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

# 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

Illustration 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

Illustration 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 in order 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.

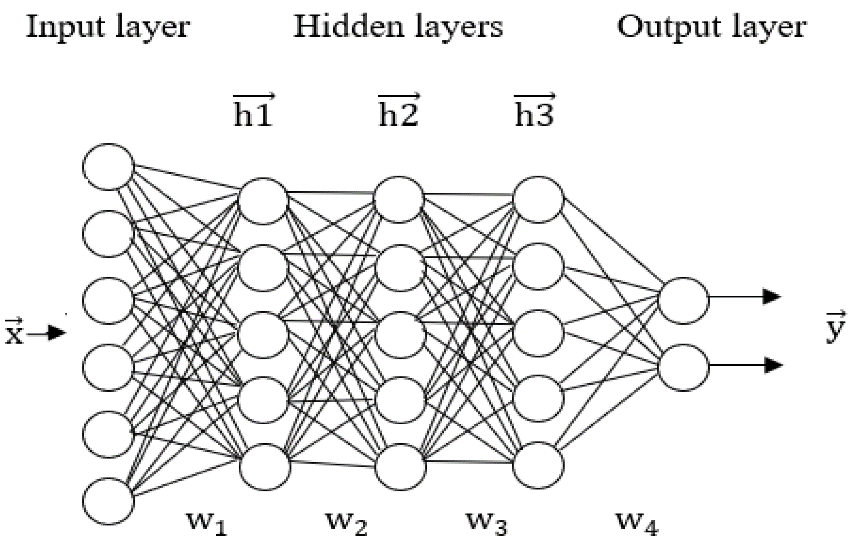


Illustration 3 A sample feedforward deep neural network.

# 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

Illustration 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

Illustration 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

Illustration 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.