

# Degree in Data Science and Engineering

---

**Title: Analysis and prediction of the Overall Equipment Effectiveness of manufacturing processes**

**Author: Sara Sánchez López**

**Advisor: Jordi Ricart**

**Co-advisor: Martí Soler**

**Tutor: Silverio Martínez-Fernández**

**Institution: ClearPeaks**

**Month and year: June of 2021**



**UNIVERSITAT POLITÈCNICA  
DE CATALUNYA  
BARCELONATECH**

**Facultat d'Informàtica de Barcelona  
Facultat de Matemàtiques i Estadística  
Escola Tècnica Superior d'Enginyeria de Telecomunicació de  
Barcelona**



# CONTENTS

---

- 1 Introduction ..... 1**
  - 1.1 Motivation and Context..... 1
    - 1.1.1 Industrial problem..... 2
  - 1.2 Background on Overall Equipment Effectiveness ..... 3
  - 1.3 Previous work..... 5
  - 1.4 Structure of this document ..... 6
- 2 Objectives ..... 7**
- 3 Data understanding and preparation ..... 9**
  - 3.1 Dataset description ..... 9
    - 3.1.1 Analysis of the metrics..... 12
  - 3.2 Data preparation ..... 14
- 4 Modelling..... 16**
  - 4.1 Modelling results..... 17
    - 4.1.1 Availability ..... 18
    - 4.1.2 Performance ..... 19
    - 4.1.3 Quality..... 20
  - 4.2 Planification of work orders..... 21
- 5 Reporting..... 23**
  - 5.1 OEE prediction..... 24
  - 5.2 Availability, performance and quality..... 25
  - 5.3 Data analysis ..... 27
  - 5.4 On-site prediction..... 28
  - 5.5 Production order planning..... 29
- 6 Proposed solution: the pipeline ..... 31**

<b>7</b>	<b>Validation of the tool .....</b>	<b>32</b>
<b>8</b>	<b>Conclusions.....</b>	<b>34</b>
<b>9</b>	<b>Acknowledgements .....</b>	<b>36</b>
<b>10</b>	<b>References.....</b>	<b>37</b>
	<b>Appendix.....</b>	<b>39</b>
	A. Project plan.....	39
	B. Replication package.....	40
	C. Economic viability.....	41
	D. Ethical implications .....	43

# ABSTRACT

---

One of the main focuses in the manufacturing industry is to diminish the costs associated with the lack of effectiveness of the machines during the production process. To be able to track and benchmark progress, the Overall Equipment Effectiveness (OEE) metric is widely used in the industry, as it takes into consideration the three main pillars when it comes to determining the effectiveness of a production process: availability, performance and quality.

This final thesis describes the work done for a company named Mapex, which creates software that helps manufacturing companies to track their OEE, amongst others. They wanted to add an advanced analytics layer to their software that was able to predict future OEE values given a set of production orders.

This project has been thought of as a Proof of Concept (PoC) from ClearPeaks, a consulting firm, to Mapex, to prove that it is possible to create a tool that is able to make predictions of future OEE values given a set of production orders and show the results in some dashboards that are easy to understand and use. The main objective of this PoC is to encourage Mapex's clients to provide their data and adapt this advanced analytics layer into their environment.

This thesis explains the process of creation of the proposed tool: the initial extraction of data from the client's database, the creation of the predictive models and the dashboards and the final validation of the tool, in which the client was asked about its conformance with it.

In this final validation, it was proved that the proposed tool met the initial expectations and that it could be used to attract Mapex's clients to acquire this new advanced analytics layer. Nonetheless, some improvement points were identified in order to make the tool even more useful.

# RESUMEN

---

Uno de los mayores objetivos en la industria de la manufacturación es disminuir los costes asociados a la falta de efectividad de las máquinas durante el proceso de producción. Para poder realizar un seguimiento de los progresos, la métrica “Overall Equipment Effectiveness” (OEE) es ampliamente usada en la industria, ya que tiene en cuenta los tres pilares que determinan la efectividad de un proceso productivo: disponibilidad, rendimiento y calidad.

Esta tesis final describe el trabajo hecho para una compañía llamada Mapex, que crea software que ayuda a las empresas manufactureras a hacer un seguimiento de su OEE, entre otros. Mapex quería añadir una capa de analítica avanzada a su software que fuese capaz de predecir valores futuros de OEE dado un conjunto de órdenes de fabricación.

Este proyecto ha sido pensado como una Prueba de Concepto de ClearPeaks, una empresa de consultoría, a Mapex, para demostrar que es posible crear una herramienta que sea capaz de predecir el OEE dado un conjunto de órdenes de fabricación y mostrar los resultados en unos reportes que sean fáciles de entender y usar. El principal objetivo de esta Prueba de Concepto es alentar a los clientes de Mapex a que proporcionen sus datos y adapten esta nueva capa de analítica avanzada en su entorno.

Esta tesis explica el proceso de creación de la herramienta propuesta: la extracción inicial de los datos de la base de datos del cliente, la creación de los modelos predictivos y los reportes y la validación final de la herramienta, en la que el cliente fue preguntado por su conformidad con ella.

En esta validación final se comprobó que la herramienta propuesta cumplía con las expectativas iniciales y podía ser usada para atraer a los clientes de Mapex a adquirir esta nueva capa de analítica avanzada. No obstante, se identificaron puntos de mejora para hacer que la herramienta sea aún más útil.

# RESUM

---

Un dels majors objectius en la indústria de la manufactura és disminuir els costos associats a la falta d'efectivitat de les màquines durant el procés de producció. Per poder realitzar un seguiment del progrés, la mètrica “Overall Equipment Effectiveness” (OEE) és àmpliament usada en la indústria, ja que té en compte els tres pilars que determinen l'efectivitat d'un procés productiu: disponibilitat, rendiment i qualitat.

Aquesta tesi final descriu el treball fet per una empresa anomenada Mapex, que crea software que ajuda a les empreses manufactureres a fer un seguiment del seu OEE, entre d'altres. Mapex volia afegir una capa d'analítica avançada al seu software que fos capaç de predir els valors futurs d'OEE per un conjunt donat d'ordres de fabricació.

Aquest projecte ha estat pensat com una Prova de Concepte de ClearPeaks, una empresa de consultoria, a Mapex, per demostrar que és possible crear una eina que sigui capaç de predir l'OEE donat un conjunt d'ordres de fabricació i mostrar els resultats en uns informes que siguin fàcils d'entendre i usar. El principal objectiu d'aquesta Prova de Concepte és motivar als clients de Mapex a que proporcionin les seves dades i adaptin aquesta nova capa d'analítica avançada en el seu entorn.

Aquesta tesi explica el procés de creació de l'eina proposta: l'extracció inicial de les dades de la base de dades del client, la creació dels models predictius i els informes i la validació final de l'eina, en què el client va ser preguntat per la seva conformitat amb ella.

En aquesta validació final es va comprovar que l'eina proposada complia amb les expectatives inicials i podia ser usada per atraure als clients de Mapex a adquirir aquesta nova capa d'analítica avançada. Tot i així, es van identificar uns punts de millora per tal de fer l'eina encara més útil.

# 1 INTRODUCTION

---

This final degree project has been developed at an external institution, ClearPeaks, which is a consulting firm specialized in Business Intelligence, Analytics and Data Management. The work under this final thesis has been done for one of their customers, Mapex.

This section is respectively divided in four subsections: motivation, background, previous work and structure of this document.

## 1.1 MOTIVATION AND CONTEXT

Mapex is a corporation which produces Manufacturing Execution System (MES) software for manufacturing companies [1]. With this software the clients can configure and plan all the different steps and components that are involved in a manufacturing process: production lines, products, phases of production, quality controls, etc.

There is a vast amount of data that is generated with every production order that is sent to the factories. Before production, we already have data regarding the planning of such order, like the number of pieces to produce. The volume of data increases even more once the order goes into production; we can know how much time it took to produce, if there were any incidences during the process, how many pieces had some kind of fault and the information collected from the sensors inside the machines, amongst others.

In order to justify the performance obtained from a production line in the past, Mapex, as many other companies in this field, uses the Overall Equipment Effectiveness (OEE) metric. This OEE is calculated based on three other metrics:

- Availability, which measures the percentage of time in which a product line was available for use with respect to the planned allocated time.
- Performance, which measures the real speed of production compared to the nominal performance of the production line.
- Quality, which measures the percentage of good pieces with respect to the total amount that were produced.

Mapex, with its software, provides the OEE metric to its clients so that they can find a reasoning behind the results obtained from their production lines. For example, if a production order was delivered with a delay with respect to the original planning, it may be because the performance during the days in which the production was held was quite low.

### **1.1.1 Industrial problem**

Being able to know in advance how a production order is going to perform is something that could benefit Mapex's clients. Following the example above, if the predicted OEE is low because of the performance metric, they can know that the production order will most likely be finished later than expected, and they can plan accordingly. If the reason why the OEE is low is because the prediction says that the availability of the machines will be low, maybe it is a good time to perform some maintenance work to avoid machine failure during production time.

Prior to this project, Mapex's clients have already tried to predict OEE on their own but have failed to obtain a useful solution. In view of this, Mapex thought it would be a good idea to add an analytics layer to its software, which would allow the prediction of OEE for a certain production order before sending it to the factory. This could help improve the planning of the production orders and even optimize the use of the resources. Moreover, this added layer of analytics could give more insights on the production process and help identify the weak points and improvement possibilities of factories, making them more efficient.

To address the aforementioned industrial problem, ClearPeaks proposed to Mapex to do a Proof of Concept project, in which the client agreed to provide data to ClearPeaks to analyse it and show whether it is possible to predict OEE. Such proposal of creating a Proof of Concept is the starting point of this Final Degree Project.

## 1.2 BACKGROUND ON OVERALL EQUIPMENT EFFECTIVENESS

Overall Equipment Effectiveness (OEE) is a standard metric used for measuring manufacturing productivity. This term was first introduced in 1982 by the Japanese scientist Seiichi Nakajima in his book “TPM<sup>1</sup> tenkai” [2], as a component of the Total Productive Maintenance concept.

This metric is obtained from the product of three other metrics: availability (A), performance (P) and quality (Q). It can be expressed with the following formula [3]:

$$OEE=A*P*Q$$

Where:

$$A=\frac{\text{Planned production time}-\text{Non planned stops}}{\text{Planned production time}}$$

$$P=\frac{\text{Ideal cycle time}^2*\text{Total count}^3}{\text{Planned production time}-\text{Non planned stops}}$$

$$Q=\frac{\text{Total count}-\text{Faulty units}}{\text{Total count}}$$

One of the goals of OEE is to identify and mitigate the 6 Big Losses [3]:

- Availability losses: those which affect the amount of time in which the machines can be in production.
  - o Unplanned Stops: these are the stops in production usually due to a failure in the equipment. The machine was scheduled for production but could not operate due to an inconvenience (like the equipment breaking or there not being enough materials or operators).
  - o Planned Stops: these kinds of stops are due to setups, adjustments or maintenance work. They can be planned and taken into consideration before sending a production order.
- Performance Losses: these are the ones which affect the velocity in which machines produce the products, after subtracting the availability loss.

---

<sup>1</sup> TPM: Total Productive Maintenance

<sup>2</sup> Theoretical minimum time to produce one part.

<sup>3</sup> Total number of units produced, including defects.

- Small Stops: they are minor stops which are not big enough to be considered an unplanned stop. Material jams, incorrect settings or quick cleanings of the machines fall into this category.
- Slow Cycles: this loss occurs when the equipment is not producing as fast as it should, based on its nominal performance. This could be due, for example, to bad lubrication of the machines or worn-out equipment.
- Quality Losses: pieces which have defaults produce a loss in quality. They will either have to be repaired, refactored or completely made again, which implies a loss in time and raw materials.
  - Production Rejects: it counts the number of faulty pieces that were produced with respect to the total amount during the steady state of the machine.
  - Start-up Rejects: it accounts the pieces which have some sort of defect, and which have been produced from the start-up of the machine until it achieves a steady state. This is accounted because machines can take some time to adapt until the production becomes steady and the probability of pieces being faulty during this period of time can be higher.

After subtracting all the losses above, we get the Fully Productive Time, which is considered as the Valuable Operating Time. That is summarized in Figure 1:



Figure 1 Effect of availability, performance and quality losses to the Planned Production Time. Source: [4]

### 1.3 PREVIOUS WORK

The analysis and prediction of the Overall Equipment Effectiveness has already been studied in the past. In this chapter I summarize some of the main contributions.

Hassani et al. used data from an automotive cable production industry [5]. The data contained information on 22 similar performing machines designed to cut wires. Each type of wire is considered a different product, and for each of them they had information about its specifications (cable class, wire length, number of terminals, number of seals, etc.). For them, a production order was defined by a quantity of one product, which was sent from the planning department automatically to each of the machines.

The raw data that Hassani et al. had included information about the OEE values per shift and machine and the different orders of fabrication with the associate specifications done by every machine. From all this data, they found that some of the variables which played an important role in predicting the OEE value were: setups (sum of all the setup values from the production orders), breakdown (information about the planned maintenances), number of orders (how many orders are there planned by machine and day), mean wire length, number of terminals, number of seals.

They tested several different models (Support Vector Machine, Random Forest, Extreme Gradient Boosting, and Deep Neural Networks) to predict the estimate OEE value and found that the better performing models were the ones using Deep Learning techniques (6.27 Mean Absolute Error) and, also, the Support Vector Machine (6.16 MSE).

Another contribution to the prediction of OEE can be found in the paper “Performance Prediction through OEE-model” by CH. Anusha and V. Umasankar [6]. They used simple moving averages and Holt’s double exponential smoothing methods to predict the future performance of OEE. The first technique measures the overall trend of the dataset and the second responds more quickly to most recent changes, hence why this method resulted in a minimum error for them.

To perform their study, they obtained data from an Auto ancillary company for one year. The data consisted of planned production time per shift, total losses, net operating time, ideal cycle time, parts per minute, total parts produced and rejected. With all this information, they obtained the respective availability, performance and quality metrics.

Their framework was able to establish the major factors affecting OEE through equipment losses (breakdown, changeover and adjustment, start up, speed loss, management loss, motion loss, defects and rework, quality process, energy loss, etc.). They observed that the simple moving average attained the minimum error at short term, but the error gradually increased past the 3-month period. The opposite happened with Holt's double exponential smoothing method.

When it comes to the major factors affecting OEE, they found that breakdown loss, management loss and distribution loss had the biggest effect towards the efficiency of the machine.

As a conclusion, they found that simple moving averages were incapable of recording sudden changes in data and Holt's double exponential smoothing method the one with minimum error.

Although the prediction of OEE has already been studied in the past, this project will focus on how that can be implemented in the specific case of a MSE company. Not only will OEE predictions be made, but some interactive reports will also be created to make sure that the results from the models can be easily interpreted by Mapex's clients.

## **1.4 STRUCTURE OF THIS DOCUMENT**

This document is structured as follows.

Section 2 shows the objectives set for this project. Section 3 aims to describe the provided database and explain the data preparation process. Section 4 describes the best results obtained from the models, done for each of the metrics that need to be estimated. Section 5 shows the developed Power BI reports which serve as a visualization of the results obtained during this project. Section 6 summarizes the whole pipeline of the proposed solution. Section 7 shows the results of the validation of the tool by the client. Section 8 describes the contributions that this project has made to Mapex and the final conclusions. Section 9 collects the acknowledgments. Section 10 contains all the referenced documentation and literature cited in the document.

## 2 OBJECTIVES

---

This section describes the main objectives to be attained during the development of this project.

**The main objective of this project is to provide a tool that Mapex can use to showcase to their clients that it is possible to add an advanced analytics layer to their software that can predict the OEE value for the next production orders.**

Therefore, given the initial goal for this project, a set of subobjectives can be defined: creating Machine Learning models to predict future OEE values and providing an actionable tool that shows the results from the models in some easy to interpret dashboards.

The first subobjective is to show whether it is possible to make predictive models that use data from a production process provided by Mapex to predict its Overall Equipment Effectiveness by means of 3 metrics: availability, performance and quality. Therefore, 3 models will need to be developed to individually predict each of these metrics:

- **Availability.** The goal is to predict the amount of time in which the machines will be available for use as a percentage of the total amount of time that they were allocated for production.
- **Performance.** This model will be the one responsible for predicting the variation of performance of the machine compared to its nominal performance.
- **Quality.** This model will predict the percentage of good pieces with respect to the total amount.

A requirement for this part of the project is that the first two models should use as input data the information about the production orders (like the planned number of units to produce, the allocated machine, the type of product, etc.), and for the quality model data from the sensors of the machines should be used (temperature, pressure, humidity, etc.).

The second subobjective of the project is to consolidate all the information obtained from modelling into a Power BI report that will provide insights on the production process. The objective of this is to make the results of this project understandable to non-experts and help those making the production orders, Mapex's clients, to optimize their planning. It can also be helpful

to the managers of the factories to know which are the weak points during the production process and try to find a fix for them.

Finally, to validate the quality and success of the proposed tool (the models and the dashboards) a new subobjective can be set: to validate the tool with Mapex and the project managers to obtain feedback and verify that the solution proposed in this final thesis meets the main objective.

Figure 2 summarizes the tasks to accomplish during the development of this project, how they are related with the subobjectives set in this section, the contributions of each of the phases of the project (see Appendix A) and the sections of the document to which they belong.

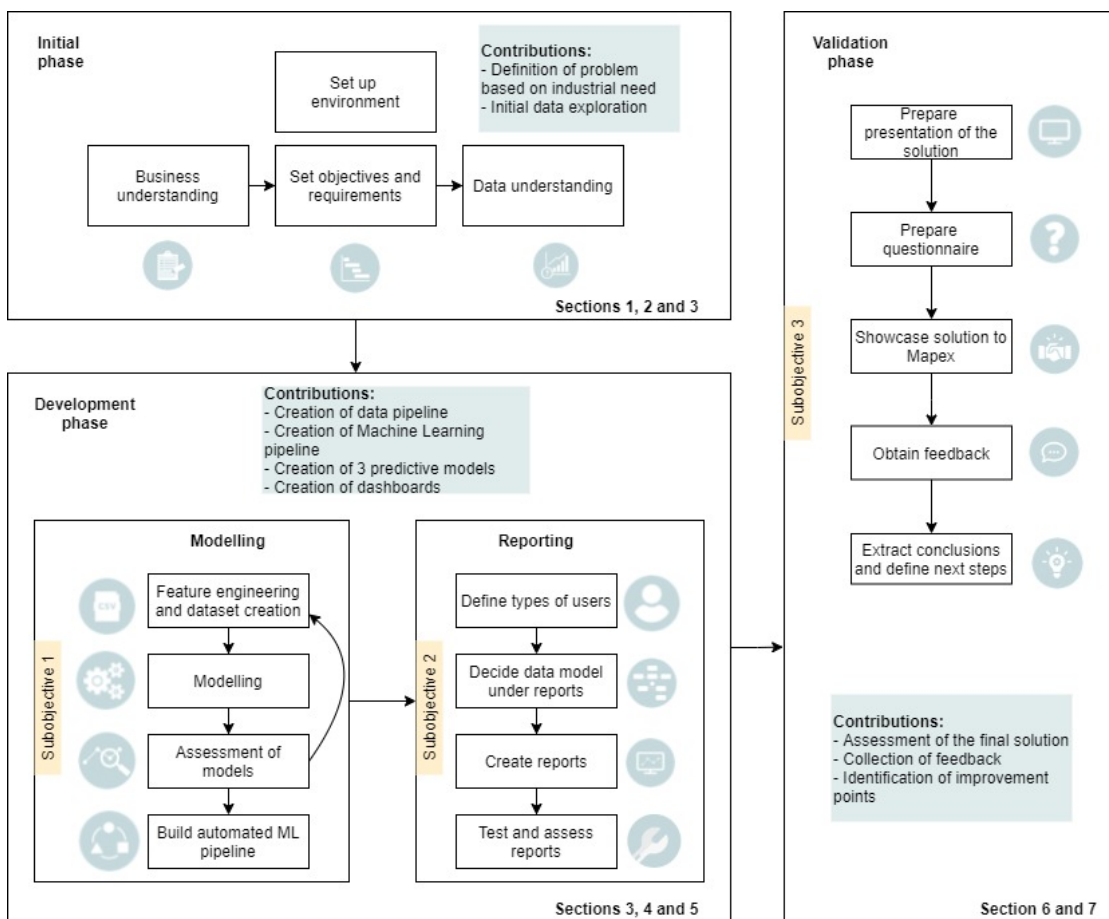


Figure 2 Diagram of tasks, objectives and contributions of the project

### 3 DATA UNDERSTANDING AND PREPARATION

---

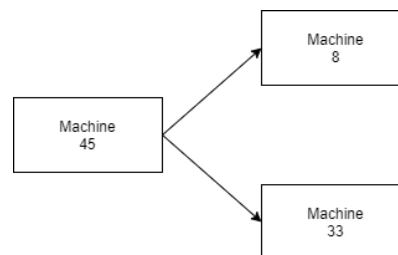
This section describes the provided database and explains how the data has been processed in order to obtain the datasets needed to train the models.

#### 3.1 DATASET DESCRIPTION

This subsection describes the most important tables in the provided database, and the relationships between them.

The raw data provided for the development of this project is located in an SQL Server database. This database consists of real data from one of Mapex’s clients, which has been anonymized.

It contains records from September of 2018 until the end of February of 2021, which log the production and planning information for three of the machines in the same factory. Figure 3 shows the order and placement of such machines.



*Figure 3 Placement of the machines in the factory of the provided database*

The machine with ID 45 produces the paste for an edible product and the machines with IDs 8 and 33 give shape to it. These two machines are set up in a parallel manner since they perform the same task. On the demo dataset available for use in this project, the machine 45 only contains sensorics data, but the information necessary to be able to calculate the availability, performance and quality metrics is only available for machines 8 and 33. Therefore, the models developed during this project only apply to these two machines.

Most of the information relevant for this project can be extracted from the tables:

- **his\_of:** it contains historic information about the production orders being sent to the machines. It contains information such as the ID of the product that must be produced, the planned quantity of products to be produced, the batch number, the state of the fabrication order, etc.
- **his\_fase:** fabrication orders can be broken down into phases, which represent the different phases that a product goes through to be produced. In the specific case of this client's dataset each product has one and only phase, but this might not be the case for every client. This table keeps track of the historic information of the production phases: the ID and description of the phase, the number of planned units to produce, the ID of the production order to which it corresponds, etc.
- **his\_prod:** differently from the two tables above, which keep track of the historic planned production orders and phases, his\_prod records the actual measurements once the order of fabrication has gone into production: the ID of the machine which is producing, the ID of the shift who is working during that period of time, the starting and ending date, the ID of the activity (that can be: production, closed, maintenance, etc.), the expected nominal performance of the machine, the number of OK and not OK pieces that were produced, the number of seconds in which there was a planned stop (PP<sup>4</sup>), a non-planned stop (PNP<sup>5</sup>) or just some time in which the machine was not planned to be in production (NAF<sup>6</sup>).
- **his\_prod\_paro:** this table has information of all the stops that the machines have experienced: the starting and ending date, the ID of the stop and the ID of the record in his\_prod to whom it is related. This way, the ID of the machine responsible for the stop in production can be retrieved. More information about the details of the stop can be found in the dimension tables cfg\_paro, cfg\_tipoparo1, cfg\_tipoparo2, cfg\_tipoparo3, cfg\_tipoparo4 and cfg\_tipoparoOEE.
- **his\_graf:** this is the table which contains information about the sensorics of the machines, which retrieve information of the state of the machine (such as its temperature). The fields that can be found in this table are the date in which the sensorics value was collected, the

---

<sup>4</sup> PP: "paro programado" (planned stop time)

<sup>5</sup> PNP: "paro no programado" (non-planned stop time)

<sup>6</sup> NAF: not available for use. That is, during weekends or holidays, for example.

value itself and the ID of the machine, amongst others. It has a corresponding dimension table called `cfg_graf` which contains extra information such as the unit of measurement or the type of sensor it is. The collection rate of this table is quite high, every 2 minutes there is a new record with measurements, so having this type of data would provide a lot of information about a production order. As explained before, the only machine which contained sensorics data in the provided dataset was the one with ID 45, which did not contain OEE information, so this data could not be used.

Figure 4 shows the relationship between the tables described above.

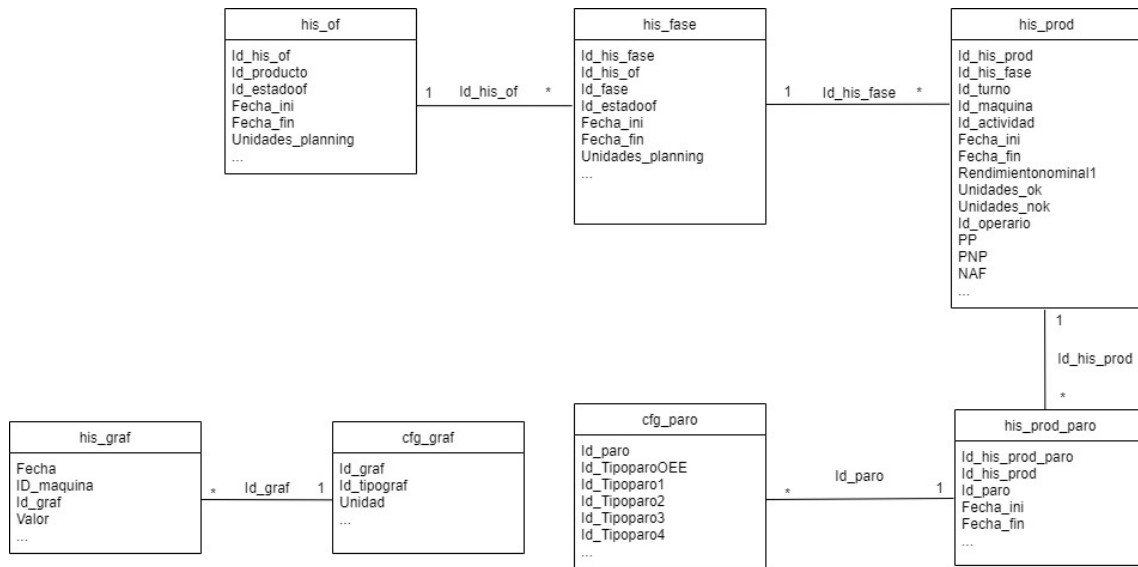


Figure 4 Reduced data model

### 3.1.1 Analysis of the metrics

As stated before, only the data related to machines 8 and 33 will be taken into account for this project. Also, as there is no sensorics data in the provided database to perform an accurate prediction for the quality metric, the focus of this analysis will be, specially, the data that can be found in the table his\_prod. To build the dataset from the provided database, only the records that belong to closed production orders were considered.

The table below describes some of the key metrics of the table his\_prod.

Metric	Machine 8	Machine 33
Number of records	10154	2640
Number of production orders	766	228
First recorded date	2018-10-04 22:36:14.000	2018-11-11 21:31:39.000
Last recorded date	2021-02-25 16:01:28.000	2021-02-18 03:56:37.000
Mean availability	88.01	83.92
Mean performance	99.94	130.127
Mean quality	99.10	97.87

Table 1 Main descriptors of table his\_prod

The figures below show the histograms and boxplots for the three target variables:

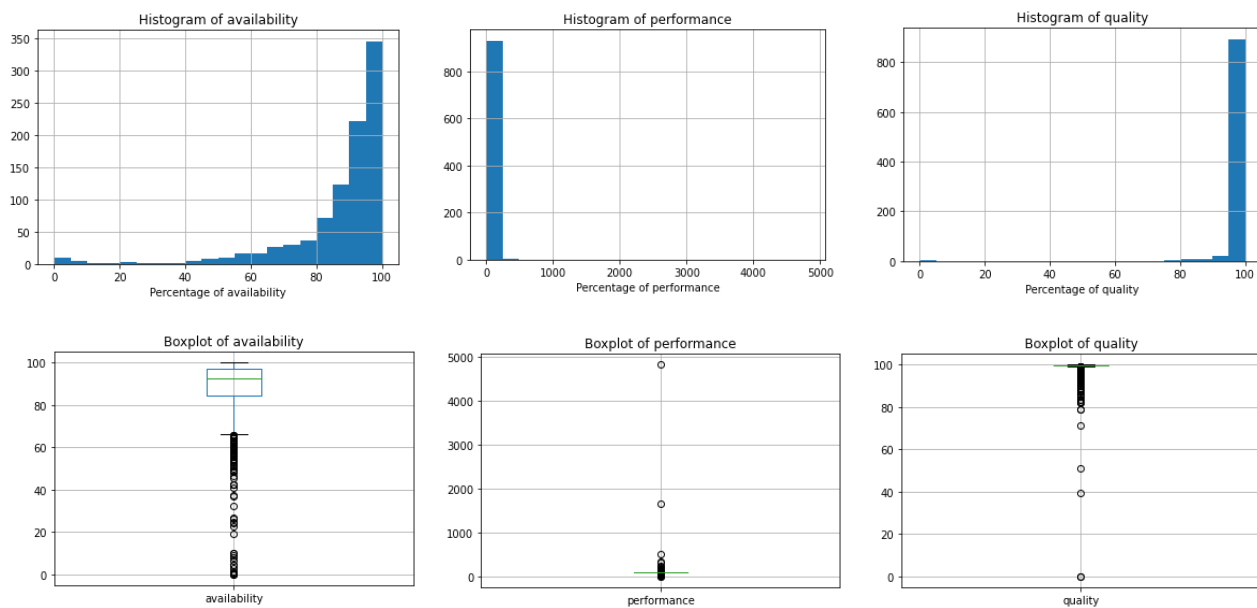


Figure 5 Histograms and boxplots of the three main metrics: availability, performance and quality

The availability metric is highly imbalanced, most of the observations have values ranging from 80 to 100 percent. This means that the predictive models developed for this metric will most likely

have a hard time learning what makes a production order have a low availability result. On the other hand, the quality metric is quite concentrated with most of its values being above 95 percent. Finally, the performance metric seems to have too high of values for some observations. After discussing this issue with the matter expert at Mapex, it was decided that any observation with a value of this metric higher than 110% could be considered an outlier, probably due to an error in the tagging of the nominal performance of the machines. After removing the outlier values, a more realistic distribution of the metric can be observed:

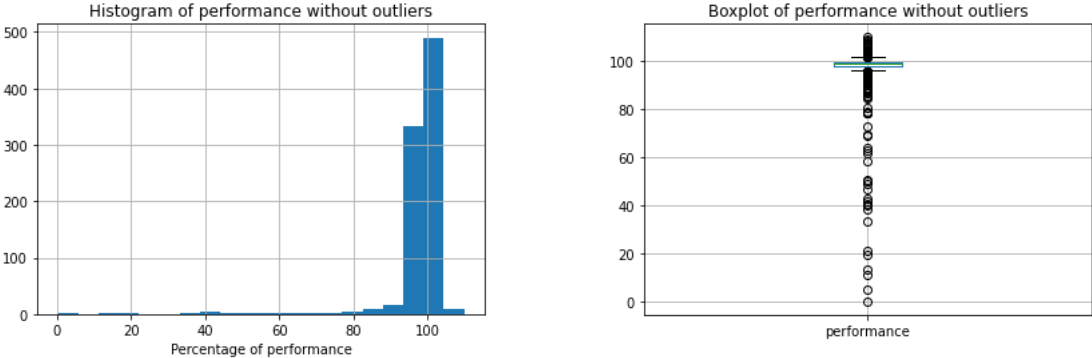


Figure 6 Histogram and boxplot of the metric performance after eliminating outliers

The values are also highly concentrated, as it happens with the quality metric, most of them revolve around 100%.

### 3.2 DATA PREPARATION

This subsection explains the data pipeline and how the final datasets for training are obtained.

After obtaining an initial understanding of the data available on the main table, his\_prod, it was time to perform some feature engineering to obtain variables that could potentially benefit the models. Figure 7 shows the implemented data pipeline using Python.

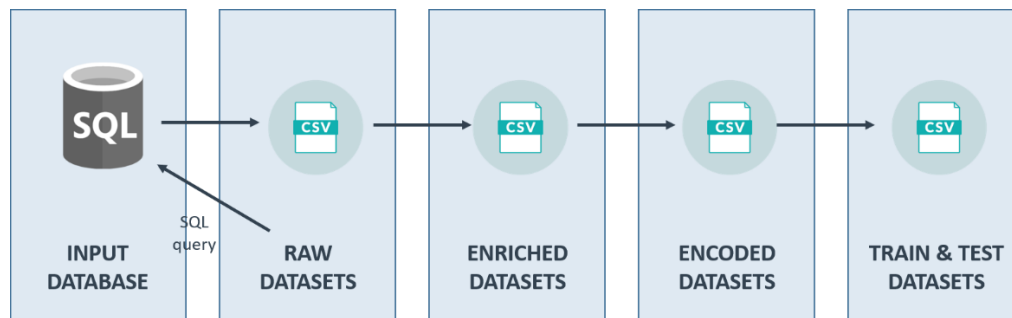


Figure 7 Data pipeline

First, a raw dataset is obtained from the database by executing an SQL query using the pyodbc module of Python [7]. This first dataset contains variables present in his\_prod or his\_fase such as the ID of the production order, the machine to which it was allocated, the nominal performance of the machine, the planned number of units, the ID of the product and its type and the day in which the production started. To add more variables into the dataset, a new set of features was created by joining the table his\_prod with the one which contains information about the stops, his\_prod\_paro. This way a new set of variables was added like the duration and the seconds passed since the last planned and non-planned stops or the number of stops in the last three days. Each row of this dataset represents a production order in a specific machine. The target metrics to be predicted are also calculated in this step by using the data available in the tables, following the same formulas already used by Mapex in their software.

Once this raw dataset is obtained, an enriched dataset is built after performing some feature engineering over the initial variables to get those features that were harder to obtain through an SQL query: moving averages (with windows of 5 and 10 observations) of the duration of stops

and time between stops; moving average of the most recent values of the target metrics and cyclic variables (sine and cosine) to represent the date of production.

Once the enriched dataset is built, the categorical variables are encoded by using the one-hot encoding technique.

Finally, train and test datasets are created. To do that, the last month with data (February of 2021) is saved for testing. Since cross-validation is used to obtain the Mean Squared Error (MSE) of the models, no validation datasets are created in this step. Instead of that, the validation sets are built during the modelling part by using Python Pandas data frames.

It is important to mention that all Mapex's clients share the same data model, which facilitates a standard extraction of the data, no matter which client ends up acquiring this solution.

## 4 MODELLING

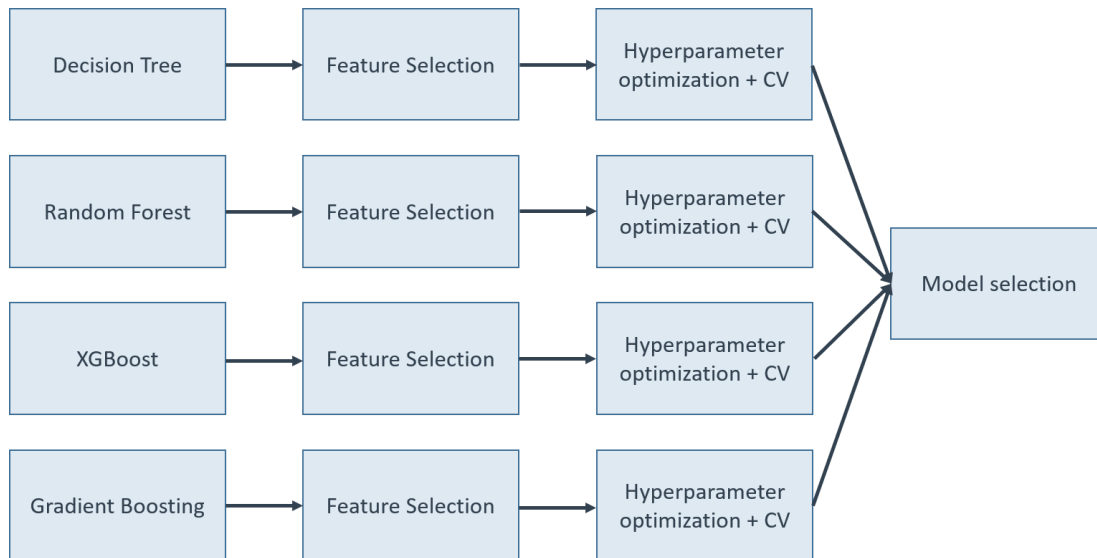
---

This section describes the pipeline that has been implemented in order to obtain the models for each of the metrics and the results that they provide.

To be able to predict the OEE for a given production order and machine, there needs to be a predictive model for each of the components of OEE: availability, performance and quality.

The target values that these models will try to predict are real numbers, usually between 0 and 100, and they indicate percentages. Therefore, we are dealing with three regression problems.

Figure 8 shows the pipeline that has been implemented to obtain these models.



*Figure 8 Diagram of the model selection pipeline*

This pipeline must be repeated and executed for each of the three metrics that need to be predicted.

At the beginning of the modelling part of this project, different models were tested to see their overall performance in predicting each of the metrics. After a few iterations, the best performing models were selected: Decision Trees, Random Forest [8], XGBoost [9] and Gradient Boosting [10].

The pipeline then uses Forward Feature Selection [11] to obtain, for each of the four baseline models, the variables that are the most relevant to them. Afterwards, once each model has its own set of input variables, a grid search is performed for each of them to obtain the best configuration of hyperparameters. To choose the best performing model amongst all the configurations, a 3-fold cross validation (CV) is performed using the Mean Squared Error (MSE) as the error that is tried to be minimized. Finally, the best model between the four best configurations is trained with the whole train dataset.

Every time that a model is trained a JSON file is generated, which stores the final results obtained by each of the four baseline models after being optimized: the train and validation errors, the hyperparameters and variables chosen for each of them, and the predictions made in the validation sets. This allows for reproducibility of the results.

This pipeline can be thought of as a case of AutoML<sup>7</sup>, which is helpful in cases like this where the solution may be implemented to clients which may not have a data scientist available to supervise the training of the models. This way, every time the models need to be trained, the best performing model will always be chosen automatically, without the need of human supervision.

## 4.1 MODELLING RESULTS

This subsection presents the results obtained for each of the models that have been fitted to predict the main metrics of OEE: availability, performance and quality.

As mentioned before, MSE has been used as the error to be minimized. The test dataset consists of the last month with data, which has 40 observations. All models have been cross validated using 3 folds.

To create all these models, only the most recent data was used, the work orders of last year (2020), since that yielded the best results. This was discussed during one of the meetings with the matter expert at the client company and it was agreed that it made more sense to only use recent data for the models, since those serve as a better representation of the current states of the machines.

---

<sup>7</sup> Automated Machine Learning (AutoML) is the process of automating the tasks of applying machine learning to real-world problems [12].

### 4.1.1 Availability

The chosen model for the availability metric is the Gradient Boosting one. It shows a MSE of 205.27 in the training dataset, 323.87 in validation and 178.43 in test.

The chosen variables for this model are the machine IDs, the type of product, the nominal performance of the machines, the day of the week, whether it is a weekend or not, the number of stops two days prior to the actual date and the seconds that have passed since the last non-planned stop.

The figures below show the predictions made by the model in each of the validation sets.

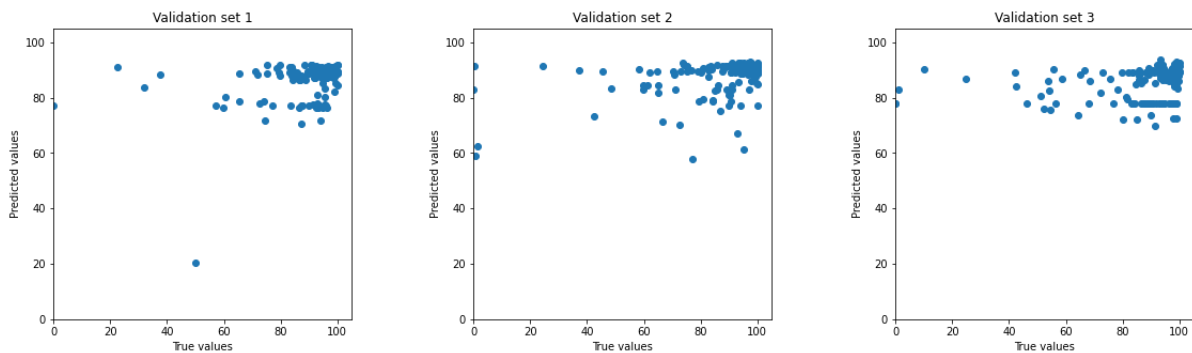


Figure 9 Scatter plots of the predicted versus true values in the validation sets for the availability model

Although the model does not seem to be overfitting, since it acts in a similar manner in all 3 datasets, it is clear to see that it does not perform well with low values of availability. That may be because the dataset is unbalanced and there are not many observations with low values of availability.

Figure 10 shows the results obtained on the test dataset for the availability metric.

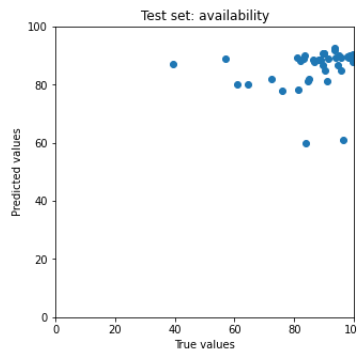


Figure 10 Scatter plot of the predicted versus true values in the test set for the availability model

For this dataset, the model is a lot more accurate with low values of availability than it was in the validation sets. Except for a couple of observations, the model usually predicts higher values of availability compared to the true values.

### 4.1.2 Performance

The chosen model for the performance metric has been XGBoost. It has an MSE of 175.83 on the train dataset, 189.43 on validation and 15.30 on test.

The chosen variables for this model are the machine IDs, the type of product, the day of the week, whether or not it is a weekend and the average, minimum and maximum duration and number of stops three days prior to the actual date.

The predictions obtained in the validation datasets are shown in the graphics below:

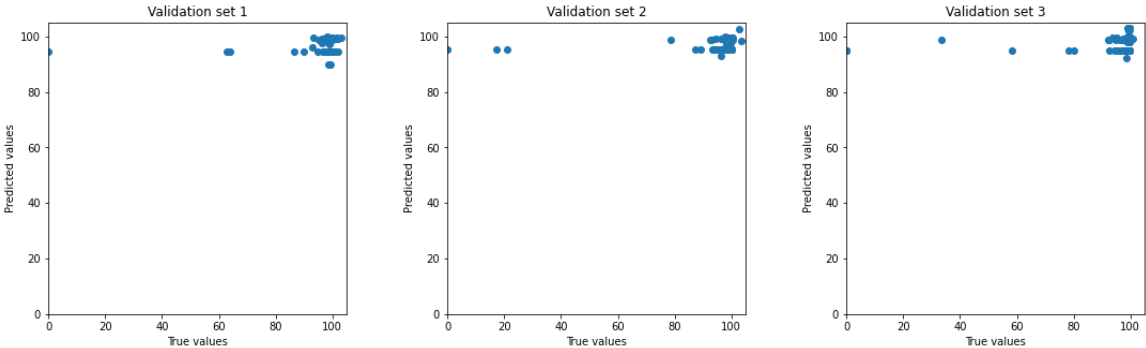


Figure 11 Scatter plots of the predicted versus true values in the validation sets for the performance model

In this case, the target values are a lot more concentrated compared to the availability metrics, except for a few values which could be outliers. This concentration of values could be the explanation to why this model is performing better compared to the availability one, since the range of error is not as big.

Figure 12 shows the predictions that this model makes for the test dataset.

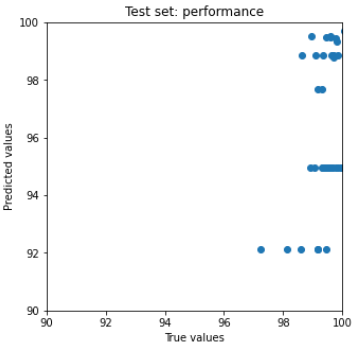


Figure 12 Scatter plot of the predicted versus true values in the test set for the performance model

### 4.1.3 Quality

Following the requirements and the initial expectation for this project the quality model should be trained using the sensorics data that can be obtained from the table his\_graf. Nevertheless, the data that was available for use for this project did not have actual sensorics data (that is, temperature, pressure or humidity of the machines, for example) but rather most of the information it contained was data from multiple counters which tracked the number of pieces that were fabricated, how many of them were faulty or not, etc. This information is directly related to target variable, quality, and therefore cannot be used in the model.

The sensorics data may not be available in some cases, like this one, since this is an extra module that not all Mapex's clients have. Because of that, a model was fitted using as input dataset the same one as for the availability and performance metrics.

The chosen model in this case was the Decision Trees, with a MSE of 97.66 in the train dataset, 137.97 in validation and 17.66 in test.

The chosen variables for this model are the ID of the machine, the product type, the day of week, the number, minimum, maximum and average duration of stops in the last 3 days and the sine and cosine representation of the month variable.

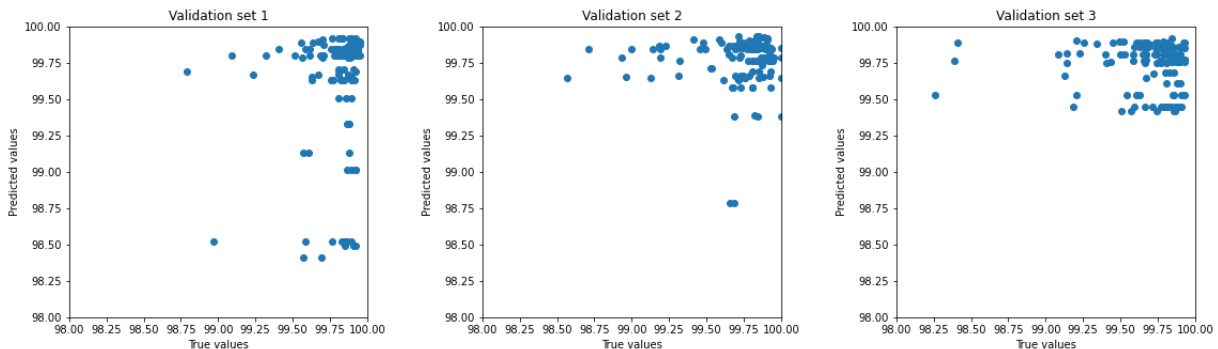


Figure 13 Scatter plots of the predicted versus true values in the validation sets for the quality model

In this case, the target value is so concentrated that inevitably, the model gives better results when it comes to MSE. An interesting point to note from looking at the validation predictions, is that in the first set the model seems to be giving lower values of quality compared to the true values, and the opposite happens on the other two sets. This may mean that the model is overfitting.

Figure 14 shows the predictions made by the model on the test dataset. Note that the axis have been cropped in order to display small variations.

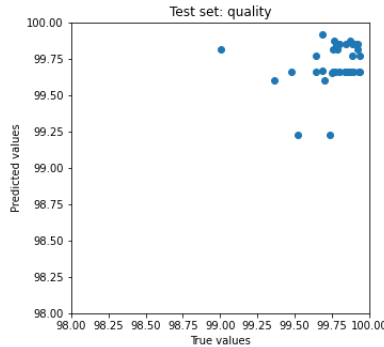


Figure 14 Scatter plot of the predicted versus true values in the test set for the quality model

The predictions on this set look much nicer, it does not seem that the model is systematically giving higher quality values that it should, as one could think from looking at the validation sets.

The table below summarizes the results obtained for each of the metrics:

	<b>Availability</b>	<b>Performance</b>	<b>Quality</b>
<b>Chosen model</b>	Gradient Boosting	XGBoost	Decision Tree
<b>Train score</b>	205.27	175.83	97.66
<b>Validation score</b>	323.87	189.43	137.97
<b>Test score</b>	178.43	15.30	17.66

Table 2 Summary of results obtained from models

## 4.2 PLANIFICATION OF WORK ORDERS

Once a model has been trained for each of the three metrics and predictions on the test dataset have been obtained, one can calculate the expected number of items that need to be produced as well as the durations of time that the machine will need to be running and available for. To be able to make such calculation, two parameters are needed: the number of units that need to be produced in a specific work order (namely PU, planned units) and the nominal performance of the machine (namely NP, nominal performance) which indicates the number of pieces that a machine can produce in one hour.

Therefore, given the predicted values of availability, performance and quality for a production order we have that:

$$\text{EU} = \text{Expected units to produce} = \left( 1 + \left( 1 - \frac{\text{predicted quality}}{100} \right) \right) * \text{PU}$$

$$\text{ERT} = \text{Expected running time of machine} = \frac{\text{EU}}{(\text{predicted performance} * \text{NP})}$$

$$\text{EAT} = \text{Expected available time of machine} = \left( 1 + \left( 1 - \frac{\text{predicted availability}}{100} \right) \right) * \text{ERT}$$

Finally, given a starting date for a set of production orders, it is easy to calculate following the formulas above the estimated starting and ending date of each of the tasks. The results obtained from these calculations are then showed in the Power BI reports, giving those in charge of the planning of the production orders a better understanding of the effect that OEE has in the planification of the work orders.

## 5 REPORTING

---

This section shows the tool that has been developed in Power BI in order to visualize the results obtained by the models and provide some helpful visuals to their users.

To be able to interpret the results obtained by the models, it is necessary to have a tool that shows some visuals which are easy to interpret, and which allow to obtain insights on the data. This way, it will be easier for the users to act and know which are the measures they need to concentrate on in order to improve their OEE.

This tool has been created with three different profiles in mind:

- The factory manager, that is the person in charge of the overall wellbeing of the factory and who is kept accountable for its performance. A profile like this will need to have, in just a glance, a complete overview of how the next production orders are going to perform. This profile also needs to have a clear understanding of which are the factors that affect the most to the final OEE result.
- The factory workers or operators, those responsible for putting the production orders into fabrication. This tool is also helpful for the people working on-site, so that they can know, before starting the machines for a production order, if it is going to do well or not. If there is a metric that is predicted to do quite bad, they could instantly take some action to prevent bad results. For example, if the predicted availability metric is low, there could be a high chance that the machine is going to fail repeatedly and could benefit from some maintenance work.
- The production order planner. The tool can be especially helpful for this profile, since it would allow them to know an estimation for the time it will take to produce all the work orders and, therefore, it could help them improve the planification of the supply chain.

## 5.1 OEE PREDICTION

The first dashboard, represented in Figure 15, shows an overview of how the next production orders are going to perform.

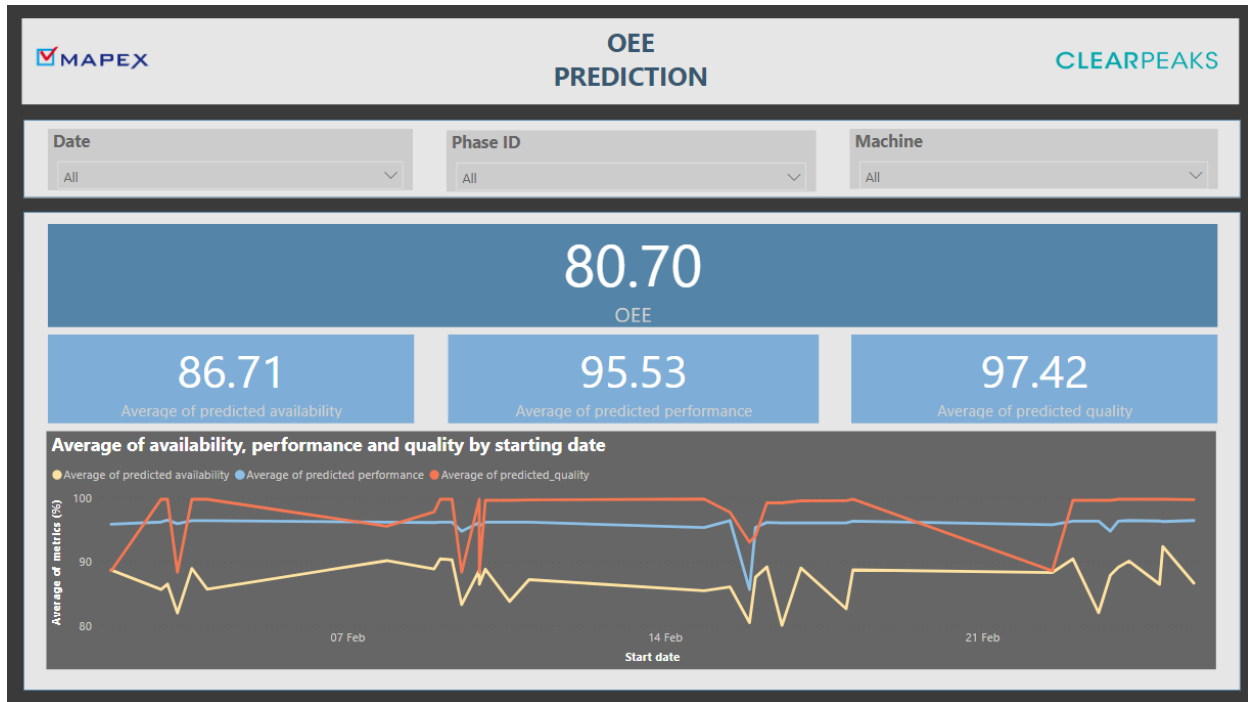


Figure 15 First dashboard: OEE prediction

At the top of the page there are some slicers that can be used to filter which production orders the user wants to focus on: by the starting date, their ID or the machine they are allocated to.

Next, we can see the average OEE value for all the production orders as well as the averages of the individual metrics: availability, performance and quality.

On the bottom of the page there is a line graph which shows the evolution of the three metrics with respect to the starting date of each of the production orders.

## 5.2 AVAILABILITY, PERFORMANCE AND QUALITY

The next three dashboards (figures 16, 17 and 18) show in more detail the information related to the work orders.

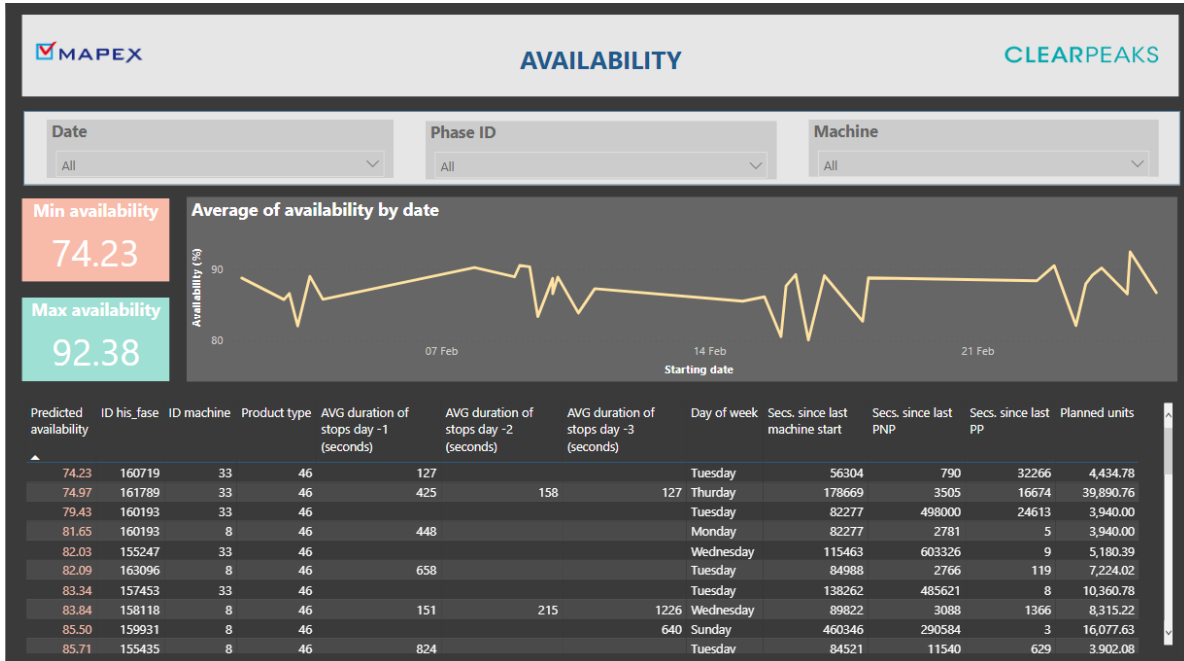


Figure 16 Second dashboard: Availability metric

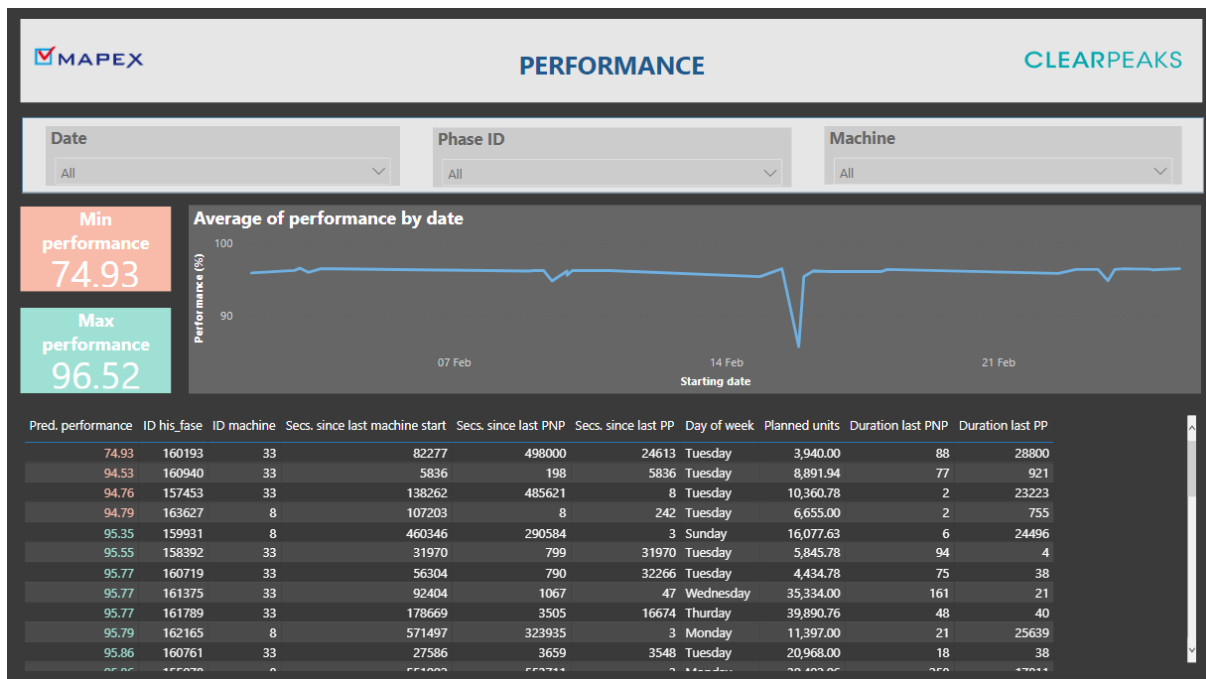


Figure 17 Third dashboard: Performance metric

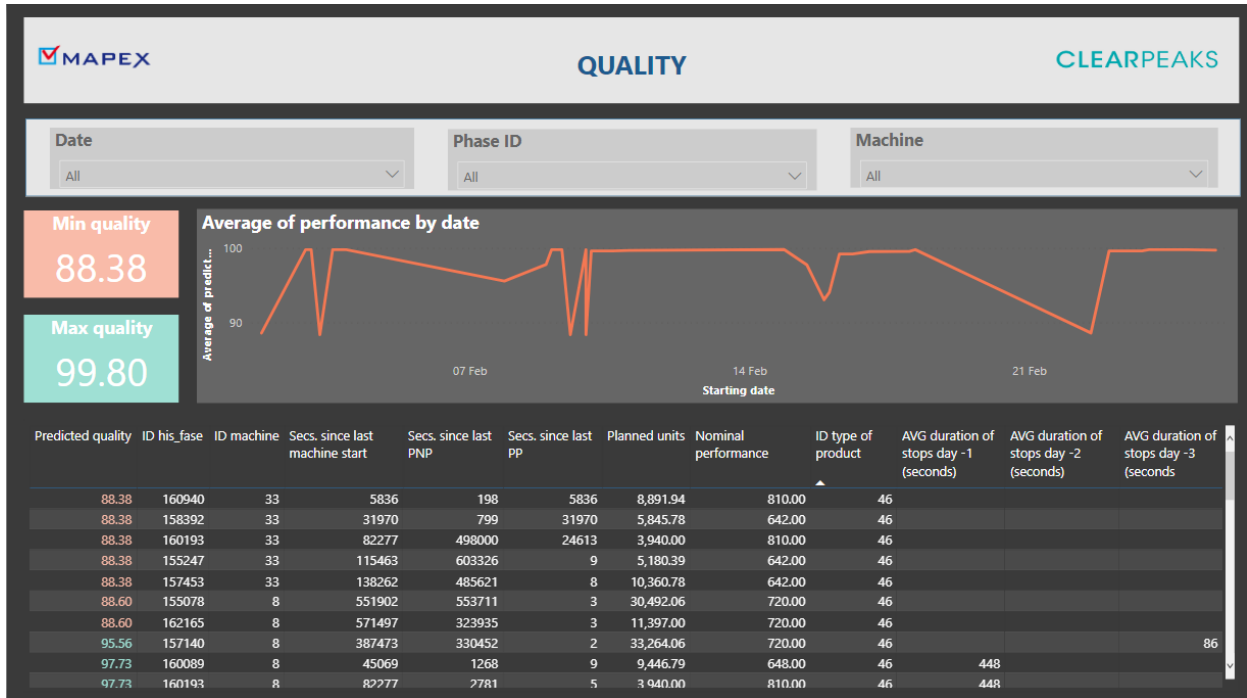


Figure 18 Fourth dashboard: Quality metric

As well as with the first report, there are some filters that we can use to focus on only some specific work orders.

Below the filters there are two cards which show the minimum and maximum values for each of the metrics and a line graph which shows the predicted values by date.

On the bottom of the page there is a table which shows the individual predictions as well as some of the data that was used to obtain such prediction. This way, the subject expert can know which are the changes that need to be made in order to improve the results obtained.

### 5.3 DATA ANALYSIS

The fifth dashboard, shown in Figure 19, allows the person in charge of planning the production orders to know how different variables affect the main metrics.

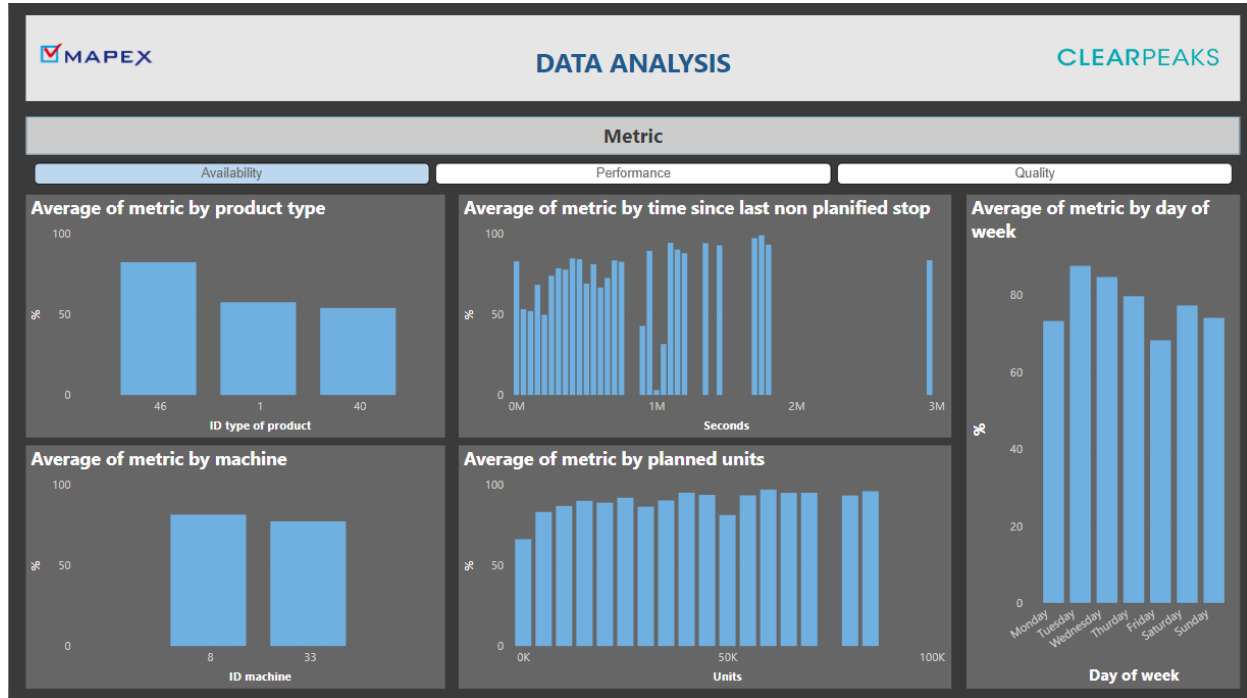


Figure 19 Fifth dashboard: Data analysis

This dashboard shows some visuals that help the user gain insights on how the metrics behave with respect to some variables like the allocated machine, the number of units that need to be produced or the time that has passed since the last non-planned stop, for example. These visuals are interactive and the user can hover over them to obtain more information, like the size of the bins or how many observations there are in each of them.

With these visualizations, the person in charge of planning the production orders can know which are the weak and strong points in the production line and establish policy changes such as planning bigger production orders or prioritizing the machine that gives better results.

The metric that wants to be studied can be changed by using the selector on the top of the dashboard.

## 5.4 ON-SITE PREDICTION

This dashboard, pictured in Figure 20, has been created with the factory workers in mind, thinking of the most important thing that they would need to know before putting an order into production.

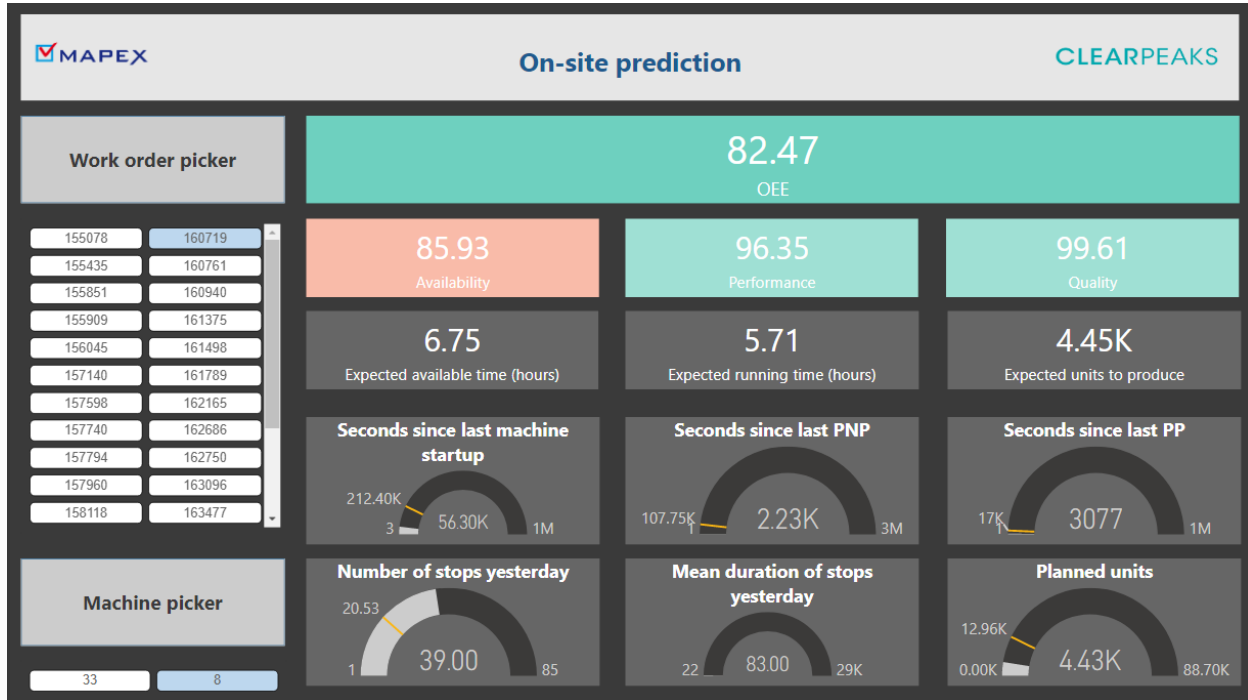


Figure 20 Sixth dashboard: On-site prediction

On the left-hand side of the report there are two slicers which act as buttons that allow to select the next work order which needs to be produced and the allocated machine. On the right side, the worker would see the predicted OEE for that order and how that breaks down into the different metrics. Below each of these main metrics, the worker would also see, from right to left: the number of expected units that will need to be produced, keeping in mind that some of them will need to be refactored because of quality issues; the expected running time of the machines, that is, the amount of time that the machines will need to be running; and, lastly, the expected amount of time that the machines should be left available, knowing that during that time some non-planned stops will occur.

Finally, on the bottom of the page the worker would see some the values for some of the variables that affect that individual work order. Under the gauge there is the actual value for the production order, on each of the sides the minimum and maximum values observed in history and, marked in

yellow, the average value. This way the worker can easily see if the observed value is above or below average and take actions, if necessary.

This dashboard would be especially helpful in preventing putting into production orders that are predicted to perform exceptionally bad and which could have an easy fix: maybe there has not been a stop in a long time or the planned number of units to produce is too high.

## 5.5 PRODUCTION ORDER PLANNING

This last dashboard, shown in Figure 21, has been thought with the production order planner in mind.

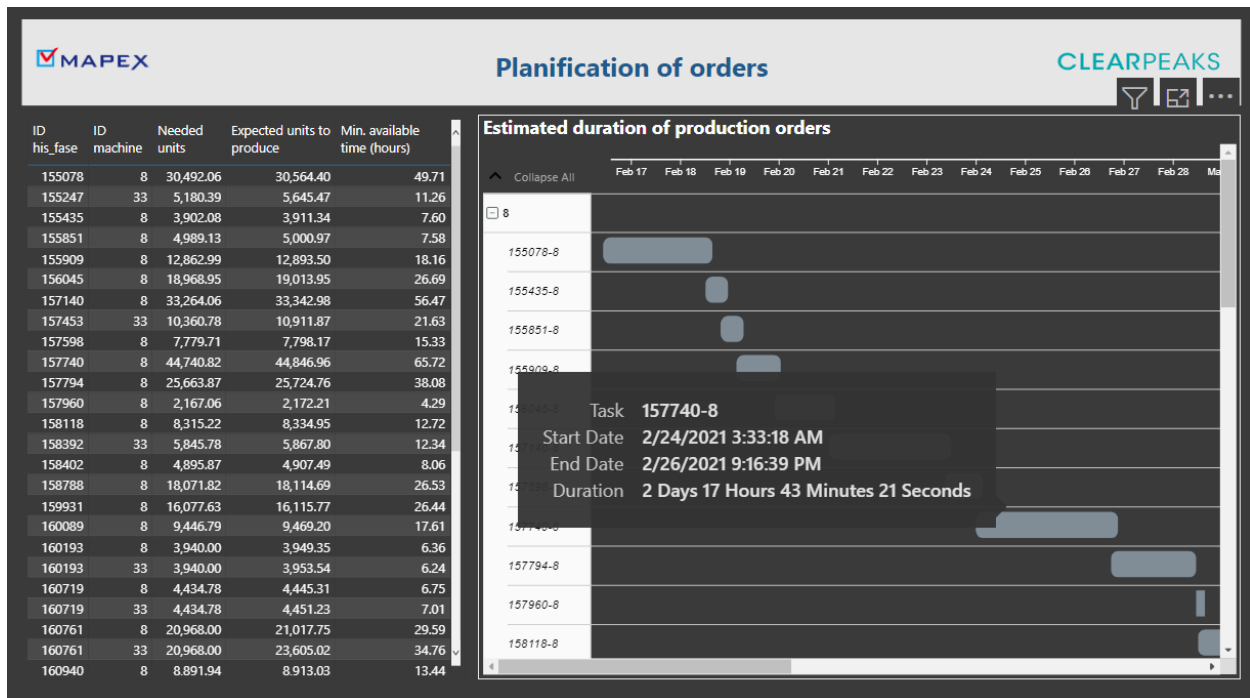


Figure 21 Seventh dashboard: Production order planning

On the left side, there is a table that shows, for each of the production orders, the number of needed units as it appeared in the original planification, the expected number of units that will need to be produced having in mind that some of them will need to be refactored, and the expected amount of time that the machines should be left available in order to be able to complete the production orders.

On the right side, there is a Gantt diagram that shows the beginning and end date of each of the work orders and, if you hover the mouse on any of the tasks, you can see the exact expected duration.

Because the functionality of this dashboard was out of the scope of this project, there are still so many features that could be added to it in order to make it even more useful. To make a good planification, some variables would need to be taken into consideration, like the duration of the workday, the number of available employees, the times in which the factory will be closed, etc. Nonetheless, it still shows the benefits that a fully developed tool like this could bring and opens the door for next steps and future work.

# 6 PROPOSED SOLUTION: THE PIPELINE

This section aims to integrate and summarize all the different components that have been presented until now into a unique Proof of Concept, ready for its validation (section 7).

Figure 22 shows a diagram of all the components of this solution and how they are related between them:

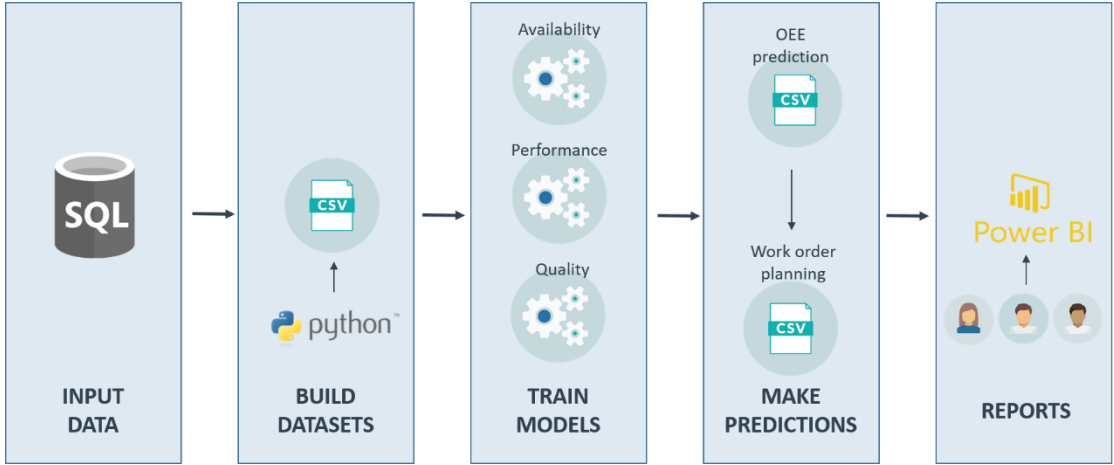


Figure 22 Pipeline of the proposed solution

At the start of the project, all there was available was a SQL Server database with the data from one of Mapex’s clients. To obtain the datasets needed for training, a pipeline was designed using Python, which connects to the database by using the pyodbc module and applies the necessary transformations until the final datasets for modelling are obtained (section 3.2).

Next, an AutoML pipeline has been implemented in order to obtain the trained models for each of the metrics. This pipeline automates the process of feature selection and hyperparameter optimization for a series of baseline models. That is an important contribution to this PoC, since it allows to implement the solution to clients which may not have a data scientist on-site able to monitor the models. The structure of the developed code can be found in Appendix B.

Once the models have been trained, they are used to make predictions of future OEE values for the test set, and a planification for the expected duration of the work orders based on the estimated OEE values (section 5).

Finally, the solution is presented in some Power BI reports that allow for simple interpretation (section 6).

## 7 VALIDATION OF THE TOOL

To be able to obtain an assessment of the success of the provided tool, a questionnaire was provided to the director and co-director of the project and the client representative. The questionnaire consisted of a list of statements to which the respondents had to give an answer using a scale from 1 to 5, in which 1 meant that they strongly disagreed with the statement, and 5 that they strongly agreed with it. The questions could be left blank if the respondent felt like the answer was not applicable. These statements aimed to respond to some of the key concepts of the “Technology Acceptance Model” [13]: usefulness, ease of use, quality of output and intention of use of the tool.

The table below summarizes the answers obtained by the three participants:

Statement	N	Min	Max	Mean	Median	Mode	Std
Using this tool would help the clients get a clearer understanding of how the OEE behaves in their own factories	3	4	5	4.67	5	5	0.47
Using this tool would help improve the outcome of the OEE.	3	3	5	4	4	-	0.81
Using this tool would help reduce the costs related with bad OEE results.	3	3	5	4	4	-	0.81
Using this tool would help the clients see which are the strong and weak points of their factories.	3	4	5	4.67	5	5	0.47
The interaction with the tool is clear and understandable.	3	5	5	5	5	5	0
I find the tool easy to use.	3	2	5	3,67	4	-	1.24
I find it easy to find the information I am looking for in the dashboards.	3	4	5	4.67	5	5	0.47
The quality of the output I get from the tool is high.	3	2	4	3	3	-	0.82
If I had access to the tool, I would use it.	2	5	5	5	5	-	0
I plan to use the tool in the next months.	3	4	5	4.67	5	5	0.47

*Table 3 Summary of results obtained from the validation questionnaire*

Moreover, the questionnaire had three open questions with the aim of getting more feedback on the proposed tool and its improvement points.

When asked about whether the proposed solution met the initial expectations, Mapex’s representative answered that it did, as it demonstrated the viability of the connection between the

two systems: Mapex's software and the advanced analytics layer proposed in this project. The director of the project also agreed that the expectations were met, and even though there were some limitations due to the original dataset provided, the implementation of the solution was the one expected. One of the weak points of the proposed solution was the low accuracy of the models, which makes the analysis a little unreliable. On a positive side, there was an extra functionality added, the planning of the production orders.

Another important point during the development of this project was that it could be used as a Proof of Concept, which would attract potential Mapex clients into providing their data and acquiring an advanced analytics module. All participants agreed that the proposed solution could be used to attract clients to implement this solution to their environment. Some positive points added by the co-director of the project included the benefits of the tool being compatible with Mapex's data model and being packaged as an add-on on top of it.

Finally, participants were asked about the possible improvement points to the proposed dashboards. To that question, the project director answered that there could have been added product and client dimensions to the reports. The co-director of the project added that more filtering options could be added to the dashboards, as well as more information on the interpretability of the models.

Overall, the response from both the project managers and Mapex was positive. They saw that the tool had the potential of attracting their clients into providing their data to add this advanced analytics layer into their current environment. The proposed dashboards seemed to meet the principal objectives measured in the "Technology Acceptance Model": usefulness, ease of use, quality of output and intention of use of the tool. Nonetheless, during a formal presentation of the project to some of the C-Level employees at Mapex, some improvement points were identified. The most important metric to predict for one of their most important clients is the quality one, and sensorics data should be used to build that model. Even though this data was not available for this project, it should be a focus point going forward. Lastly, a more in-depth explanation should be provided about the key influencers that affect the three metrics: availability, performance and quality.

## 8 CONCLUSIONS

---

This section offers a summary of the work done in this project, its main contributions and the identified improvement points and next steps.

The Overall Equipment Effectiveness is a widely used metric in the manufacturing industry that can be used to track and bench-mark the effectiveness of production processes. As of now, Mapex has been providing this metric to their clients to help them explain the results obtained on their factories. To be able to improve on the planning of work orders and reduce costs during production, Mapex commissioned us to add an Advanced Analytics layer to their software that would be able to predict the OEE values for the next production orders. This project was born as a Proof of Concept from ClearPeaks to Mapex to show that it is possible to make such predictions and present them in a way that is attractive and easy to understand to their clients, in order to convince them to acquire this new layer of analytics.

Therefore, in this project we studied the Overall Equipment Effectiveness of manufacturing processes by the creation of three predictive models to individually predict each of the metrics that compose the OEE: availability, performance and quality. Also, to be able to visualize the results obtained by the models and get a better understanding of the data and the key influencers, a set of dashboards was provided using Power BI.

To create each of the models that would be used to predict the expected OEE values, an AutoML pipeline was designed with the aim of making the code easy to use and as automatized as possible, so that any client can benefit from this solution even if they do not have a data scientist on-site. The pipeline uses forward feature selection and hyperparameter optimization by grid search in order to always obtain the model that gives the best results, without the need of human supervision.

Using the models obtained after the execution of this pipeline, a prediction of the OEE value for each production order and machine was provided, as well as a planification of the duration of such work orders based on the expected OEE obtained by the models. This functionality, although it was not considered as a requirement for the Proof of Concept, ended up serving as a new use case which provided a lot of value to the proposed solution. By knowing an estimation of the units that will need to be refactored or the minimum amount of time that the machines should be left available for, the clients could directly use this information to calculate the impact that OEE would

have on their cost of production. They can use this information to improve the planning of their production orders, which would positively affect the wellbeing of the whole supply chain.

Finally, some Power BI dashboards were provided which showed the prediction of OEE for the next production orders, the data used to make such predictions, some visualizations to obtain better insights on what affects the OEE and a Gantt diagram to show how the next production orders should be planned considering the predicted OEE values.

The dashboards were presented to the client, Mapex, and the overall feedback was positive. The tool was found to be useful and easy to use and the initial expectations for this Proof of Concept were met. After the validation of the tool, it was clear that the dashboards could be used to attract potential Mapex clients to provide their data and implement this new module into their current architecture, which was a goal of the project. Nonetheless, some improvement points were identified: the reports should have more filtering options and show some more visualizations in order to obtain better insights of which are the key influencers when it comes to OEE. Also, the accuracy of the models was not as high to make the tool reliable, so it would need to be improved.

In sight of this, a set of next steps can be defined for this project.

- Improve the models. One of the reasons why the error on the models is high could be that there are not enough observations of low values of the metrics, so the models have a hard time identifying what makes a bad production order. To make the predictions more accurate, more data would be needed, especially observations that belong to production orders with bad OEE results, which are exactly the types of production orders that we are interested in predicting well.
- Add new visualizations in the dashboards that give better insights of the main influencers of OEE.
- Obtain a dataset with sensorics data and train the quality metric using these types of variables.
- An enablement session should be provided to Mapex to show them how to demonstrate this tool to their clients.

## 9 ACKNOWLEDGEMENTS

---

I would like to thank ClearPeaks for giving me the opportunity to work on this project, and for the incredible learning experience that working here as an intern for the past two years has been. More specifically, I want to thank Jordi Ricart, who proposed and trusted this project to me, and Martí Soler, who helped me with all the technical aspects of it. Your help, guidance and support during these last few months has been highly appreciated.

Also, this project would not have been possible without the guidance of Xavier Coll from Mapex, with whom weekly meetings were held to discuss the state of the project. Thank you, Xavier, for always being so helpful, answering all the questions that I had along the way.

I would also like to thank Silverio Martínez-Fernández for helping me with the documentation, always giving me ideas of how to improve it and making sure that all the work that has been done was correctly reflected in this document.

Finally, I want to give a special thank you to my classmates and friends, Júlia Sala and Zoé Meini, and to family, especially to my mother, Rosa, for always being by my side and for the moral support.

## 10 REFERENCES

---

- [1] Mehta, B.R. & Reddy, Y.J.. (2015). Manufacturing execution systems. 10.1016/B978-0-12-800939-0.00023-1.
- [2] Seiichi Nakajima. (1982) TPM tenkai.
- [3] Okpala, Charles & Anozie, Stephen. (2018). Overall Equipment Effectiveness and the Six Big Losses in Total Productive Maintenance.
- [4] OEE Factors, Understand Availability, Performance, Quality | OEE. (2002–2019). oee.com. <https://www.oee.com/oee-factors.html>
- [5] Hassani, Ibtissam & Mazgualdi, Choumicha & Masrour, Tawfik. (2019). Artificial Intelligence and Machine Learning to Predict and Improve Efficiency in Manufacturing Industry.
- [6] Anusha, Chintada & Umasankar, Venkata. (2020). Performance Prediction through OEE-Model. International Journal of Industrial Engineering and Management. 11. 93-104. 10.24867/IJIEM-2020-2-256.
- [7] (Last accessed: 23-06-2021) Python's module pyodbc. <https://github.com/mkleehammer/pyodbc/wiki>
- [8] Ali, Jehad & Khan, Rehanullah & Ahmad, Nasir & Maqsood, Imran. (2012). Random Forests and Decision Trees. International Journal of Computer Science Issues(IJCSI). 9.
- [9] Chen, Tianqi & Guestrin, Carlos. (2016). XGBoost: A Scalable Tree Boosting System. 785-794. 10.1145/2939672.2939785.
- [10] Ayyadevara, V. (2018). Gradient Boosting Machine. 10.1007/978-1-4842-3564-5\_6.
- [11] (Last accessed: 23-06-2021) Feature Selection. [https://scikit-learn.org/stable/modules/feature\\_selection.html](https://scikit-learn.org/stable/modules/feature_selection.html)
- [12] He, Xin & Zhao, Kaiyong & Chu, Xiaowen. (2021). AutoML: A survey of the state-of-the-art. Knowledge-Based Systems. 212. 106622. 10.1016/j.knosys.2020.106622.

[13] Venkatesh, Viswanath & Bala, Hillol. (2008). Technology Acceptance Model 3 and a Research Agenda on Interventions. Decision Sciences - DECISION SCI. 39. 273-315. 10.1111/j.1540-5915.2008.00192.x.

[14] Chapman, Pete & Clinton, Julian & Kerber, Randy & Khabaza, Tom & Reinartz, Thomas & Shearer, Colin & Wirth, Rüdiger. (1999). CRISP-DM 1.0 step-by-step data mining guide.

[15] (Last accessed: 23-06-2021) Cookiecutter Data Science.

<https://drivendata.github.io/cookiecutter-data-science/>

# APPENDIX

---

## A. PROJECT PLAN

To be able to reach the objectives of this project, a project plan has been designed inspired by the CRISP-DM [14] methodology, with the task breakdown below:

- **Initial phase:** it comprises the installation of the software and database, the understanding of the industrial need, the analysis of data and the definition of the requirements and objectives of the project.
  
- **Development phase:**
  - **Modelling:** to fulfil the first subobjective, a predictive model will need to be developed for each of the three metrics. The main tasks during their development will be: feature engineering, modelling, assessment of models and documenting. The interpretability of these models will also be a relevant task that will help provide insights on OEE.
  - **Reporting:** to fulfil the second subobjective, a report will need to be created to successfully represent all the insights obtained during the study of the data as well as the results of the models.
  
- **Validation phase:** once the modelling and reporting are done, the reports will be shown to members of the company and the client to ask for feedback and improvement points to make sure that the usability of the tool is being exploited to its fullest extent.
  
- **Documentation phase:** although the documenting of this project will be an ongoing task during its development, some time will be allocated to it on the final weeks to make sure that all the work that has been done is correctly documented and reported.

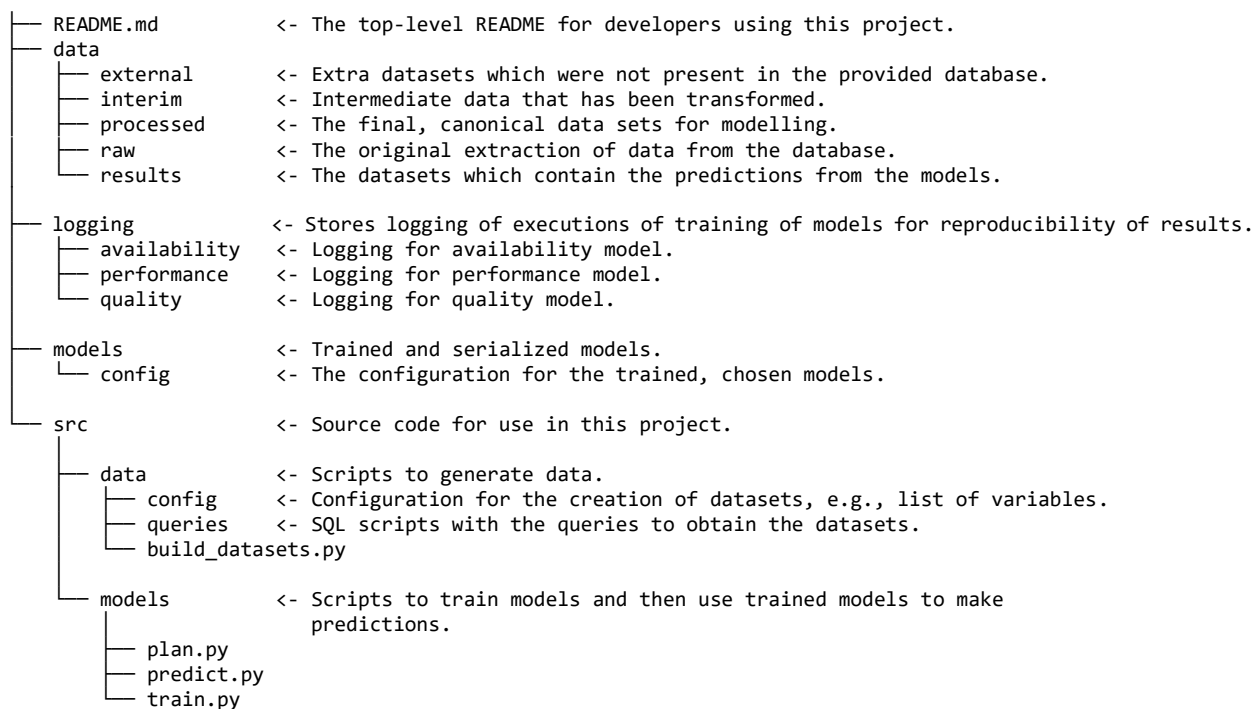
## B. REPLICATION PACKAGE

This subsection explains how the scripts used for the development of this project have been structured.

When developing a data science project, it is of extreme importance to keep the code quality high, to make sure that any developers that may continue your work in the future can reproduce and replicate your results, for example.

For that matter, an adaptation of the CookieCutter Data Science [15] project structure was used. It provides a template for the structure of the directories that hold all your files.

The diagram below summarizes the structure of the code and datasets:



For confidentiality reasons, the code developed during this project is not public, but private access can be requested for the grading of this thesis if necessary.

### C. ECONOMIC VIABILITY

This subsection describes the economic viability of this project.

The cost of this project can be analysed from two different points of view: the investment that ClearPeaks has made to be able to create this Proof of Concept and the cost that the implementation of this solution would have for a new client.

When it comes to the actual cost of the creation of this PoC, only personnel costs are considered, since the software and tools needed to develop the solution were already available at ClearPeaks.

The human resources involved for the development of this project consisted of a Business Intelligence (BI) Analyst who developed the solution, a BI Consultant who took the lead as the technical manager of the project, a project manager for the overall wellbeing of the project and the communication with the client, and the Mapex representative with whom weekly meetings were held to discuss the progress of the project. The table below summarizes the costs associated with the use of this personnel resources, considering the rates that clients are usually charged for, depending on the job title.

Job title	Dedication (man-days)	Daily rate	Cost
BI Analyst	62	400 €	24800 €
BI Consultant	5	700 €	3500 €
Project manager	2	800 €	1600 €
Matter expert (Mapex representative)	2	500 €	1000 €
<b>TOTAL COST</b>			<b>30900 €</b>

*Table 4 Summary of estimated costs of the proposed Proof of Concept*

Although this cost may seem quite high, it is considered as an investment by ClearPeaks. This project has been developed as a Proof of Concept and it is general enough that with small modifications to the code or models, it could be implemented in Mapex clients’ environments quickly. This means that the solution can be used in many different scenarios with little effort.

Let us now make an estimation of what would the cost be for a client that wants to implement this solution in their environment. The table below shows the costs associated with the adaption of the tool to a new environment, considering that only one person with a daily rate of 500 € is dedicated to these tasks.

<b>Task</b>	<b>Duration (man-days)</b>	<b>Cost</b>
Adapt queries to obtain data from the new database	2	1000 €
Analyse data and obtain insights	10	5000 €
Re-train models	10	5000 €
Adapt dashboards to new data	3	1500 €
<b>TOTAL COST</b>		<b>12500 €</b>

*Table 5 Summary of costs of adoption of tool in new environment*

The most time-consuming task would be to analyse the new data and obtain insights that can be later shown in the dashboards as well as re-training the models. Even though the current pipeline always chooses the best performing model, the first time that the models are trained there should be a more intensive study of which baseline models should be used, whether new features need to be engineered or not, which hyperparameter grid is the most appropriate, etc.

Moreover, the potential new clients may want to make some modifications to the proposed tool to adapt it better to their needs. Some examples of tasks that these clients could ask for are:

- Create a quality model based on the sensorics data. In this case, data would need to be collected and analysed and the model would have to be trained with the necessary adjustments to the hyperparameter grid. This would take at least 15 man-days, which would have a cost of 7500 €.
- Create additional dashboards with more in-depth data analysis. Dedicating 10 man-days to this task, it would have a cost of 5000 €.
- Optimization of the production order planner. This would consist of making the production order planner more intelligent by optimizing the workload on the machines or considering non-working days in the Gantt planification, for example. For the described tasks, 20 man-days would be needed, with an associated cost of 10000 €.

To summarize, a basic implementation of the tool to a new client would cost them about 12500 € and adding new functionalities to it would make that cost increase.

It is important to note that an accurate prediction of the OEE values could translate in a big diminishment of costs for Mapex's clients, so the Return on Investment that the clients would get from it would be high.

## **D. ETHICAL IMPLICATIONS**

This subsection aims to describe how this project can be related with the Sustainable Development Goals set by the United Nations (UN).

One of the objectives of this project has been to build predictive models that are able to determine the Overall Equipment Effectiveness of manufacturing processes. An accurate OEE prediction and the analysis of the main contributors to bad manufacturing results would lower the costs of production by lowering waste, both in materials and energy, improve the planning of the production orders and the whole supply chain, identify and bench-mark progress, etc. Being able to reduce these costs and learning how to make the factories more efficient would have many benefits to our environment, which can be related to some of the goals set by the UN:

- Goal 9: Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation.
- Goal 12: Ensure sustainable consumption and production patterns.

Adding an analytics layer to the software that is used in production plants can be thought of as a form of innovation, and it would help make the factories more sustainable, for example: the waste could be lowered if the key contributors to bad quality results can be identified, and the energy consumption could be diminished if one can avoid starting machines with a high probability of failure by doing maintenance jobs before that happens.