.

Rationale: An optimized maintenance plan will reduce unnecessary interventions and ensure that resources are allocated effectively, prioritizing the most critical equipment.

# Literature Review

Maintenance engineering is a practice increasingly used in productive industrial processes or not, this is used for the productivity and flexibility of the systems involved in the production lines of factories. Nowadays, with the advances in studies using BIG DATA and Machine Learning (ML) techniques, human participation in the diagnostic processes on physical assets has been minimized to improve the early detection of potential failures. In this sense according to (Brik, et al., 2019), work previously focused on programming and reprogramming to improve system performance, however, few tasks dealt with disturbance monitoring due to the lack of real-time data, something that is nowadays increasingly common in order to decrease failure times and frequencies.

The digital era with the advances it has brought to us, has helped to adapt new technologies to industrial environments, as could be the case of the Internet of Things (IoT), which adapted industrial environments, such as the Industrial Internet of Things (IIoT), this brought us a new industrial revolution, is what is now called Industry 4. 0, which according to (Pinto & Cerquitelli, 2019), is a new concept that using interconnected sensors helps to generate large volumes of data on physical assets that favour automatic learning systems when making decisions for the associated preventive maintenance.

Today there are companies dedicated to this, to take manufacturing companies to the next level by upgrading or modernizing facilities, a clear example could be the company (DINAMOX, 2021) that offers a set of solutions to integrate the industrial environment with data analytics, for which there is a whole subcategory of this technology (IIoT) that includes applications oriented to specific customers and their requirements. This can improve supply chains, facility management and maintenance activities by monitoring parameters such as oil analysis, vibration, ultrasound, and thermography. To exemplify the advantages, I will mention below some success stories based on this tool that are also present in the above-mentioned Blog:

* <https://dynamox.net/en/blog/predictive-maintenance-at-nexa-generates-a-saving-and-18-days-of-production>
* <https://dynamox.net/en/blog/nexa-avoids-a-corrective-maintenance-cost>

The purpose of condition monitoring is to proactively classify the condition of machines or machine components to predict the time to next failure (TTF), so that an early warning can be generated before downtime is generated that affects the productivity of the factory. In their study, (Kraus & Feuerriegel, 2019) propose a structured effect neural network to predict the replacement service life of equipment that combines the approach based on historical failures and the prediction of the service life of the machines, and the approach of machine learning based on the historical data available, although in this second option for the use of neural networks, they refer to this as a black box since the interpretation of the results is reduced.

I will now proceed to explain a little about the mechanisms I will use for the development of this research.

In the first instance it will be appropriate to perform an exploratory analysis in order to determine relevant factors for the research, such as recurrent failures, symptoms of degradation in parts of the machines and actions taken in the most relevant cases, comparing it with the results obtained to determine which patterns may be relevant in the symptoms and which are not, you can use documents such as the international standard ISO 10816-3 "Mechanical vibration - Evaluation of machine vibration by measurements on non-rotating parts".

It is also important how you approach the project from a statistical point of view as there are many ways to collect and treat the available data depending on the purpose you want to use it for, because it will be different for each case if the data will be used for e.g. reliability calculation or survival analysis.

In this sense, the Weibull distribution is widely used for reliability analysis, as explained by (Scheu, et al., 2017), which is applied in reliability engineering due to its versatility when representing characteristics and adjusting parameters, allowing different analyses such as the estimation of the mean time to failure (MTTF), as well as allowing the estimation of the average useful life of the components of the machines.

Other statistical methods that could be applied are regressions where you can study for example the relationship between the hours of operation and the number of failures, to determine if there is a relationship between them.

And once the information has been obtained and processed, the next step in the research will be to apply machine learning methods that help to study it in order to make accurate predictions based on the information obtained. As (VanderPlas, 2016) points out, this is achieved through the construction of mathematical models that help determine the behaviour of the data by means of algorithms that allow the system to learn from itself. When this learning is achieved correctly, this helps the system to automatically adjust its parameters by adapting to the results obtained from the training data set to make the prediction or classification of the new observations.

In this sense, we will obviously have to go into more depth, due to the variety of algorithms and types of learning that exist, and we will also have to touch on neural networks and their application, as they will be an indispensable tool for this investment, as they are the most important complements to the machine learning.

The final expected result of this research will be to support the theory that the integration of advanced technologies can mean an improvement in the activities that could bring an operational improvement for manufacturing companies, taking NATURA as a specific case, this could help to redefine the way in which assets are operated and managed, as well as optimising processes tied to the idea of "It was always done this way" which is one of the biggest biases in industrial processes, this could put the company closer and closer to the ideology of Industry 4.0, and therefore be better prepared to meet the demands and challenges of modern industry.

## Machine Learning

As I mentioned earlier, machine learning focuses on the creation of algorithms and models capable of learning from data, allowing computers to improve their performance on specific tasks without being explicitly programmed. These systems identify patterns in large volumes of data, allowing them to make predictions or decisions based on experience. As described by (Goodfellow, et al., 2016), ‘machine learning is at the heart of the revolution in artificial intelligence, facilitating the development of systems that can learn to perform complex tasks such as speech recognition, computer vision and machine translation’.

For (Bishop, 2006) the two categories of Machine Learning methods are:

Unsupervised learning: where the algorithm is trained using datasets that do not have labels, the main objective of this type of learning is to identify patterns or structures in the data without any prior guidance, instead of predicting a label unsupervised learning helps to discover the structure of the data. An example of this could be clustering which helps to group data sets based on their similarities, such as customers in a market segment or individuals residing in the same area.

Supervised learning: the algorithm is trained using a labelled dataset, meaning that each training example is accompanied by a ‘label’ or desired outcome. The goal of the algorithm is to learn a function that maps inputs to correct outputs based on this labelled dataset. Once trained, the model can make predictions or classifications on new, unlabeled data. For example, in an image classification problem, the algorithm could learn to classify images of cats and dogs if given many images labelled ‘cat’ or ‘dog’.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Figure 1 Classification of machine learning

(Duc Le, et al., 2019) supports what was explained above and also presents a classification of the most common algorithms when facing prediction problems, it is worth noting that just like learning times, algorithms also have a greater use or effectiveness depending on the problem they face, and these are classified as classification or regression problems. It is important to note that although these are some of the most common, they are not the only ones that exist.

Diagrama

Descripción generada automáticamente

Figure 2 Classification of the most common machine learning algorithms.

### Description of the Machine Learning Algorithms

In this section I will take the opportunity to go a little deeper into the algorithms I used to train my model and additionally make a comparison between them in order to get a closer look at why they were selected.

#### Random Forrest (RF)

(Koehrsen, 2018) explains that RF is a learning algorithm used that works by creating a ‘forest’ of decision trees during training. Each tree in the forest is trained with a random subset of data, and the predictions of all the trees are combined to obtain the final prediction. In classification, the top-ranked class is the model prediction, while in regression the results are averaged. This approach improves model accuracy and reduces the risk of overfitting, making Random Forest robust and effective especially when working with non-linear data.

As I mentioned earlier, its ability to handle non-linear data is what makes it useful in this case for the problem I was faced with regarding the failure of equipment in a production factory, since although the failures could be recurrent, this does not imply that the time between them is constant or that the increase or decrease of failures is equally distributed over the period of time studied. On the other hand, another important aspect is that RF is able to handle unbalanced data, this implies that it is able to balance datasets that present a very noticeable difference between their classes, in my case for the classification I chose to predict between NO FAILURE and FAILURE, based on the days that comprise the period studied, it is clear that the NO FAILURE class will be much higher in number, but this is something that is manageable by the algorithm in order to control the bias that this can produce.

Not least, it is important to note that it is a model with very good scalability and is not prone to overfitting.

#### Logistic regression

According to (Gustavo, 2019), this model is especially useful for binary rankings when you want to predict an outcome based on characteristics or independent variables, the main advantage of this model is its simplicity of interpretation and its low probability of overfitting.

Unlike linear regression which seeks to predict continuous values, logistic regression predicts the probability that an observation belongs to a specific class, in this case NO FAIL or FAIL. This model is a valuable tool when information is not abundant or when classification characteristics are limited. On the other hand, one of its main disadvantages may be its limited effectiveness in complex relationship problems where the different variables that can affect a prediction are relevant but not independent of each other. Even so, it is still a good option for the problem I am studying in this case.

#### Decision Tree

(Koli, 2023) suggests that Decision Trees are a type of supervised learning model that are a perfect fit for classification problems like mine. The model works by dividing the data into increasingly specific subsets based on the features that provide the greatest information gain at each step.

Its main advantages are ease of interpretation as its graphical structure allows to visualize how the dataset is divided as it becomes more specific allowing to understand what the decision is based on which makes it transparent. It also allows the handling of both numerical and categorical variables without the need for additional processing, which makes it very versatile in terms of application.

Although it should also be noted that it has certain limitations such as its sensitivity to changes in the data, causing minimal changes in the structure of the data to generate a significant change in the results and the ease of overfitting that can generate a bias when making predictions.

#### Support Vector Machine (SVM)

Like the algorithms explained above Support Vector Machines (SVM) is a supervised learning algorithm widely used for classification tasks, although it can also be applied to regression. SVM works by finding the optimal hyperplane that maximizes the distance between classes in the data. This approach allows it to be particularly effective in scenarios where classes are well separated, this hyperplane can be used to classify new data based on their relative position.

As for its uses, being an algorithm capable of working with high-dimensional data makes it ideal for dealing with complex problems, such as text classification or pattern recognition. Moreover, as it works by maximizing the distance between classes, it has a good capacity to avoid overfitting, making it very reliable in that sense. Although looking at it from another point of view, that very capability could be a limitation as the processing needs are quite heavy, which makes the requirements to carry out its processing higher than the models explained above as the data becomes more abundant.

In conclusion, as (Fagbuyiro, 2023) comments, Support Vector Machines are powerful and flexible models, particularly effective in high dimensionality scenarios and in the classification of non-linearly separable data through the use of kernels. However, their high computational complexity, difficulty in interpreting the results and dependence on the correct choice of parameters may limit their applicability in certain contexts.

#### Neuronal Networks

As far as neural networks are concerned, they have been used for a long time and for multiple applications from predicting maneuvers in aerial combat to, as in my case, assessing the reliability of physical assets. Currently it is essential to implement modern technologies such as this in order to proactively detect potential failures that can lead to downtime in industrial environments that may represent a significant expense in the maintenance in factories, in this sense (Chen, et al., 2019) says that deep learning has been used in modern times as one of the main tools as far as predictive maintenance is concerned.

As (Lindholm, et al., 2019) states, the real descriptive power of a neural network is achieved when we stack multiple layers, which is known as a deep learning neural network (DNN). The figure below shows an example of a deep neural network, which consists of an input layer with six memory units, three hidden layers with five memory units each, and an output layer with two memory units; This configuration of layers allows the modelling of complicated relationships, positioning it as one of the most recent methods with the greatest number of applications in automatic learning.

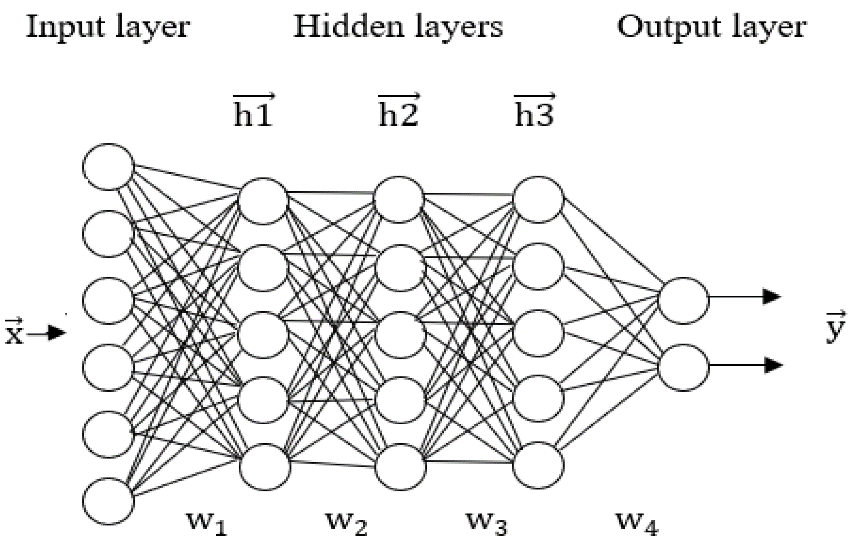


Figure 3 A sample feedforward deep neural network.

##### Artificial Neuronal Networks (ANN)

These are the most basic type of neural networks, inspired by the structure and functioning of the human brain. They are very versatile and are used to solve problems ranging from classification to regression and can solve complex prediction tasks.

Diagrama

Descripción generada automáticamente con confianza media

Figure 4 Diagram of MLP

(Source: <https://www.ibm.com/cloud/learn/recurrent-neural-networks>)

##### Recurrent Neuronal Networks (RNN)

are identified by their feedback loops. These learning algorithms are primarily leveraged when using time-series data to make predictions about future outcomes, such as stock market predictions or sales forecasting.

Icono

Descripción generada automáticamente

Figure 5 Diagram of Recurrent Neural Network

(Source: <https://www.ibm.com/cloud/learn/recurrent-neural-networks>)

##### Convolutional Neuronal Networks (CNN)

are similar to feedforward networks, but they’re usually utilized for image recognition, pattern recognition, and/or computer vision. These networks harness principles from linear algebra, particularly matrix multiplication, to identify patterns within an image.

Interfaz de usuario gráfica, Diagrama

Descripción generada automáticamente

Figure 6 Diagram of Convolutional Neural Network

(Source: <https://skyengine.ai/se/skyengine-blog/125-what-is-a-convolutional-neural-network>)

In general terms, I consider that an ANN is the most suitable option for my problem because, as it is a classification problem with structured data that does not require complex processing but is efficient, this type of network is adequate and sufficient in relation to the requirements I have, without adding unnecessary complexity to the system.

## Regression VS Classification

In terms of (Lindholm, et al., 2019), regression refers to the problem of learning the relationships between input variables (qualitative or quantitative) and a quantitative output variable, the goal being to find a model that relates the input variables to the output variable.

While classification for (Zoumana, 2024) is a supervised machine learning method where the model tries to predict the correct label of a given input data. In classification, the model is fully trained using the training data, and then it is evaluated on test data before being used to perform prediction on new unseen data. According to (Lindholm, et al., 2019), in a statistical approach, we understand classification as the problem of predicting class probabilities. Classification models can be divided into two types, when one of two possible classes is assigned, it is considered a binary classification problem and multi-class classification applies when all observations are assigned one of three or more classes.

In the context of my project, I am interested in predicting whether a failure will occur or not, which implies classifying each observation into one of two possible categories: ‘failure’ or ‘no failure’. This aligns perfectly with the definition of classification provided by (Zoumana, 2024), the binary nature of the problem, as there are only two possible outcomes, further reinforces that this is a classification problem, it focuses on predicting a discrete class, which is exactly what I need to achieve with my analysis.

So far in this paper I have attempted to explore the crucial role that machine learning techniques play in the modernization of industrial maintenance processes within the framework of Industry 4.0. In the next section I will discuss in more depth my research findings and compare the results of the models developed, which not only facilitate failure prediction, but also optimize asset management by reducing downtime and improving operational efficiency.

Each of these models brings unique advantages depending on the nature of the problem, with Random Forest standing out for its ability to handle non-linear and unbalanced data, while Logistic Regression offers interpretability and simplicity. Decision Trees are valued for their ease of use and visualization, while SVMs are valued for their effectiveness in high dimensionality problems, and Neural Networks for their ability to model complex relationships in large volumes of data.

The comparative analysis performed suggests that there is no single model that is superior in all aspects; rather, my choice of model depended on the specific context supported by the results obtained.

It is important to note that going forward, it will be essential to continue to investigate how the integration of emerging technologies, such as deep learning and hybrid models, could further improve predictive maintenance strategies.

In conclusion, the implementation of machine learning models in predictive maintenance not only offers a significant competitive advantage for industries, but also marks an important step towards the full realization of Industry 4.0, where technology and data drive smarter, more efficient and proactive decision-making.

# Validity

As for the validation of the results, they will be evaluated based on the performance of the trained models. Since the decision as to which model is more suitable or better adapted to the context of my research depended on how the models performed with the data and the quality of the predictions made.

## Strategy

For the evaluation of the models, I decided to divide the data into two sets, one of 80% to be used for training the models and 20% to be used for testing the results obtained to measure the accuracy of each model and thus be able to determine which of them had a better performance.

Regarding the classification of faults, due to the different stoppages that may exist in the production plant, take as a fault any stoppage with a duration of more than 600 seconds (10 minutes), since any stoppage with a shorter duration may be qualified as a micro-stop, a necessary adjustment or configuration in the machine, but not necessarily as a fault.

In addition to this I think it is important to highlight that due to the results obtained in the parameters measured in each model, I added a cross-validation stage in which I applied k-fold, this consisted of generating 5 sets of data, which went through the same process within the prediction model in order to verify that the values obtained were similar between the 5 sets of data used for the test, the data and values obtained from all the parameters measured I will share in the next section in which I will explain my step by step in the development of my study.

## Metrics

To measure the results, I used metrics such as accuracy, recall, F1-score and AUC-ROC to evaluate the performance of the models. Precision provided a measure of the accuracy of positive predictions, while recall indicated the model's ability to correctly identify faults. F1-score provided a balance between precision and recall, and AUC-ROC gave an overview of the model's ability to distinguish between classes.

## Results

Although, as mentioned above, all the results will be discussed and explained in the next section, I would like in this section to give a brief introduction to the findings of my research, so I will briefly mention them below, considering the two prediction models used, for the two variables that I consider relevant based on the information I had:

### Results using stop durations

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **AUC-ROC** |
| Random Forrest | 1 | 1 | 1 | 1 | 1 |
| Logistic Regr | 0,9914 | 0,9387 | 0,9903 | 0,9638 | 0,9998 |
| Decision Tree | 0,9996 | 1 | 0,9968 | 0,9984 | 0,9984 |
| SVM | 0,9843 | 0,8803 | 1 | 0,9364 | 0,9989 |
| Neural Network | 0,9698 | 0,7923 | 1 | 0,8841 | 0,9918 |

Table 2 Results (Stop Duration)

### Results using uptime

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **AUC-ROC** |
| Random Forrest | 0,8985 | 0,5508 | 0,7778 | 0,6449 | 0,9229 |
| Logistic Regr | 0,9167 | 0,6977 | 0,5242 | 0,5986 | 0,7480 |
| Decision Tree | 0,8985 | 0,5508 | 0,7785 | 0,6451 | 0,9236 |
| SVM | 0,9168 | 0,6982 | 0,5242 | 0,5989 | 0,7454 |
| Neural Network | 0,8961 | 0,5428 | 0,7785 | 0,6396 | 0,8578 |

Table 3 Results (Uptime)

The validation results show that Random Forest presented the best balance between accuracy and recall (AUC-ROC), which positions it as the most robust model for this specific problem, these results suggest that this model will provide reliable predictions in a real production environment, although this will be expanded upon in the step-by-step explanation of my research.

# Sampling strategy

Regarding the population to be studied in this project, considering that the information I work on is based on the production lines of the company NATURA for the prediction of failures and the proposal of improvements on the maintenance plan applied in its plant in Buenos Aires, I have determined that each machine to be studied will represent a unit in my population, since each one will have its own record of failures and scheduled maintenance events. Each production line is made up of different machines that perform different processes, therefore I took each machine to be studied as a unit in my population, as each one has its own record of failures and scheduled or unscheduled maintenance events, which served as input information or record with event details such as duration, type of event and the cause of the event. By analyzing these records my objective was to identify patterns in these failures in order to predict when a future event is most likely to occur on the line or equipment under consideration. At the beginning I took the whole population for the preparation and analysis of the information, although later I focused on Line 33 (L33), because when performing the analysis, I noticed that it was the line that presented more failures and represents a good opportunity to apply improvements.

Now regarding my sampling method, the method used was probabilistic, as this technique is more related to quantitative sampling, which is one of the main factors in this study, knowing that the number of occurrences of failures or events that can happen in each machine is a very important aspect when deciding which options are the most relevant. By initially allowing all equipment to have a probability of being chosen, I ensured that the sample would be truly representative and would allow the results obtained by the subsequent sampling to be more easily generalizable with greater assertiveness.

And within probability sampling I used cluster sampling. Considering that trying to randomly select individual units could be complicated because many of the production lines have similar equipment, grouping these machines by characteristics such as cost, criticality, failure rate or even obsolescence could be a safer way to target a more important segment of machinery. in this case my decision was to group the machines into clusters.

Since when studying the equipment through the block diagrams of these production lines, not all equipment has the same importance, as some might generate production stoppages, others might only slow down the process and therefore decrease the production rate, and on the other hand some might not generate any impact and could be bypassed. In general, less variability in the sample will result in higher reliability of the results and throughout the research in more efficient analysis times by reducing the population to be studied to a single group of interest, although creating a model that can be transferred to each case.

# Primary Research and Methodology

For my research as a primary source of information, I used the information obtained through a quantitative research technique, such as observation, as my main source is the equipment failure history, which is a record obtained directly from the machine and its memory for the purpose of my research in order to verify the number of failures of the studied population, their duration, reason for the event and actions taken in each case. This implies that these records were used by me to try to determine a pattern between usage and actions taken, corrective or preventive, for the purpose of proposing improvements in the maintenance plan applied by the company Natura for the production lines of a factory in the cosmetics industry.

In case my study serves as a spark for further progress, other relevant data that could be obtained directly from the machines for decision making, although not addressed in this research, could be the operational data of the equipment, such as daily production for example, since a variation in this data could mean a deterioration that could precede a breakdown that could be avoided.

Having the data firsthand and oriented to my research contributes to all data being specific and relevant to the context of my research and influential for the company NATURA which is the owner of the equipment to be studied.

As for the reasons, I think it is important to stress that the selection of this information responds to specific reasons such as relevance, as it refers directly to the studied population, control and quality of the data, as it is collected directly from the operation without any previous treatment, avoiding possible human errors or modifications, making it accurate, reliable and adapted to the context.

# Explanation and procedure for research

## Data preparation and graphic analysis

My study, as I mentioned earlier, was designed to predict failures in an industrial production environment using machine learning models. This began with the preparation and cleaning of the data, followed by the modelling and application of various machine learning models such as Random Forest, Logistic Regression, Decision Tree, Support Vector Machine, and an Artificial Neural Network (ANN). Finally, an evaluation and comparison of the results was carried out to determine which of the models performed best in terms of failure prediction.

I will now proceed to explain step by step how I progressed in the research and consequently how the decision-making process developed as the research progressed.

In a first instance the databases shared by the company are segmented by year, taking into account that the years I studied were 2022, 2023 and 2024 (the latter only until the month of May) was necessary at the time of reading the files on my Jupyter notebook, concatenate them in order to have a single unified file where the information for the 3 years in question was found and then proceed to the debugging of the information. And by using the parameter ‘ignore\_index=True’ I made sure that the dataframe indexes were reindexed continuously instead of keeping the original indexes of each dataframe producing duplicate columns.

Texto

Descripción generada automáticamente con confianza media

Figure 7 Reading and concatenation of df\_log

The next step was to review the columns resulting from this process and drop the columns that I considered would not provide relevant information to the research, to concentrate on what I was really interested in. As a result, I went from having 23 columns to having only 11 columns: 'maquina', 'linea', 'causa\_parada\_descripcion', 'parada\_fecha', 'parada\_hora', 'resolucion\_fecha', 'resolucion\_hora', 'parada\_duracion (SEC)', 'min', 'causa', 'detalle'. As can be seen, these databases were in Spanish, so in order to facilitate understanding for third parties, throughout the research, I made a translation as I went along, in the first instance I translated the names of the columns by means of an index:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Figure 8 Index of translation columns df

Of the columns I decided to work with. Below, I will detail the content of these columns:

|  |  |  |
| --- | --- | --- |
| **Original Name** | **Translated Name** | **Content** |
| maquina | machine | Machine under study |
| linea | line | Production line |
| causa\_parada\_descripcion | stop\_cause\_description | Main cause of shutdown |
| parada\_fecha | stop\_date | Date of shutdown |
| parada\_hora | stop\_time | Time of shutdown |
| resolucion\_fecha | resolution\_date | Date of resolution |
| resolucion\_hora | resolution\_time | Time of resolution |
| parada\_duracion(SEC) | stop\_duration\_sec | Duration of stop in seconds |
| min | stop\_duration\_min | Duration of stop in minutes |
| causa | cause | Explanation of the shutdown |
| detalle | detail | Extra Information |

Table 4 Meaning and content of the df columns

Once at this point, the next step was to replicate the previous step for the translation of the content of the columns of the dataset:

|  |  |  |  |
| --- | --- | --- | --- |
| **Machime** | **Translated Name** | **Machime** | **Translated Name** |
| Llenadora | Filler | Coloca Tapas | Cap Placer |
| Etiquetadora | Labeler | Paleta | Pallet |
| E33 Valvula | Valve | Roscadora | Screw Capper |
| Laser | Laser | Coloca Bombas | Pump Placer |
| Brazo transporte | Transport Arm | Falla en la tapa | Cap Failure |
| Etiquetadora de fondo | Bottom Labeler | Embalador | Packer |
| Estuchadora | Cartoner | Coloca Bolillas | Ball Placer |
| LogiPack | LogiPack | Abastecedor de bolillas | Ball Feeder |
| Molino | Mill | Girador de pucks | Puck Rotator |
| Celofanadora | Cellophaner | Balanza | Scale |
| Torqueadora | Torqueing Machine | Enhebrado/Coloca Bombas | Threading/Pump Placer |
| Abastecedor pinceles | Brush Feeder | Valvula | Valve |
| Termosellado | Heat Sealer | Bomba Neumatica | Pneumatic Pump |
| Bomba de vacío | Vacuum Pump | Picos | Nozzles |
| Axon | Axon | Abastecedor tubos | Tube Feeder |
| Transporte | Conveyor | Bomba Moyno | Moyno Pump |
| Crimpadora | Crimper | E38 Valvula | Valve |
| Inkjet | Inkjet | Sobretapa | Overcap |
| Llenadora L38 | Filler | E20 Agitador | Agitator |
| Abastecedor de bombas | Pump Feeder | Tanque Stock | Stock Tank |
| Crimpadora L11 | Crimper | E24 TK Purgado | Purging Tank |
| Etiquetadora L33 | Labeler | E24 Agitador | Agitator |
| Camara | Camera | Agitador | Agitator |
| Bajador tetina | Nipple Lowerer | E24 Valvula | Valve |
| Abastecedor de tapas | Cap Feeder | Abastecedor tetina | Nipple Feeder |
| E20 Valvula | Valve | Mordazas | Jaws |
| Martillo | Hammer | Abastecedor de mecanismos | Mechanism Feeder |
| Transfer | Transfer | Picking Forzado | Forced Picking |
| Llenadora L37 | Filler | Bajador de tapas | Cap Lowerer |
| Tolva | Hopper | No Larga | Doesn’t Start |
| Temperatura Equipo | Equipment Temperature | Cozolli | Cozolli |
| Celofanadora L10 | Cellophaner | Soldador | Welder |
| Abastecedor de envases | Container Feeder | Llenadora L16 | Filler |
| Pick and play | Pickand Place | Estrella | Star |

Table 5 Translation of the column ‘Machime’.

|  |  |
| --- | --- |
| **stop\_cause\_description** | **Translated Name** |
| Mantenimiento Mecanico | Mechanical Maintenance |
| Mantenimiento Electrico | Electrical Maintenance |
| Mantenimiento Preventivo | Preventive Maintenance |
| Quebra/Falha Mecânica | Mechanical Maintenance |
| Quebra/Falha elétrica e Eletronica | Electrical Maintenance |

Table 6 Translation of the column stop\_cause\_description

|  |  |
| --- | --- |
| **Detail** | **Translated Name** |
| ajuste | adjustment |
| micro parada | micro stop |
| rotura | breakdown |
| sin informacion | no information |
| no aplica | not applicable |
| microparada | micro stop |
| ajuste leve | minor adjustment |
| quiebra falla | breakdown |
| fusible | fuse |
| nan | no information |

Table 7 Translation of the column detail

Once everything was translated and substituted into the original dataframe, the form of the information was verified, resulting in 14648 observations and 11 columns. And in the next step I converted the columns ‘stop\_date’ and ‘resolution\_date’ from text to a datetime format, using a specific date format (%d-%m-%Y). After the data I printed the data types of each column to confirm the type of data I had, as well as to verify that the date conversion was successful, this was very important because in the following steps of the study much of the information and analysis was done based on the date.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Figure 9 Data conversion and data types

My next step was to verify the missing data in my dataset, to be able to make a determination of how I would treat them, although when I checked the result showed that in a total of 14648 observations, only 51 had missing values, this represents 0.35% of the total observations, so being such a low value I decided to remove them from the database so the final database to be treated was left with a form of 14597 observations and 11 columns and in order to have this file in case it was necessary for another step I added a line of code to save it in Excel format.

Now with this information I proceeded to move forward with the analysis of the data present in order to determine the next steps or the information that is relevant to the research. For this reason I grouped the information of the failures by number of failures occurred in each month, in each specific production line, this with the intention of determining if there was a pattern in the failures that obeyed for example to seasonality, some period of the year where the machines fail more, and on the other hand this also allowed me to verify in which of the production lines occur more failures, which in later steps I used to narrow a little the study using the technique of clustering for the example of my analysis and predictions.

Imagen que contiene Tabla

Descripción generada automáticamente

Table 8 Failures per month and line

And with the help of the graph based on this information, I could notice that the data did not obey any temporal pattern, but rather had a distribution that I could call random. For this reason, there are prediction models that I avoid using, such as time series-based models using for example ARIMA or SARIMA which are models designed to capture patterns with temporal or seasonal trends, as well as recurrent neural networks which are specifically designed to process sequences of data with temporal dependencies over time. With this information clear, I focus on classification models that are flexible to this type of data because they do not depend on the presence of temporal patterns in the data.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figure 10 Failures per month and line

Continuing with the exploratory graphical analysis, I took the information present in the dataset and continued segmenting it in different ways to obtain as much relevant information as possible before moving forward with the predictive analysis, this was important to determine where I would put the focus of the research.

Tabla

Descripción generada automáticamente

Figure 11 Failure count per machine

Gráfico

Descripción generada automáticamente

Figure 12 Failure count per line

Once in this point I proceeded to look for the main cause of stops that affect the equipment, in this case for the company they qualify it in two big sets that are stops by electrical maintenance or stops by mechanical maintenance, although inside these causes more information is added in order to have subsets to make it more specific in case of being necessary for the different analyses that they carry out emphasizing that in some cases that subset is not coupled by what in the data base appears like ‘’no information‘’.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Figure 13 Number of stops for each cause

Gráfico, Gráfico de barras

Descripción generada automáticamente

Figure 14 Frequency of details

My next step was to make a box plot to analyse the duration of the stops of the machine and I obtained the following graph as a result which explains in general terms the stops:

Gráfico

Descripción generada automáticamente con confianza baja

Figure 15 Box Plot of stop duration (seconds)

This graph suggests that stoppage durations are generally short, but there are some occasions when stoppages last much longer, generating significant outliers. The box plot is very narrow compared to the overall range of values, suggesting that most outage durations are concentrated in a small range of values, but there is a large dispersion in the higher values. This suggests to me that the maintenance response to downtime is good, so that it is not prolonged, although there are some cases where it is too long, which could be indicative that the machine involved in the breakdown may have suffered significant damage.

I then continued with the analysis by calculating the mean duration of stoppages for each machine regardless of the production line to which they belong, based on the results I could notice that there are machines with very high mean duration of stops, although with a low number of stoppages, so the impact on production downtime is not so relevant, as well as the opposite case, machines whose mean duration of stoppages is low, although the number of failures they have is high, so the impact on downtime if it is relevant.

Gráfico

Descripción generada automáticamente

Figure 16 Mean stop duration by machine

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Machine** | **Mean Stop Duration (Min)** | **Failure Count** | **Machine** | **Mean Stop Duration (Min)** | **Failure Count** |
| Filler | 52,23 | 3861 | Hopper | 38,24 | 54 |
| Labeler | 46,59 | 1942 | Transfer | 52,14 | 51 |
| Cellophaner | 68,38 | 1588 | Ball Placer | 36,80 | 50 |
| Cartoner | 51,22 | 847 | Cap Lowerer | 29,74 | 50 |
| Torqueing Machine | 55,85 | 816 | Ball Feeder | 26,93 | 48 |
| Conveyor | 51,11 | 555 | Tube Feeder | 34,49 | 48 |
| Cap Feeder | 46,02 | 378 | Moyno Pump | 52,02 | 47 |
| Laser | 43,38 | 370 | Nipple Lowerer | 33,66 | 43 |
| Crimper | 29,24 | 321 | Jaws | 51,74 | 42 |
| Inkjet | 33,33 | 302 | Nozzles | 51,23 | 38 |
| Packer | 54,14 | 299 | Pneumatic Pump | 29,03 | 33 |
| LogiPack | 80,72 | 293 | Nipple Feeder | 72,73 | 30 |
| Hammer | 62,57 | 226 | Mechanism Feeder | 46,19 | 24 |
| Pick and Place | 34,21 | 217 | Mill | 137,59 | 20 |
| Container Feeder | 34,02 | 199 | Valve | 94,53 | 18 |
| Overcap | 74,04 | 191 | Welder | 28,96 | 8 |
| Brush Feeder | 42,05 | 183 | Equipment Temperature | 304,82 | 6 |
| Camera | 36,47 | 181 | Heat Sealer | 9,46 | 5 |
| Bottom Labeler | 37,80 | 164 | Cozolli | 52,19 | 5 |
| Screw Capper | 45,27 | 162 | Agitator | 95,01 | 4 |
| Pump Placer | 63,76 | 154 | Vacuum Pump | 337,93 | 3 |
| Axon | 63,32 | 149 | Purging Tank | 115,43 | 2 |
| Puck Rotator | 33,02 | 130 | Star | 15,50 | 2 |
| Transport Arm | 39,70 | 130 | Forced Picking | 0,44 | 2 |
| Pump Feeder | 32,61 | 123 | Cap Failure | 239,27 | 2 |
| Scale | 102,33 | 63 | Stock Tank | 20,48 | 1 |
| Cap Placer | 62,38 | 59 | Pallet | 1016,13 | 1 |
| Threading/Pump Placer | 39,96 | 57 |  |  |  |

Table 9 Mean stops duration and failure count

Once we had this data, we calculated the impact to calculate the Impact Ratio of stops per machine It simply consisted of multiplying the average number of machines stops by the number of machines stops.

Gráfico, Tabla

Descripción generada automáticamente con confianza media

Figure 17 Impact Ratio of stops per machine (min)

Calculating the impact that each machine has on the productive performance of the factory is important as it helps me to know how important or how much attention should be paid to each machine according to the impact, as the greater the impact, obviously the greater the effect this failure will have on the productive time, also knowing the number of times that each machine fails helps me to narrow my focus of attention and put a special care on the machines that have the highest number of failures and take action on them, for example by increasing the frequency of preventive maintenance or on the other hand making a special study for that machine to verify that there is no external problem that could cause the machine to fail.

## Statistical analysis

The metrics used are suitable for working with non-parametric data and do not assume a linear relationship between variables, which makes them useful in situations where the relationship between variables may be monotonic but not linear.

### Null and alternative hypotheses

In this case I decided to study the relationship between the age of the machines and the failures, so for this purpose I set up the following hypotheses

* (H₀): There is a relationship between the age of the machine and the number of failures ("the older the machine gets, the more it fails").
* (H₁): There is no relationship between the age of the machine and the number of failures.

And these were demonstrated using the following statistical methods:

#### Kendall Tau

According to (De Veaux, et al., 2018), this is a robust measure to assess the strength and direction of the association between two variables, especially when dealing with ordinal data or when the assumption of normality is violated. This is one of the reasons why I use this statistical method because logic tells me that the older the machine, the greater the possibility of failure, although this is not necessarily so as this will depend on other factors such as the quantity and quality of maintenance and the frequency of use on the machines.

value ranges from -1 (complete discordance) to 1 (complete agreement). A value of 0 indicates no correlation.

For this analysis I obtained the following results:

Kendall tau: -0.236, P-value: 0.000 - The correlation is significant, so we can reject the null hypothesis.

#### Spearman rho

As explained by (Frost, s.f.) Spearman's correlation is a non-parametric alternative to Pearson's correlation. It is used for data that follow curvilinear and monotonic relationships and for ordinal data. Statisticians also refer to Spearman's ordinate correlation coefficient as Spearman's ρ (rho). This measure assesses the extent to which the relationship between two variables can be described by a monotonic function, which makes it suitable for data that do not meet the assumptions of parametric tests.

For the analysis of the results the following should be considered:

* Coefficient close to 1: Indicates a strong positive association between the two variables. As one variable increases, the other also tends to increase.
* Coefficient close to -1: Indicates a strong negative association between the two variables. As one variable increases, the other tends to decrease.
* Coefficient close to 0: Indicates little or no association between the variables.

For this analysis I obtained the following results:

Spearman rho: 0.038, P-value: 0.000 - The correlation is significant, so we can reject the null hypothesis.

Other reasons supporting the use of these statistical methods could be, for example, that the data are not continuous and contain outliers that could significantly affect the correlation between variables, and these methods are less sensitive to these outliers.

## Predictive models

In this section I will show the results obtained from the training of my prediction models. but first I will proceed to briefly explain how I modelled the information to train the models.

The first step was to convert the columns with date to date format to be sure that this would not bring problems with the way of handling the information, on the other hand this is important because when performing operations with times such as calculating the duration of the stops the formats must be similar in order to avoid errors.

The next step was to encode the categorical variables such as machine name or line number, since the prediction models require numerical inputs so using the LabelEncoder, a number was assigned to each variable to have manageable formats for the prediction models.

I created a column that would be the objective or the information sought for my predictions, this was the failure column in this I set a parameter for the stops of 600 seconds (10 minutes), where if the stop lasts more than this time, it is qualified as a “failure” and is marked with the number 1 and in case of being equal or less than the determined time it is marked with 0 which indicates “No failure”.

Next, add temporal information that is important and could influence the prediction such as the date of the previous stop to calculate the running time between stops.

Then preparing the variables “X” and “Y”, by creating the matrix in “X” the irrelevant columns are eliminated, and “Y” is defined as the target variable, to later divide the data set, this allows me to use 80% of the sample to train the model and the remaining 20% is used to evaluate the performance of the model.

### FAILURE PREDICTION (Stop Duration)

#### Random Forrest

Since I have already explained the reason for my choice of prediction models and how they work, I will now proceed to discuss the results obtained in each case.

The first model is the Random Forrest model:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Label** | **Precision** | **Recall** | **F1-score** | **Support** |
| 0 (No Failure) | 1.00 | 1.00 | 1.00 | 752 |
| 1(Failure) | 1.00 | 1.00 | 1.00 | 2168 |
|  |  |  |  |  |
| accuracy |  |  | 1.00 | 2920 |
| macro avg | 1.00 | 1.00 | 1.00 | 2920 |
| weighted avg | 1.00 | 1.00 | 1.00 | 2920 |

Table 10 Random Forrest Results. (Stop Duration)

This table shows that the model performs perfectly, with 100% accuracy, recall and F1-score in both classes, indicating that all samples for the classes are perfectly identified and that the predicted classes are also 100% accurate, with these results I assumed that the model was functional.

Although on the other hand I had the doubt that the model could possibly be overfitted so I decided to review the data and use a cross validation process to be sure that the whole process was correct, for this validation I used the following code:

Texto

Descripción generada automáticamente

Figure 18 Cross Validation

Basically, what this code does is to use the model I trained previously to perform a series of tests on the data set, taking this data set and dividing it into 5 subsets of which 4 are taken to train the model again and then compare the results with the remaining part. The process is repeated once for each subset rotating the part of the data used to evaluate. It also uses the AUC-ROC as an evaluation metric, this is a measure that indicates how well the model can distinguish between classes. A value of 1 indicates a perfect model, while a value of 0.5 indicates a model with no discriminating ability. The following results were obtained from this process:

AUC-ROC mean across folds: 0.9999945904410634

AUC-ROC std across folds: 1.0819117873062113e-05

These values suggest that the model has a high ability to distinguish classes across all subsets of information and since the variance is very low, this indicates that the variability in performance is virtually zero so that the results are consistent in each comparison.

With this data, my next step was to perform a second test using a correlation matrix, as the results obtained in the prediction model could also be due to the comparison between variables that are closely related, which could make the model highly accurate but irrelevant. This matrix is essentially a table showing the correlation between the variables present in the dataframe, yielding a munerical value representing the relationship between the features:

* 1 indicates a perfect positive correlation.
* 0 indicates no correlation.
* -1 indicates a perfect negative correlation.

Gráfico

Descripción generada automáticamente

Figure 19 Correlation Matrix

In the matrix we can see that there are no features that are so strongly related as to be redundant, so with these tests done I proceeded to continue my analysis.

In this next stage I proceeded to take a sample to narrow down my research, so within the information I had, I put the focus now on Line L33 as it is the line that has more failures in the database I have, and I proceeded to perform again the Random Forrest analysis, taking an additional consideration compared to the previous process. Taking into account that all the observations are failures and there are no observations for the days that the machine does not fail, I considered that the information was unbalanced among the classes since the proportion was very unequal Failure: 11140; No failure: 2274, I thought it was a good idea to balance the information, so I used ‘smote’ to generate a balanced training set, creating synthetic examples for the minority class and trained the model again, obtaining the following results:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Label** | **Precision** | **Recall** | **F1-score** | **Support** |
| 0 (No Failure) | 1.00 | 1.00 | 1.00 | 2374 |
| 1(Failure) | 1.00 | 1.00 | 1.00 | 309 |
|  |  |  |  |  |
| accuracy |  |  | 1.00 | 2683 |
| macro avg | 1.00 | 1.00 | 1.00 | 2683 |
| weighted avg | 1.00 | 1.00 | 1.00 | 2683 |

Table 11 Random Forrest Results. (Stop Duration) balanced

I then proceeded to train the selected models using this balanced database in order to compare the results between the models.

#### Logistic regression

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Label** | **Precision** | **Recall** | **F1-score** | **Support** |
| 0 (No Failure) | 1.00 | 0,99 | 1.00 | 2374 |
| 1(Failure) | 0,94 | 0,99 | 0,96 | 309 |
|  |  |  |  |  |
| accuracy |  |  | 1.00 | 2683 |
| macro avg | 0,97 | 0,99 | 0,98 | 2683 |
| weighted avg | 0,99 | 0,99 | 0,99 | 2683 |

Table 12 Logistic Regression Results. (Stop Duration) balanced

#### Decision Tree

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Label** | **Precision** | **Recall** | **F1-score** | **Support** |
| 0 (No Failure) | 1.00 | 0,99 | 1.00 | 2374 |
| 1(Failure) | 0,94 | 0,99 | 0,96 | 309 |
|  |  |  |  |  |
| accuracy |  |  | 1.00 | 2683 |
| macro avg | 1.00 | 1.00 | 1.00 | 2683 |
| weighted avg | 1.00 | 1.00 | 1.00 | 2683 |

Table 13 Decision Tree Results. (Stop Duration) balanced

Imagen que contiene Diagrama

Descripción generada automáticamente

Figure 20 Decision Tree Results. (Stop Duration) balanced

#### Support Vector Machine

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Label** | **Precision** | **Recall** | **F1-score** | **Support** |
| 0 (No Failure) | 1.00 | 0,98 | 0,99 | 2374 |
| 1(Failure) | 0,88 | 1.00 | 0,94 | 309 |
|  |  |  |  |  |
| accuracy |  |  | 0,98 | 2683 |
| macro avg | 0,94 | 0,99 | 0,96 | 2683 |
| weighted avg | 0,99 | 0,98 | 0,98 | 2683 |

Table 14 SVM Results. (Stop Duration) balanced

#### Neuronal Network

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Label** | **Precision** | **Recall** | **F1-score** | **Support** |
| 0 (No Failure) | 1.00 | 0,97 | 0,98 | 2374 |
| 1(Failure) | 0,79 | 1.00 | 0,88 | 309 |
|  |  |  |  |  |
| accuracy |  |  | 0,97 | 2683 |
| macro avg | 0,90 | 0,98 | 0,93 | 2683 |
| weighted avg | 0,98 | 0,97 | 0,97 | 2683 |

Table 15 NN Results. (Stop Duration) balanced

In the neural network I used there is one aspect to highlight, an early stop was added which is a technique to prevent overfitting during training. It monitors the loss in the validation set and stops training if it does not improve after 10 epochs.

### Results. (Stop Duration) balanced

This is the table with the compilation of the results obtained for my trained predictive models:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **AUC-ROC** |
| Random Forest | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 |
| Logistic Regression | 0,99 | 0,94 | 0,99 | 0,96 | 1,00 |
| Decision Tree | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 |
| SVM | 0,98 | 0,88 | 1,00 | 0,94 | 1,00 |
| Neural Network | 0,97 | 0,79 | 1,00 | 0,88 | 0,99 |

Table 16 Results. (Stop Duration) balanced

Gráfico, Gráfico de barras

Descripción generada automáticamente

Figure 21 Results. (Stop Duration) balanced - Accuracy

Gráfico, Gráfico de barras

Descripción generada automáticamente

Figure 22 Results. (Stop Duration) balanced – Precision

Gráfico, Gráfico de barras

Descripción generada automáticamente

Figure 23 Results. (Stop Duration) balanced - Recall

Gráfico, Gráfico de barras

Descripción generada automáticamente

Figure 24 Results. (Stop Duration) balanced - F1 Score

Gráfico, Gráfico de barras

Descripción generada automáticamente

Figure 25 Results. (Stop Duration) balanced - AUC-ROC

Based on the results obtained I was able to highlight that the tree-based models are clearly the ones that perform best in relation to the data I have and the context of my research, that added to their clear interpretation makes them in my opinion the most convenient, although simple, The other models are valid and effective options that are perhaps not good options, either because of the need for processing capacity or because they add unnecessary complexity, at least at this stage of the study in which I am conducting my research.

### FAILURE PREDICTION (Uptime)

As I mentioned before, I used a second method for my prediction analysis, not based on the duration of the stops but on the time of operation of the machines between failures, calculating the time difference between the date of the stop of a failure and the date of the previous failure, this characteristic of uptime, although it is not explicitly marked in the following steps, it is taken for the prediction in the model, for the calculation of the prediction of the probability of failure by means of the function ‘predict\_proba’ which returns a probability of occurrence of the failure, based on the prediction of the trees in the forest.

As in the previous case, I will now show the results of each model predicted for this case:

#### Random Forrest

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Label** | **Precision** | **Recall** | **F1-score** | **Support** |
| 0 (No Failure) | 0,97 | 0,91 | 0,94 | 11825 |
| 1(Failure) | 0,55 | 0,78 | 0,64 | 1589 |
|  |  |  |  |  |
| accuracy |  |  | 0,90 | 13414 |
| macro avg | 0,76 | 0,85 | 0,79 | 13414 |
| weighted avg | 0,92 | 0,90 | 0,91 | 13414 |

Table 17 Random Forrest Results. (Uptime) balanced

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

#### Logistic regression

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Label** | **Precision** | **Recall** | **F1-score** | **Support** |
| 0 (No Failure) | 0,94 | 0,97 | 0,95 | 11825 |
| 1(Failure) | 0,70 | 0,52 | 0,60 | 1589 |
|  |  |  |  |  |
| accuracy |  |  | 0,92 | 13414 |
| macro avg | 0,82 | 0,75 | 0,78 | 13414 |
| weighted avg | 0,91 | 0,92 | 0,91 | 13414 |

Table 18 Logistic Regression Results. (Uptime) balanced

Gráfico, Gráfico en cascada

Descripción generada automáticamente

#### Decision Tree

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Label** | **Precision** | **Recall** | **F1-score** | **Support** |
| 0 (No Failure) | 0,97 | 0,91 | 0,94 | 11825 |
| 1(Failure) | 0,55 | 0,78 | 0,65 | 1589 |
|  |  |  |  |  |
| accuracy |  |  | 0,90 | 13414 |
| macro avg | 0,76 | 0,85 | 0,79 | 13414 |
| weighted avg | 0,92 | 0,90 | 0,91 | 13414 |

Table 19 Decision Tree Results. (Uptime) balanced

Gráfico, Gráfico en cascada

Descripción generada automáticamente

#### Support Vector Machine

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Label** | **Precision** | **Recall** | **F1-score** | **Support** |
| 0 (No Failure) | 0,94 | 0,97 | 0,95 | 11825 |
| 1(Failure) | 0,70 | 0,52 | 0,60 | 1589 |
|  |  |  |  |  |
| accuracy |  |  | 0,92 | 13414 |
| macro avg | 0,82 | 0,75 | 0,78 | 13414 |
| weighted avg | 0,91 | 0,92 | 0,91 | 13414 |

Table 20 SVM Results. (Uptime) balanced

Gráfico, Gráfico en cascada

Descripción generada automáticamente

#### Neuronal Network

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Label** | **Precision** | **Recall** | **F1-score** | **Support** |
| 0 (No Failure) | 0,97 | 0,91 | 0,94 | 11825 |
| 1(Failure) | 0,54 | 0,78 | 0,64 | 1589 |
|  |  |  |  |  |
| accuracy |  |  | 0,90 | 13414 |
| macro avg | 0,76 | 0,85 | 0,79 | 13414 |
| weighted avg | 0,92 | 0,90 | 0,90 | 13414 |

Table 21 NN Results. (Uptime) balanced

Gráfico, Gráfico en cascada

Descripción generada automáticamente

### Results. (Uptime) balanced

As can be seen in the results obtained in the different variables for the different prediction methods, these are less accurate than those obtained when comparing the stop times, for this reason I decided to add a confusion matrix to show in a clearer way where the discrepancies were either false negatives or false positives. this matrix can be found under the table of results shown in each case in the previous sections.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **AUC-ROC** |
| Random Forest | 0,899 | 0,551 | 0,778 | 0,645 | 0,923 |
| Logistic Regression | 0,917 | 0,698 | 0,524 | 0,599 | 0,748 |
| Decision Tree | 0,899 | 0,551 | 0,778 | 0,645 | 0,924 |
| SVM | 0,917 | 0,698 | 0,524 | 0,599 | 0,745 |
| Neural Network | 0,896 | 0,543 | 0,778 | 0,640 | 0,858 |

Table 22 Results. (Uptime) balanced

Among the models I analyzed, Random Forest and Decision Tree again presented themselves as the most robust options for my research, given the good balance between precision and recall. Random Forest because of its ability to handle unbalanced data and its resistance to overfitting. I consider that Neural Network could be a viable option if more data and training time were available, as its performance could have been improved with a larger amount of data or some more specific data that I do not have.

On the other hand, in my opinion the use of Logistic Regression and SVM is less advisable in this context, especially with my objective of maximizing the failure detection, since both models showed inferior performance in this aspect.

# Conclusion

Given the results I obtained throughout the research between prediction based on failure duration and prediction using uptime, I consider that prediction based on failure duration is a better option as it is more accurate than prediction based on uptime.

One of the main reasons for this is due to the consistency of the results obtained by analyzing the duration of the failure, since the performance of the models was in all cases very good and consistent, which can be an advantage in the future in case of extending this study, since no additional configuration or major changes in the models are needed if you wish to add features or information that may be relevant. Also, these good results are a sign that the data used for this prediction are more adequate and provide clearer patterns for the machine learning algorithms.

Although uptime may be a good metric in certain contexts, the results suggest that failure duration provides a more robust model in the face of inherent variations and complexities in the data. It is also important to remember that during my statistical analysis in a previous section I was able to show that machine failures in maquilas are not tied to machine age, so perhaps thinking about uptime in this way may lose relevance in cases where there is preventive or scheduled maintenance that extends the life of machines in a factory environment.

And looking at the data in a practical way, prediction by failure duration is directly related to the duration of a shutdown, as it is what will determine how each event is classified, which is directly actionable for maintenance strategies. This is especially useful in a predictive maintenance context where failures need to be identified and acted promptly.

# Recommendations

It would be important to integrate the predictive models with the maintenance management system that Natura uses, to be able to generate early warnings and base the frequency of maintenance on the predictions made by the models. As I mentioned earlier, with the advent of the Industry 4.0 concept, it will be important to train personnel to understand and analyze the data provided by the models to take full advantage of the benefits they can bring.

And last but not least, before applying definitive changes in the work philosophy for maintenance personnel, it is of vital importance to analyze the cost-benefit ratio of this implementation, in order to analyze advantages, disadvantages, opportunities for improvement and potential savings, either by extending the life of the machines or reducing downtime, and thus determine whether it is feasible or worth adjusting the plan based on the results of the predictions obtained.

And as a last recommendation I propose to use the new data obtained with this research such as the uptime or the probability of failure prediction, to enrich the information you have today to have a more extensive information base that helps to develop more accurate models using as a basis the ones I already proposed to further improve their performance.

# Bibliography

Bishop, C., 2006. *Pattern Recognition and.* s.l.:Springer.

Brik, B., Bettayeb, B., Sahnoun, M. & Duval, F., 2019. Towards Predicting System Disruption in Industry 4.0: Machine Learning-Based Approach. *Procedia Computer Science,* Volume 151, pp. 667-674.

Chen, C. et al., 2019. *Automobile Maintenance Prediction Using Deep Learning with GIS Data.* Ljubljana, Slovenia, s.n.

De Veaux, R. D., Velleman, P. F. & Bock, D. E., 2018. *Intro Stats.* Fifth ed. s.l.:Pearson.

DINAMOX, 2021. *https://dynamox.net/en/blog/.* [Online]   
Available at: https://dynamox.net/en/blog/iiot-transforms-industries-and-predictive-maintenance  
[Accessed 05 2024].

Duc Le, T., Garcia, R., Casiri, P. & Östberg, P.-O., 2019. Machine Learning Methods for Reliable Resource Provisioning in Edge-Cloud Computing: A Survey. *ACM Computing Surveys,* 52(5), pp. 1-39.

Fagbuyiro, D., 2023. *https://medium.com/.* [Online]   
Available at: https://medium.com/@davidfagb/understanding-the-support-vector-machine-svm-model-c8eb9bd54a97  
[Accessed 15 07 2024].

Frost, J., n.d. *https://statisticsbyjim.com.* [Online]   
Available at: https://statisticsbyjim.com/basics/spearmans-correlation/  
[Accessed 01 09 2024].

Goodfellow, I., Yoshua, B. & Courville, A., 2016. *Deep Learning.* s.l.:MIT Press.

Gustavo, 2019. *https://towardsdatascience.com/.* [Online]   
Available at: https://towardsdatascience.com/understanding-logistic-regression-step-by-step-704a78be7e0a  
[Accessed 15 07 2024].

Javaid, M. et al., 2022. Understanding the adoption of Industry 4.0 technologies in improving environmental sustainability. *Sustainable Operations and Computers,* Volume 3, pp. 203-217.

Koehrsen, W., 2018. *https://towardsdatascience.com/.* [Online]   
Available at: https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76  
[Accessed 15 07 2024].

Koli, S., 2023. *https://medium.com/.* [Online]   
Available at: https://medium.com/@MrBam44/decision-trees-91f61a42c724  
[Accessed 15 07 2024].

Kraus, M. & Feuerriegel, S., 2019. Forecasting remaining useful life: Interpretable deep learning approach via variational Bayesian inferences. *Decision Support Systems,* Volume 125.

Lasi, H. et al., 2014. Industry 4.0. *Business & Information Systems Engineering,* Volume 6, pp. 239-242.

Lindholm, A., Wahlström, N., Lindsten, F. & Schön, T. B., 2019. *Supervised Machine Learning.* Uppsala, Sweden: s.n.

Palmer, R. D., 2015. *Maintenance Planning and Scheduling Handbook.* s.l.:McGraw-Hill Education.

Pinto, R. & Cerquitelli, T., 2019. Robot fault detection and remaining life estimation for predictive maintenance. *Procedia Computer Science,* Volume 151, pp. 709-716.

Ran, Y. et al., 2019. A Survey of Predictive Maintenance: Systems,Purposes and Approaches. *IEEE COMMUNICATIONS SURVEYS & TUTORIALS,* XX(XX).

Scheu, M. N., Kolios, A., Fischer, T. & Brennan, F., 2017. Influence of statistical uncertainty of component reliability estimations on offshore wind farm availability. *Reliability Engineering & System Safety,* Volume 168, pp. 28-39.

VanderPlas, J., 2016. *Python Data Science Handbook.* s.l.:O'Reilly Media, Inc..

Zoumana, K., 2024. *https://www.datacamp.com/.* [Online]   
Available at: https://www.datacamp.com/blog/classification-machine-learning  
[Accessed 01 09 2024].