* **Proposed title and topic area of Data Analysis project (10%)**
* **Title**:

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

* **Research Objectives/statement of hypothesis/eses (including alignment to your Topic area) (15%)**
* **Topic:**

In the field of industrial production, it is well known that one of the main aspects that can affect the performance of production lines is downtime due to the different failures that can occur in the equipment that make up these lines. For this reason, the study to determine the reliability of the machinery and the prediction of failures due to deterioration and use, is important to maximize the productive times and translate this into a better performance that could positively affect the costs of the production and the deadlines of compliance with the final consumer.

According to Richard D. Palmer (2015) in his book “Maintenance Planning and Scheduling Handbook”, maintenance costs can vary widely ranging from 5% to 15% of total production costs, depending obviously on the industry, the complexity of the processes, but also in highly specialized industries, this cost can rise to 30%.

Considering these costs, I consider it relevant to highlight that preventive maintenance is important as it could help to avoid additional expenses incurred in corrective actions only, which generate higher expenses and downtime that increase production costs.

Regarding the choice of the topic, I find it very interesting due to several reasons, the first and main one is that beyond the variants that may occur depending on the type of industry, it is something very applicable to all production processes of factories for goods or consumables alike. Another very important reason is related to the way in which I face the subject, since I do not have a related background, I consider that my look will be free of biases allowing me to focus more on what is related to the analysis of the data, instead of focusing on the problems related to the industrial nature of the subject.

Due to the above mentioned, the processes of data analysis, learning and prediction are highly applicable to this topic since the data are based on time series to predict the degradation of both machinery and miscellaneous in the production processes, through the study of the degradation model called Remaining Useful Life (RUL).

* **A comprehensive literature review of previous research in the area and justification for further research. The word count in your Capstone Project for your literature review is to be between 3600 and 4400 words\*. (15%)**

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 and Cerquetelli (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 DYNAMOX (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. In addition to the ARIMA model which is a powerful tool to make predictions on data sets that have time series among the information that is available.

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 aforementioned 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 mean 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.

* **Proposed sampling strategy (20%) – you are required to identify the appropriate population(s) of interest, sampling method(s), type(s) and a support paragraph outlining the reasons for your choices in each of these areas.**

Regarding the population to be studied in this project, considering that we will work on the production lines of the company NATURA for the prediction of failures and the design of the maintenance plan that adapts to this. Each production line is made up of different machines that perform different processes, thus I determine that each machine to be studied will represent a unit in my population, since each one will have its own record of failures and maintenance events whether scheduled or not, which will have information about the event such as date of occurrence, duration, type of event and cause of this. By analyzing these records my objective will be to identify patterns in these failures to predict when a future event is most likely to occur on the equipment in consideration.

Now referring to my sampling method, the method to be used will be probabilistic, since this technique is more related to quantitative sampling, which is one of the main factors in this study, meanwhile 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, this ensures that the sample is truly representative and will allow the results obtained to be more easily generalized with greater assertiveness.

And within probability sampling I will use 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.

Since when studying the equipment through the block diagrams of these production lines, not all the equipment has the same importance, because some could generate downtime in production, others could only delay the process and therefore decrease the production rate, and on the other hand some would not generate any impact could be by-passed. In general, the lower variability in the sample will result in greater reliability in the results and throughout the research in greater efficiency in the analysis times by reducing the population to be studied to only one group of interest.

* **Proposed primary research methodology (20%) - you are required to identify the appropriate research methodology/ies you plan to use and a support paragraph outlining the reasons for your choice/s. (PLEASE NOTE: You are not required to undertake any primary research for the Research Proposal)**

For my research as a primary source of information, I will use the information obtained through a quantitative research technique, such as observation, since my main source will be the history of equipment failures, which will be a record obtained directly from the machine and it’s registration 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 will be used to try to determine a pattern between the use and the actions taken, corrective or preventive, for the design of the appropriate maintenance plan for production lines in the cosmetics industry.

Other relevant data that could be obtained directly from the machines to be used in my research could be the operational data of the equipment, such as the daily production for example, since a variation in this data could mean a deterioration that could precede a breakdown that could be avoided.

Having all these data firsthand and oriented to my research contributes to make all the data 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, it seems important to me to emphasize that the selection of this information responds to specific reasons such as relevance, as it refers directly to the population studied, control and quality of the data, since they are collected directly from the operation without any previous treatment, this avoids possible human error or modification, making them accurate, reliable and adapted to the context in which they will be obtained.

* **You are required to discuss the ethical and legal/regulatory considerations you expect to encounter in both the sourcing and the analysis of appropriate data for this project. (20%)**

From my point of view the only area of my project where I could face a problem of ethics, I consider that it could be in the obtaining of the information since the primary source of information is the company that in this case is the owner of the machines, and this is due to the fact that the company could consider this information as sensitive or simply keep it for internal use, and the biggest challenge for me in this case could be in the way in which this information will be obtained. In this sense I have taken the corresponding actions to ensure that I am not incurring in any fault, by consulting this information directly with one of the Heads of reliability and maintenance of the company NATURA in its subsidiary in Argentina, by doing this I am taking into account the implications that sharing this information could entail, and from the company's point of view, it is only necessary to sign a document, which establishes as a commitment that the use of the information will be solely and exclusively for educational purposes, so that no profit will be obtained from it and on the other hand that it will not be shared or disclosed outside the scope covered by this assignment. On the other hand, it is important to bear in mind that communication with the company always must be clear, as this demonstrates the good faith and intentions that are intended with this research.

I think it is important to note that the absence in this sense of a greater influence of what may be ethical behaviors are somewhat relegated in the research as it does not deal with sensitive or personal data for individuals so it does not transgress the rights of any of the parties involved, this means that most of the limitations when sharing information do not apply, in short the only limitations that may exist are those expressed by the company in its agreement to disclose information because my study population will be the machines of the production lines.

**References**

Brik, B., Bettayeb, B., Sahnoun, M. & Duval, F. (2019), ‘Towards predicting system disruption in industry 4.0: Machine learning-based approach’, Procedia Computer Science 151, 667–674.

DYNAMOX. (2021). IIoT transforms industries and predictive maintenance. DYNAMOX Blog. Retrieved from [https://dynamox.net/en/blog/iiot-transforms-industries-and-predictive-maintenance].

Kraus, M. & Feuerriegel, S. (2019), ‘Forecasting remaining useful life: Interpretable deep learning approach via variational bayesian inferences’, Decision Support Systems.

Pinto, R. & Cerquitelli, T. (2019), ‘Robot fault detection and remaining life estimation for predictive maintenance’, Procedia Computer Science 151, 709–716.

Scheu, M., Kolios, A., Fischer, T. & Brennan, F. (2017), ‘Influence of statistical uncertainty of component reliability estimations on offshore wind farm availability’, Reliability Engineering and System Safety 168, 28–39.

VanderPlas, J. (2016), Python Data Science Handbook, O’Reilly Media, Inc, US.