A Data-Driven Decision-Support Framework for Predictive and Preventive Maintenance in a Continuous-Process Plant

João Duarte de Carvalho Antunes

Master's Dissertation

Supervisor: Prof. Masoud Golalikhani

Master in Mechanical Engineering

July 3, 2026

Abstract

Maintenance in continuous-process industry is both costly and consequential. Production lines are tightly coupled, so a single unplanned stoppage propagates quickly, and a large share of preventive effort is misdirected.

This dissertation, carried out with Sonae Arauco at its Mangualde medium-density fibreboard plant, addresses two linked problems. First, four years of computerised maintenance records serve only retrospective reporting, leaving the team without early warning of failures. Second, the preventive plans, many inherited from another site after a merger, have never been calibrated against the plant's own failure history. The aim is to turn these existing records into guidance the maintenance team can act on, without new instrumentation.

A decision-support framework is developed that brings three systems together on a single data foundation. A machine-learning layer estimates, for each equipment location, a calibrated probability of corrective work within forty-five days, presented as a ranked and explained list of alerts. Under a leakage-free out-of-time evaluation, it concentrates the real failures at the top of the ranking, at a level in line with comparable event-log studies that use no condition sensors. A reliability layer then characterises each location's failure process and judges whether its preventive plan is adequate. It reveals substantial over-maintenance, even under a conservative threshold, alongside plans that absorb effort without reducing failures. Where wear-out is statistically confirmed, it recommends a replacement interval and a quantified annual saving against current practice. A dashboard consolidates the three systems and is designed to align with the plant's daily data cycle.

Validation is convergent: a prospective backtest, a population-level hazard model, and the judgement of the plant's engineers all agree, and the exercise itself improved recording practice. The work shows that substantial maintenance value can be drawn from the records a plant already keeps, and that the surest route to extending it is richer data rather than a more complex model.

Resumo

A manutenção na indústria de processo contínuo é simultaneamente dispendiosa e crítica. As linhas de produção estão fortemente acopladas, pelo que uma única paragem não planeada se propaga rapidamente, e uma parte considerável do esforço preventivo está mal direcionada.

Esta dissertação, realizada em colaboração com a Sonae Arauco na sua fábrica de aglomerado de fibras de média densidade (MDF) em Mangualde, aborda dois problemas interligados. Em primeiro lugar, quatro anos de registos informatizados de manutenção servem apenas para relatórios retrospetivos, deixando a equipa sem aviso prévio de falhas. Em segundo lugar, os planos preventivos, muitos herdados de outra unidade após uma fusão, nunca foram calibrados face ao historial de falhas da própria fábrica. O objetivo é transformar estes registos existentes em orientação sobre a qual a equipa de manutenção possa atuar, sem nova instrumentação.

Desenvolve-se uma estrutura de apoio à decisão que reúne três sistemas sobre uma única base de dados. Uma camada de aprendizagem automática estima, para cada localização de equipamento, uma probabilidade calibrada de intervenção corretiva num horizonte de 45 dias, apresentada como uma lista ordenada e explicada de alertas. Sob uma avaliação fora do tempo (*out-of-time*) e sem fuga de dados, esta camada concentra as falhas reais no topo da ordenação, a um nível alinhado com estudos comparáveis baseados em registos de eventos que não utilizam sensores de condição. Uma camada de fiabilidade caracteriza em seguida o processo de falha de cada localização e determina se o respetivo plano preventivo é adequado. Revela um excesso de manutenção substancial, mesmo sob um critério conservador, a par de planos que consomem esforço sem reduzir falhas. Onde o desgaste é estatisticamente confirmado, recomenda um intervalo de substituição e uma poupança anual quantificada face à prática atual. Um painel (*dashboard*) consolida os três sistemas e foi concebido para se alinhar com o ciclo diário de dados da fábrica.

A validação é convergente: um *backtest* prospetivo, um modelo de risco ao nível da população e o parecer dos engenheiros da fábrica concordam entre si, e o próprio exercício melhorou as práticas de registo. O trabalho demonstra que é possível extrair valor substancial para a manutenção a partir dos registos que uma fábrica já mantém, e que a via mais segura para o ampliar passa por dados mais ricos e não por um modelo mais complexo.

Acknowledgements

I would like to thank Sonae Arauco for the opportunity to work with them, for all their support throughout this project, and for their readiness to provide any material I needed.

I am also deeply grateful to my supervisor at the Faculty of Engineering of the University of Porto, Professor Masoud Golalikhani, for all the guidance, support, and valuable insights throughout this dissertation.

My thanks also go to my supervisor at the company, Marco Ferreira, for all his help and support during the project.

Finally, I wish to express my sincere gratitude to my family and friends for being by my side on this journey; without you, none of this would have been possible. And to my girlfriend, for the constant support and help at every moment.

To all of you, thank you.

Contents

Acronyms

CI	Confidence Interval
CM	Corrective Maintenance
CMMS	Computerised Maintenance Management System
HR	Hazard Ratio
IID	Independent and Identically Distributed
KPI	Key Performance Indicator
MDF	Medium-Density Fibreboard
MLE	Maximum Likelihood Estimation
MTBC	Mean Time Between Corrective Events
MTBF	Mean Time Between Failures
MTBP	Mean Time Between Preventive Executions
PHM	Proportional Hazards Model
PM	Preventive Maintenance
PR-AUC	Area Under the Precision-Recall Curve
RCM	Reliability-Centred Maintenance
ROC	Receiver Operating Characteristic
ROC-AUC	Area Under the Receiver Operating Characteristic Curve
RUL	Remaining Useful Life
SFC	Shop Floor Control
SHAP	SHapley Additive exPlanations
SWOT	Strengths, Weaknesses, Opportunities and Threats
UML	Unified Modelling Language

List of Figures

List of Tables

Chapter 1

Introduction

The present dissertation was developed within the Master's Programme in Mechanical Engineering, with specialisation in Production Management, of the Faculty of Engineering of the University of Porto (FEUP), in partnership with Sonae Arauco. It addresses a practical problem at one of the company's plants: turning the maintenance records the plant already collects into early warning of equipment failures and into evidence on the adequacy of its preventive maintenance policy.

1.1 Background and motivation

Maintenance is the set of technical, administrative, and managerial actions that keep an item fit for its required function, or restore it to that state, across its life cycle (EN 13306, 2017). In capital-intensive industry the cost is high: depending on the industry, maintenance is estimated to absorb between 15% and 60% of the cost of goods produced, with heavy industries such as iron and steel or pulp and paper at the upper end of that range (Mobley, 2002). Surveys of maintenance effectiveness further indicate that about one third of this expenditure is wasted on unnecessary or improperly performed maintenance (Mobley, 2002). In continuous-process plants the stakes are higher still, since the lines are tightly coupled and a single unplanned stoppage propagates fast, at considerable cost for each hour of lost output. Containing such stoppages, while avoiding wasted preventive effort, is therefore a central concern of the maintenance function.

Poor maintenance strategies are estimated to reduce an asset's overall productive capacity by 5% to 20%, and unplanned downtime is estimated to cost industry around $50 billion each year (Deloitte, 2025). The opportunity is correspondingly large: McKinsey reports that digitally enabled maintenance and reliability transformations in heavy industries can raise asset availability by 5% to 15% and cut maintenance costs by 18% to 25% (Bradbury et al., 2018). Predictive maintenance is, accordingly, the largest industrial application of artificial intelligence, accounting for about a quarter of the industrial AI market and ranking ahead of quality inspection and assurance (IoT Analytics, 2019).

Maintenance is conventionally divided into two families, corrective and preventive, according to whether action follows or precedes failure. Figure 1.1 summarises this classification.

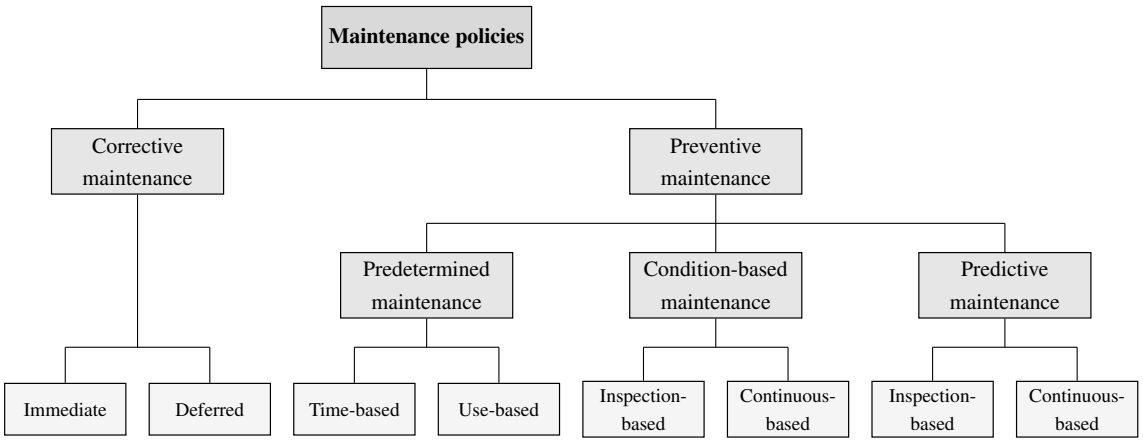

Figure 1.1: Classification of maintenance policies. Adapted from Guimarães and Leitão (2025), based on EN 13306.

Corrective maintenance is carried out after a fault has been recognised, to restore the required function. It is immediate when a breakdown forces an emergency repair, and deferred when a fault is detected but the repair is postponed to a more convenient time (EN 13306, 2017). Preventive maintenance, by contrast, acts before failure to reduce its probability. EN 13306 separates it into predetermined maintenance, which follows fixed intervals of time or usage without examining the equipment, and condition-based maintenance, which is governed by monitoring or inspection of the equipment's actual state; predictive maintenance is the condition-based form in which the intervention follows a forecast of how the significant degradation parameters are expected to evolve (EN 13306, 2017). Given its growing importance, predictive maintenance is commonly presented as a category in its own right, as in Figure 1.1.

Predictive maintenance is the focus of the present work. Its recent rise is bound up with artificial intelligence, machine learning, and data analytics, which turn operational data into early indications of failure. By flagging likely breakdowns before they occur, it tells the maintenance team what is degrading, when intervention is due, and how the repair should be carried out, reducing downtime and raising production efficiency at lower cost (Erbiyik, 2022). Several condition-monitoring techniques put this into practice: failure early warning systems, vibration analysis, infrared thermography, acoustic emission, oil and particle analysis, corrosion monitoring, and performance monitoring (Erbiyik, 2022). This work adopts the first, a failure early warning system built on the failure and intervention records already held in the plant's computerised maintenance management system (CMMS).

1.2 The industrial partner and project context

The work was carried out in partnership with Sonae Arauco, a Portuguese manufacturer of wood-based panels. The company employs around 2,600 people across 23 industrial and commercial units in 9 countries, has a production capacity of 3.9 million cubic metres of wood-based solutions, and sells its products in 68 countries. Its product range covers three panel families that differ in

how the wood is bonded: medium-density fibreboard (MDF), pressed from defibrated wood fibres into a dense and homogeneous panel; particleboard (PB), formed from wood chips and particles bonded under heat and pressure for general furniture and construction use; and oriented strand board (OSB), made from layered strands deliberately aligned to give higher mechanical strength for structural applications.

This work focuses on the plant in Mangualde, an MDF production site designated throughout this dissertation by the internal site code P058. The site is a continuous-process facility that operates multiple production lines, each built from chains of electromechanical equipment such as pumps, motors, conveyors, presses, and cutting machines, supported by auxiliary systems for water, steam, and hydraulics. Maintenance is managed by the site's Asset and Maintenance Management team through IBM Maximo, a widely deployed CMMS that records every corrective work order, preventive schedule, and condition evaluation in structured form. More than four years of operational history have accumulated in the system, beginning in 2020 and reaching reliable, consistent coverage from 2021.

1.3 Problem statement

Several recurring problems at site P058 motivated this work. The first problem is that the data already accumulated in the CMMS are barely exploited for predictive purposes. The corrective work orders and preventive schedules recorded there serve retrospective reporting rather than anticipation. As a result, the maintenance team is left without early indication of likely failures in critical production equipment and without guidance on where to direct inspection effort before a breakdown occurs.

The second problem is closely related and concerns the preventive policy itself, which is poorly calibrated. The team has no systematic means of judging whether the preventive frequency assigned to each location is adequate, too low to prevent failures, or excessive relative to the failure history. During the project's scoping, technicians and engineers reported numerous locations where the preventive plan was misaligned with the failure history. Many preventive plans were transferred largely unchanged from another plant after the merger with Arauco, with little adjustment to the local equipment. Recalibrating them empirically, across thousands of locations and without dedicated tools, is a slow and labour-intensive task that has so far remained out of reach.

These two problems map onto the three analytical systems that the framework develops. The absence of early warning is met by a failure prediction system, which ranks the equipment by its short-term risk so that inspection is directed to the most exposed items before a breakdown occurs. The miscalibrated preventive policy is addressed in two complementary ways: a plan adequacy system, which sets the preventive frequency of each location against its failure history to reveal where prevention is excessive or insufficient; and an interval optimisation system, which, for the most critical equipment whose failures show genuine wear, recommends the replacement interval that the failure history justifies. Together these systems turn the recorded history into the early warning and the policy evidence the plant lacks. The framework groups them into two

analytical layers: the prediction system as the predictive layer, and the adequacy and interval systems together as the reliability layer. The objectives defined next follow this division.

1.4 Objectives and methodology

The two problems identified above define the aim of this dissertation: to develop and validate a data-driven decision-support framework that converts the plant's accumulated maintenance records into guidance the maintenance team can act on, anticipating equipment failures over a short horizon and assessing whether the current preventive policy is matched to the observed failure behaviour. This aim is pursued through five specific research objectives (RO).

RO1: Diagnose the plant's maintenance data environment and assess its fitness to support a data-driven maintenance policy.
Any analytical layer is only as reliable as the records beneath it. The first objective is therefore to examine the systems that hold the plant's maintenance history, the structure and granularity of the records they contain, and the limits those records impose, in order to establish what can be exploited for failure anticipation and for policy assessment. This diagnosis shapes every methodological choice made in the remainder of the work.

RO2: Develop a predictive layer that anticipates short-term equipment failures.
The second objective is to supply the early warning that the maintenance team currently lacks. A machine-learning pipeline is to be developed that estimates, for each equipment location, a calibrated probability of corrective intervention within a short horizon and presents the result as a ranked list of alerts, each accompanied by the factors that explain it. The pipeline must be reproducible and constructed so that no information unavailable at prediction time can influence the model. Because not all corrective work is equally predictable, the objective also covers delimiting the model's scope, including the distinction between unplanned and planned corrective work and the level of aggregation at which prediction is most reliable, so that the model is applied where the records support it.

RO3: Develop a reliability layer that assesses the adequacy of the preventive policy and recommends intervention intervals.
The third objective addresses the calibration of the preventive plans. A statistical layer is to be developed that characterises the failure process of each location, compares the preventive frequency assigned to it with the failure history actually observed, and flags the locations that are over-maintained, where prevention is frequent yet failures are absent, and under-maintained, where failures recur yet prevention is sparse. Where the data support it, this layer should also recommend an intervention interval, testing the applicability of classical interval-optimisation models case by case rather than assuming that a single model holds across the entire equipment population.

RO4: Consolidate both layers into a decision-support tool for the maintenance team.
Analytical results only influence maintenance practice if they reach the team in a usable form. The fourth objective is therefore to combine the outputs of the predictive and reliability layers in

a single decision-support dashboard, designed to refresh and issue alerts on the plant's daily data cycle so that the guidance can stay in step with the maintenance records as they are logged.

RO5: Validate the framework on the plant's operational history and derive managerial and methodological insights.
The final objective is to apply the framework to the plant's full operational history, evaluate the quality of its predictive and reliability outputs, and translate the results into insights for the maintenance team. It also includes documenting the investigative paths that were tested and rejected during development, so that the work delimits what does and does not work in this class of problem, both for the plant and for comparable industrial settings.

These objectives are pursued through three methodological steps, summarised in Figure 1.2 together with the objectives that each step addresses.

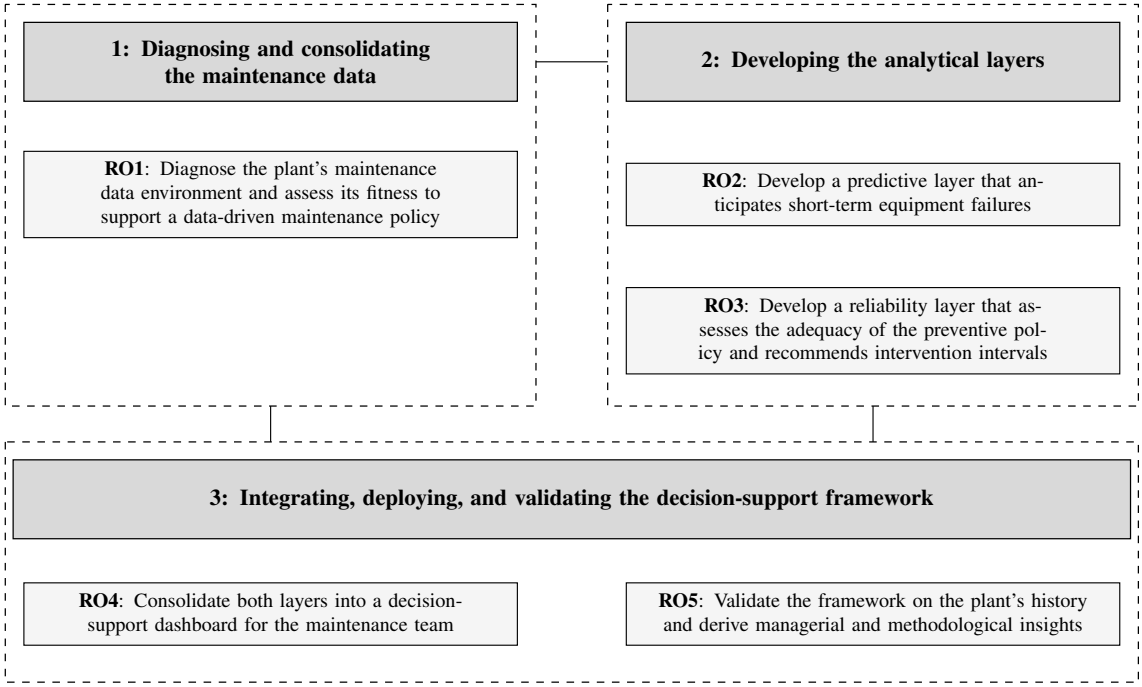

Figure 1.2: Objectives of the dissertation and the methodological steps that address them.

The first step diagnoses the data environment and consolidates the relevant records into a single analytical dataset, resolving the differences in coding and granularity between the source systems (RO1). The second step develops the two analytical layers on this dataset: the predictive layer, through feature engineering, model selection, temporally consistent validation, and probability calibration; and the reliability layer, through the statistical characterisation of each location's failure process, the diagnosis of preventive adequacy, and the recommendation of replacement intervals where confirmed wear-out justifies them (RO2 and RO3). A series of methodological investigations, concerning the level at which prediction is reliable, the definition of the target event, and the statistical conditions each reliability model must meet, shapes the design of both layers. The third step consolidates the layers into the decision-support dashboard, designed to refresh and

report on the plant's daily data cycle, and evaluates the framework on the plant's historical data and through feedback from the maintenance team (RO4 and RO5).

1.5 Structure of the dissertation

The remainder of this dissertation is organised in five chapters, each tied to the objectives set out in the previous section; Figure 1.3 gives an overview of this structure.

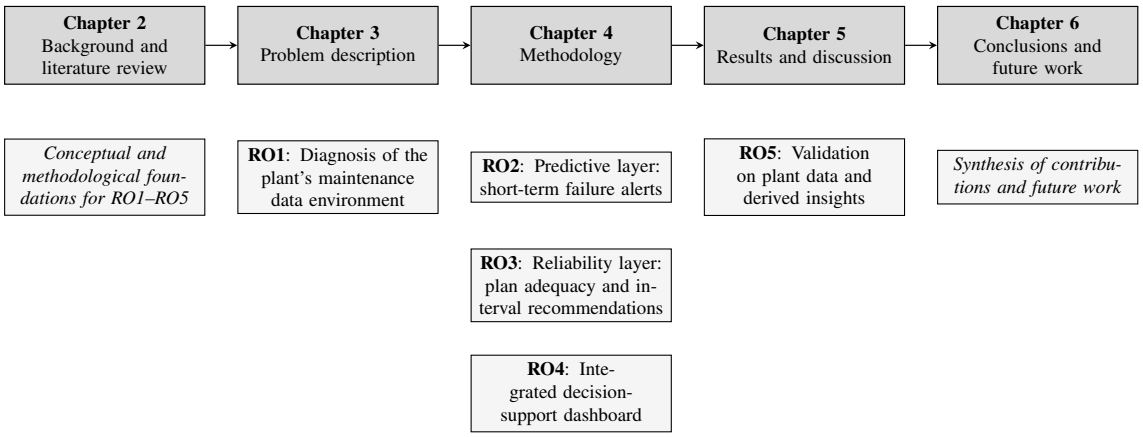

Figure 1.3: Structure of the dissertation and its mapping to the research objectives.

Chapter 2 reviews the literature on which the work rests: maintenance strategies, machine learning for failure prediction, the reliability of repairable systems, and the maintenance optimisation and hazard models built on it. The review establishes the conceptual foundations for the five objectives and identifies the gap that this dissertation addresses.

Chapter 3 characterises the case study and its data environment, in line with RO1. It describes the production process and its equipment, the two systems that hold the maintenance-relevant records, the way equipment is coded and located, and the structural gap between the two systems that leads this work to build on the Maximo records. The chapter closes with an analysis of strengths, weaknesses, opportunities, and threats (SWOT) that consolidates the diagnosis and motivates the methodology.

Chapter 4 develops the framework itself, covering objectives RO2 to RO4. It presents the preparation of the data, the predictive machine-learning layer together with the methodological investigations that shaped it, the statistical reliability layer, and the integration of both layers into the decision-support tool delivered to the plant.

Chapter 5 reports and discusses the consolidated results of applying the framework to the plant's operational history, addressing RO5. Finally, Chapter 6 synthesises the contributions and limitations of the work and outlines directions for future research.

Chapter 2

Background and Literature Review

This chapter establishes the foundations on which the work rests and positions it within the literature. Section 2.1 develops the economic and operational reasoning that leads from reactive repair to predictive and prescriptive maintenance, and identifies the conditions under which each strategy is justified. Section 2.2 reviews machine learning approaches to failure prediction in industrial settings, organised by the kind of data each approach requires. Section 2.3 presents the statistical theory of repairable systems, the classical maintenance-optimisation models that support interval recommendations, and the hazard models that explain risk differences across the equipment population. Section 2.4 synthesises the review and states the gap that this dissertation addresses.

2.1 From reactive to predictive and prescriptive maintenance

Chapter 1 introduced the taxonomy of maintenance policies. This section develops the reasoning behind that taxonomy: what each strategy costs, where it is justified, and why industry has been moving from reaction towards anticipation and, more recently, towards prescription. The purpose is to establish the strategic ground on which the two analytical layers of this work stand.

2.1.1 Corrective and predetermined maintenance and their limits

Under a corrective strategy, equipment runs until a fault occurs and is then repaired or replaced to restore its function (Molęda et al., 2023). The appeal is the minimal direct cost: no inspection or planning effort is spent in advance, and no functioning part is renewed before failure demands it. The literature considers it a defensible policy for non-critical, redundant, or easily repairable equipment, where the consequences of failure are modest (Molęda et al., 2023). Its costs are indirect. An unplanned repair takes longer and costs more than a planned one, and a primary failure can damage or destroy associated equipment (Molęda et al., 2023). Because root causes are rarely addressed when repairs are made under time pressure, the same failures also tend to recur, depressing the time between failures (Erbiyik, 2022). In continuous production a further effect is documented: components may operate permanently in a degraded state, their parameters

(efficiency, vibration, temperature) deteriorating so slowly that the decay goes unnoticed precisely because production does not stop (Molęda et al., 2023).

Predetermined maintenance counters these risks with inspections and replacements scheduled at fixed intervals of time or usage, typically derived from manufacturer recommendations or from aggregate statistics such as the mean time between failures (Molęda et al., 2023; Erbiyik, 2022). When the plan matches the failure behaviour, the benefits are substantial: minor defects are resolved before they escalate, asset life is extended, spare parts can be planned, and workplace safety improves (Erbiyik, 2022). The policy nevertheless has two structural weaknesses. First, it offers no protection against failure modes that the plan does not cover or that develop between interventions (Molęda et al., 2023). Second, replacing parts too often is itself harmful: a new part carries a higher probability of being defective or incorrectly assembled than a part already in stable service, so excessive renewal re-exposes the equipment to early-life failures (Molęda et al., 2023). This behaviour is summarised by the bathtub curve, developed in Section 2.3 (Molęda et al., 2023; Guimarães and Leitão, 2025). A time-based replacement is therefore statistically justified only for equipment whose failure risk genuinely increases with age, a condition that must be verified rather than assumed (Guimarães and Leitão, 2025).

The critical decision in a predetermined policy is thus the calibration of the interval. Intervals shorter than the failure behaviour warrants waste resources and may even introduce early-life failures; intervals longer than warranted leave failures unprevented. Moreover, the appropriate policy is not a plant-wide choice but an equipment-level one. Reliability-centred maintenance (RCM) formalises this principle: each failure mode is assigned the task justified by its consequences and its failure pattern, in an optimised mix that ranges from condition-based intervention to deliberately accepting run-to-failure on non-critical items (Molęda et al., 2023; Erbiyik, 2022). For a plant with thousands of maintainable locations, this case-by-case calibration is precisely the task that, as Chapter 1 described, remains out of reach when attempted manually, and it is the task that the reliability layer of this work sets out to automate.

2.1.2 Predictive maintenance and the conditions for its success

Predictive maintenance carries the logic one step further: servicing is performed when it is required, usually shortly before a failure is expected, on the basis of a forecast built from current and historical indicators of equipment condition (Molęda et al., 2023). Because interventions are tied to predicted need rather than to the calendar, the approach reduces unplanned downtime, by anticipating failures, and planned downtime, by eliminating interventions that the equipment state does not justify (Molęda et al., 2023). The reported gains are considerable: compared with predetermined maintenance, studies compiled by Erbiyik (2022) indicate reductions of 25–30% in maintenance costs, 70–75% in breakdowns, and 35–45% in downtime.

The conventional route to these gains is condition monitoring through the dedicated techniques introduced in Chapter 1: vibration analysis, infrared thermography, oil and particle analysis, acoustic emission, corrosion monitoring, and performance monitoring. Each, however, carries

a cost per monitored asset: vibration monitoring requires the installation of additional measurement equipment, and oil analysis involves costly instrumentation together with non-trivial sampling and interpretation (Molęda et al., 2023). The investment is most defensible for small fleets of high-value assets with well-understood failure modes.

Practice confirms that these conditions are restrictive. A multiple-case study of thirteen industrial implementations found that few companies succeed in deploying predictive maintenance effectively (Tiddens et al., 2022). Bradbury et al. (2018) reach a converging conclusion: advanced, model-based prediction is economically viable only for a subset of cases, namely machines with well-documented failure modes and high downtime impact, or large fleets of identical assets across which development costs can be spread. Where a machine can fail in hundreds of distinct and individually rare ways, building enough predictive models of sufficient quality is impractical, and simpler instruments (monitoring against thresholds, or the consistent and rigorous application of data-driven reliability analysis) deliver more value (Bradbury et al., 2018).

The same authors emphasise a second, frequently overlooked point: most industrial organisations already record maintenance and reliability data in their enterprise systems, yet the value of this asset is undermined by poor housekeeping, with inconsistent asset descriptions, free-text records, and information locked away in spreadsheets. Digitisation by default therefore delivers little of its potential (Bradbury et al., 2018). The computerised maintenance management system (CMMS) is the natural repository of this history, holding the records of repairs, diagnostics, and equipment interventions accumulated over years of operation (Molęda et al., 2023). The failure early warning system, one of the recognised implementation routes of predictive maintenance, builds exactly on this resource (Erbiyik, 2022). It supports the existing computerised maintenance programme, detects potential failures before they occur, helps establish priorities among maintenance needs, and progressively builds a maintenance database of value for future analyses (Erbiyik, 2022). For a continuous-process plant without dedicated instrumentation per equipment location (the setting of this work) this is the viable route that the literature points to, and following it requires the two analytical instruments reviewed in the remainder of this chapter.

2.1.3 From prediction to prescription

The strategies above can also be read as steps on a ladder of analytical maturity. Descriptive analysis answers what happened, through reports, key performance indicators (KPIs), and aggregate metrics such as the mean time to failure. Diagnostic analysis answers why it happened, covering the detection of a fault, the isolation of its location, and the identification of its nature. Predictive analysis answers what will happen, anticipating failures and estimating the remaining useful life (RUL). Prescriptive analysis, the most mature level, answers how to do better: it simulates or optimises the available decision alternatives and recommends the action that best serves a target function, such as the maintenance interval that minimises cost (Molęda et al., 2023; Guimarães and Leitão, 2025). Each step up the ladder increases the value of the insight delivered, and also the demands placed on data, infrastructure, and analytical capability (Molęda et al., 2023).

Most industrial organisations remain at the bottom of this ladder, monitoring and reporting on asset behaviour without systematic diagnosis or prediction (Guimarães and Leitão, 2025; Bradbury et al., 2018). The plant studied in this dissertation is, as Chapter 1 described, a case in point: years of structured maintenance records serve retrospective reporting only. The framework developed in this work is designed to climb the ladder without new instrumentation. Its predictive layer delivers calibrated short-horizon failure alerts; its reliability layer is diagnostic and prescriptive, assessing the adequacy of the current preventive policy and recommending intervention intervals where the data support them. The next two sections review the foundations of each layer in turn.

2.2 Machine learning for failure prediction

Among the predictive approaches introduced in the previous section, data-driven methods learn the patterns that precede failure directly from historical data, without an explicit physical model of the degradation mechanism (Molęda et al., 2023). The literature formulates the learning task in two main ways. The first estimates the RUL of a unit as a regression problem, the closing step of a prognostic programme that runs from data acquisition through health indicator construction and health stage division to prediction (Lei et al., 2018). The second casts prediction as binary classification: given what is known at a reference date, will the unit fail within a defined future window (Susto et al., 2015). The window length is a design choice with economic content, since it governs the balance between unexpected breakdowns and the remaining lifetime sacrificed by acting early (Susto et al., 2015).

Both formulations rest on the same physical premise, formalised in RCM as the P-F curve, shown in Figure 2.1.

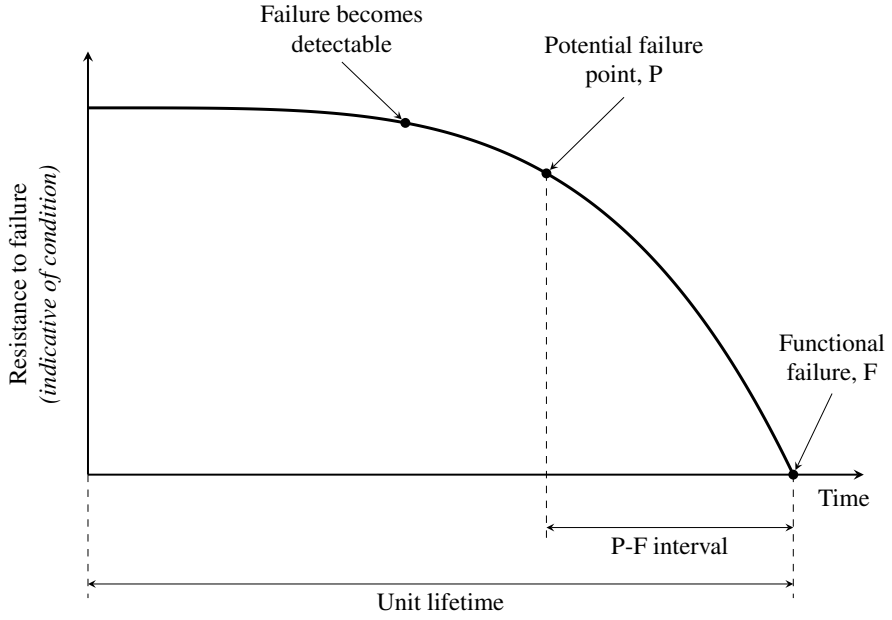

Figure 2.1: The P-F curve. Adapted from Ochella et al. (2021).

From the moment a developing failure becomes detectable, a finite interval separates the potential failure point P from the functional failure F, and a warning is only useful if detection, decision, and intervention all fit inside it (Ochella et al., 2021). In practice this interval is difficult to determine, and most equipment degrades non-linearly, which makes direct extrapolation unreliable (Ochella et al., 2021). When no condition-monitoring signal exists, the curve cannot be observed at all; a classifier with a fixed alert window then operationalises the same idea on event data, the window standing in for the P-F interval as the lead time offered to the planner.

2.2.1 Data regimes in failure prediction

What separates the published applications is less the algorithm than the data available to learn from. Three regimes can be distinguished, and Table 2.1 places representative studies along this axis.

Table 2.1: Data-driven failure-prediction studies by data regime.

Study	Context	Data	Model(s)	Key result
Ochella et al. (2021)	Aero-engine fleet (simulated)	Run-to-failure multi-sensor channels	Linear health index; k-means	Health-state grouping to prioritise life extension
Susto et al. (2015)	Semiconductor manufacturing	Process and logistic event variables	One classifier per horizon	Cost-based selection of actions; handles censored data
Allah Bukhsh et al. (2019)	Railway switches	Maintenance work-order records	Decision tree, random forest, gradient boosting	Boosted trees most accurate for maintenance need
Dai and Liu (2023)	Rail-transit signalling	Trouble-call event history	Supervised classifiers on window features	One third of failures flagged one month ahead at 10% of locations
AlGanem and Abdallah (2022)	Heavy-vehicle fleet	38,000 CMMS work orders	Tree ensembles	Boosted trees best for failure and issue type
Snider and McBean (2020)	Water distribution mains	Historical break records	Gradient boosting; survival models	Accuracy governed by depth of failure history

At the richest end sit condition-monitoring data: continuous multivariate sensor measurements, ideally with complete run-to-failure trajectories. Ochella et al. (2021) illustrate the regime on a simulated fleet of one hundred turbofan engines, fusing the informative sensor channels into a single health indicator and clustering units by health state to prioritise life-extension actions; the review by Lei et al. (2018) documents the breadth of this literature. Its reach into industrial practice, however, is bounded by its own premise: instrumenting thousands of heterogeneous assets is expensive and often impractical (Allah Bukhsh et al., 2019).

A second regime uses no sensors at all. It draws on the work-order history already held in the maintenance management system, encoding the recency, frequency, and context of past orders as features for supervised classification. Allah Bukhsh et al. (2019) predict the maintenance need

of Dutch railway switches from such records, motivated precisely by the cost argument above, and find gradient-boosted trees the most accurate of the tree-based models compared. Dai and Liu (2023) build window-based features from the trouble-call history of 18,623 rail-transit signal units and, under severe class imbalance, flag roughly one third of failures one month in advance while concentrating attention on ten per cent of locations. AlGanem and Abdallah (2022) reach a similar conclusion on 38,000 corrective work orders from a heavy-vehicle fleet, with gradient-boosted trees again ahead. In manufacturing, Susto et al. (2015) learn the footprint of degradation from process and logistic variables recorded during production, with no direct measurement of degradation itself, and train one classifier per prediction horizon so that maintenance actions can be selected on operating cost.

The third regime is defined by scarcity: histories exist, but each unit contributes few recorded failures, so the learning problem is constrained by sample size before any modelling choice is made. Comparing gradient boosting with survival models on municipal water-main break records, Snider and McBean (2020) show that predictive accuracy is governed largely by the depth of the available failure history, a constraint no algorithm removes. The present work sits at the meeting point of the last two regimes: event data from the maintenance system, with short per-unit histories.

2.2.2 Learning algorithms for tabular maintenance data

Whatever the regime, the observation that reaches the model is tabular: a fixed-length vector of counts, ratios, recency measures, and categorical attributes describing one unit at one date. Tree-based ensembles are the established model family for this kind of input, and benchmark evidence supports the preference: across forty-five medium-sized tabular datasets, tree-based models remained state of the art against deep neural architectures (Grinsztajn et al., 2022). Random forests aggregate de-correlated decision trees grown on bootstrap samples (Breiman, 2001), while gradient boosting builds the ensemble sequentially, each new tree fitted to the residual errors of the trees before it (Chen and Guestrin, 2016). CatBoost is a gradient-boosting implementation whose ordered target statistics encode categorical variables natively while limiting the target leakage that arises when encodings are computed on the same data used for training (Prokhorenkova et al., 2018). The pattern in the event-data studies above, where gradient-boosted trees repeatedly outperformed the alternatives (Allah Bukhsh et al., 2019; AlGanem and Abdallah, 2022), is consistent with this benchmark picture. Which member of the family suits a given dataset remains an empirical question, answered for the present case under a common evaluation protocol in Chapter 4.

2.2.3 Evaluation, calibration, and interpretation

Failure prediction operates under class imbalance, since at any reference date only a small fraction of units fail within the window (Dai and Liu, 2023). Imbalance changes how performance should be read. The receiver operating characteristic (ROC) curve can give a deceptively favourable impression, because the large negative class sustains specificity almost regardless of how many

false alarms accompany each true one; precision-recall analysis exposes exactly that cost and is the more informative lens on imbalanced problems (Saito and Rehmsmeier, 2015).

How the data are split to estimate performance matters as much as the metric itself. The standard procedure, K-fold cross-validation, shuffles observations at random across folds, which is sound only when they can be treated as interchangeable. Maintenance records cannot: they carry serial correlation, and the equipment population is not stationary over time. Random folds would then train the model on observations later than those it is tested on, letting it learn from the future of the very cases it must score and inflating the measured performance. The validity of cross-validation under serial dependence is in fact conditional, established for autoregressive models only when their errors are uncorrelated; where that cannot be assumed, forecasting practice favours out-of-sample evaluation (Bergmeir et al., 2018). The conservative alternative trains on records up to a cut-off date and tests on the period that follows, reproducing the operational task of predicting failures that have not yet happened.

When predictions feed planning decisions, ranking units is not enough: the predicted probabilities must themselves be reliable. Boosted ensembles, as maximum-margin methods, push predicted scores away from zero and one, distorting them as probability estimates; post-hoc calibration with Platt scaling or isotonic regression corrects the distortion, after which boosted trees are among the best probability estimators available (Niculescu-Mizil and Caruana, 2005).

For interpretation, SHapley Additive exPlanations (SHAP) values decompose each prediction into additive feature contributions with theoretical guarantees (Lundberg and Lee, 2017) and have become a widely adopted attribution tool for tree ensembles. The prognostics literature nevertheless documents that post-hoc attribution can produce unstable, sometimes contradictory importance rankings (Salinas-Camus et al., 2025); Section 2.3.2 returns to this limitation and to the stable inferential complement that hazard regression offers.

Taken together, this body of work delivers a short-horizon risk ranking of individual units, learned from whatever data regime is available, together with the apparatus to evaluate, calibrate, and explain it. What it does not deliver is an account of the failure process itself: how risk evolves with time and accumulated history, and which structural factors drive it across a population. Those questions belong to the reliability analysis of repairable systems, reviewed next.

2.3 Reliability analysis of repairable systems

Where the previous section asked which unit is likely to fail next, reliability analysis characterises the failure process itself. Its vocabulary rests on three equivalent descriptions of the same random variable: the lifetime distribution, which assigns probability to the time between failures; the survival function, the probability of not failing up to a given time; and the hazard function, the conditional probability of failure per unit of time given survival until that moment (Guimarães and Leitão, 2025).

A second distinction concerns the object of analysis. A component is an item removed from the system upon failure, and its lifetimes are naturally described by a lifetime distribution; a repairable

system is an item restored to service by replacing the components that caused the failure, so what is observed is a sequence of failures in time rather than a single lifetime (Guimarães and Leitão, 2025). The two levels should not be conflated: the failure rate of a system reflects the height of its components' hazard functions rather than their shape, and a system can show a constant failure rate even while its components individually wear out (Guimarães and Leitão, 2025).

Over the operating life of a population of items, the hazard rate traces the bathtub profile of Figure 2.2: a wear-in region of decreasing hazard dominated by early defects, a long region of approximately constant hazard in which failures strike at random, and a wear-out region of increasing hazard as ageing mechanisms take over (Molęda et al., 2023; Guimarães and Leitão, 2025). The curve describes a population rather than the path of any single item, and each of its regions corresponds to a value of the Weibull shape parameter introduced below.

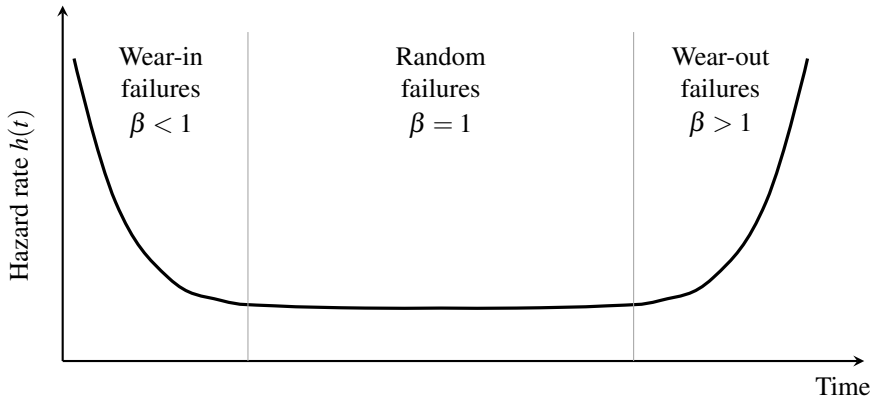

Figure 2.2: The bathtub curve. Adapted from Molęda et al. (2023) and Guimarães and Leitão (2025).

For a repairable system, the first analytical question is whether the failure process is stable. Treating the inter-failure times as a sample from a single distribution presumes that they are independent and identically distributed (IID), and this presumption fails when the process carries a trend: a system whose failures arrive ever faster is deteriorating, one whose failures thin out is improving, and in either case the order of the observations matters. Testing for trend before fitting any lifetime model is therefore the recommended practice for repairable-system data, since it provides the formal basis for choosing between a renewal representation and a non-stationary one (Louit et al., 2009). The standard instrument is the Laplace test (Guimarães and Leitão, 2025). Its null hypothesis H_0 is that the failure occurrences are IID, which is to say the process carries no trend; the alternative H_1 is that the occurrences are not IID, that is, that a trend is present. Under H_0 the test statistic is approximately standard normal once at least four failures have been observed, and for an observation window truncated at time t_0 it is given by Equation 2.1.

$$\mathrm{TS} = \sqrt{12N}\left[\frac{\sum_{i=1}^{N} t_i}{N t_0} - 0.5\right] \tag{2.1}$$

Here, N is the number of failures and t_i the time of the i-th failure measured from the start of service; the statistic is compared against the standard-normal critical values at a chosen significance level α (equivalently, its p-value against α). Two variants of the statistic exist, for observation windows truncated at a fixed time and truncated at a failure; the time-truncated form above applies throughout this work, since observation always ends at the data extraction date. A significantly positive value indicates failures concentrating late in the window, an increasing trend that marks a deteriorating process; a significantly negative value indicates a decreasing trend, the signature of reliability growth. When H_0 is not rejected, the process shows no trend, so the inter-failure times are treated as IID and passed as a renewal sequence to a lifetime model (Guimarães and Leitão, 2025).

When a renewal representation is admissible, the Weibull distribution is the workhorse lifetime model of reliability engineering, owing to the variety of shapes it can take: a single parameter moves its hazard between the decreasing, constant, and increasing forms (Guimarães and Leitão, 2025). With shape parameter $\beta > 0$ and scale (characteristic life) $\eta > 0$, and with the location parameter of the general three-parameter form set throughout this work to its usual value $\gamma = 0$ (Guimarães and Leitão, 2025), the Weibull probability density function is given by Equation 2.2.

$$f(t) = \frac{\beta}{\eta^{\beta}} t^{\beta-1} \exp\left[-\left(\frac{t}{\eta}\right)^{\beta}\right], \qquad t > 0 \tag{2.2}$$

The distribution function follows by integration, as in Equation 2.3.

$$F(t) = \int_0^t f(u)\,\mathrm{d}u = 1 - \exp\left[-\left(\frac{t}{\eta}\right)^{\beta}\right] \tag{2.3}$$

The survival function is its complement, the probability of operating beyond time t, and is given by Equation 2.4.

$$S(t) = 1 - F(t) = \exp\left[-\left(\frac{t}{\eta}\right)^{\beta}\right] \tag{2.4}$$

The hazard function is the ratio of the density to the survival function, and takes the form of Equation 2.5.

$$h(t) = \frac{f(t)}{S(t)} = \frac{\beta}{\eta^{\beta}} t^{\beta-1} \tag{2.5}$$

The shape parameter governs its form. A value $\beta < 1$ produces a decreasing hazard, $\beta = 1$ a constant hazard, and $\beta > 1$ an increasing hazard, reproducing in turn the three regions of the bathtub curve, while the scale η is the characteristic life, by which roughly 63 per cent of a population will have failed. The mean of the fitted distribution, used as the component mean time between failures (MTBF), is its first moment and is given by Equation 2.6.

$$\mathrm{MTBF} = \eta\,\Gamma\left(1 + \frac{1}{\beta}\right) \tag{2.6}$$

Here Γ denotes the gamma function. Industrial failure records are rarely complete: many units are still in service at the time of analysis, so their running times enter the data as right-censored observations, lower bounds on lifetimes not yet realised (Guimarães and Leitão, 2025). Maximum likelihood estimation (MLE) accommodates censoring directly and is the standard fitting procedure (Guimarães and Leitão, 2025). For r failures observed at times $t_1, \ldots, t_r$ and k censoring times $T_1, \ldots, T_k$, with C_j units censored at each T_j, the shape parameter is the solution of the likelihood equation of Equation 2.7.

$$\frac{\sum_{i=1}^{r} t_i^{\beta} \ln t_i + \sum_{j=1}^{k} C_j T_j^{\beta} \ln T_j}{\sum_{i=1}^{r} t_i^{\beta} + \sum_{j=1}^{k} C_j T_j^{\beta}} - \frac{1}{\beta} = \frac{1}{r} \sum_{i=1}^{r} \ln t_i \tag{2.7}$$

This equation has no closed-form solution and is solved numerically for β, from which the scale parameter follows in closed form, as in Equation 2.8.

$$\eta = \left(\frac{\sum_{i=1}^{r} t_i^{\beta} + \sum_{j=1}^{k} C_j T_j^{\beta}}{r} \right)^{1/\beta} \tag{2.8}$$

For complete data the censoring terms vanish and both reduce to the standard estimators (Guimarães and Leitão, 2025).

In small samples the point estimate of the shape parameter is a weak basis for declaring an increasing hazard, since values above one can arise from sampling variability alone. The usual safeguard is to require the lower bound of a one-sided confidence interval (CI) to exceed unity instead. This lower bound divides the estimate by a sample-size-dependent critical value F_β tabulated by Bain (1972), as in Equation 2.9.

$$\beta_L = \frac{\hat{\beta}}{F_\beta} \tag{2.9}$$

Wear-out is declared only when $\beta_L > 1$. The critical value F_β is tabulated as a function of sample size for the usual confidence levels, and the level adopted governs the balance between declaring wear-out on insufficient evidence and rejecting a genuine one (Bain, 1972). The chart of confidence limits from which F_β is read is reproduced in Figure D.1 of Appendix D; the level adopted in this work is set, with its justification, in Chapter 4.

A system can fail through several distinct failure modes, each with its own degradation path. In the classical competing-risks setting of a non-repairable item, the observed lifetime is the shortest of the latent times to failure, and the standard likelihood treats every cause other than the one realised as right-censored at the observed failure time (Crowder, 2001). A repairable system departs from this construction in one essential respect: repairing the component that failed leaves the ages of the other components unchanged, so a failure of one mode neither ends nor censors the exposure of the others, and each mode is analysed on its own timeline, with a separate lifetime distribution fitted from its own failures, its own preventive renewals, and a censoring at the end of observation (Guimarães and Leitão, 2025). In the present application the repairable system is the

equipment, and each of its components is treated as a distinct failure mode. This component-level decomposition is what lets the analysis reach below the equipment to the part that actually wears out.

The shape of the hazard carries the maintenance decision. Preventive replacement of a component is only justified when its hazard is increasing; under a constant or decreasing hazard, renewing a functioning part buys nothing, because the new part is no less likely to fail than the one it replaces (Guimarães and Leitão, 2025). Where the hazard does increase and a planned replacement costs less than a failure, an optimal replacement age exists: the age that minimises the expected cost per unit of time, balancing the cost of renewing parts early against the cost of the failures thereby avoided (Guimarães and Leitão, 2025). Classical maintenance-optimisation theory turns this principle into operational interval models, reviewed next.

2.3.1 Cost-optimal preventive intervals

The classical interval models share one construction: a preventive action performed on a fixed schedule defines a renewal cycle, and the interval is chosen to minimise the long-run cost per unit of time over that cycle. They differ in the kind of preventive action considered and in what happens between interventions, and the three models reviewed here cover the three kinds of preventive action that recur in industrial maintenance plans: the replacement of individual components, periodic group actions such as lubrication and cleaning, and inspection.

Under age-based replacement, the first policy of Barlow and Hunter (1960), a component is replaced preventively when it reaches age t_p or correctively when it fails, whichever comes first. A cycle therefore ends either with a planned replacement, at cost C_p after exactly t_p time units, or with a failure, at cost C_f after the mean life of the units that fail before t_p. This mean truncated life is given by Equation 2.10.

$$M(t_p) = \frac{1}{1 - S(t_p)} \int_0^{t_p} t\, f(t)\, \mathrm{d}t \tag{2.10}$$

Weighting the two endings by their probabilities gives the expected cost per unit of time, as in Equation 2.11.

$$C(t_p) = \frac{C_p S(t_p) + C_f \left(1 - S(t_p)\right)}{t_p S(t_p) + M(t_p) \left(1 - S(t_p)\right)} \tag{2.11}$$

Here, S and f are the survival and density functions of the component lifetime (Equations 2.4 and 2.2), in practice supplied by the fitted Weibull distribution (Guimarães and Leitão, 2025; Jardine and Tsang, 2013). Under the two conditions established above, an increasing hazard and $C_p < C_f$, Equation 2.11 has an interior minimum that defines the optimal replacement age.

Group actions such as lubrication and cleaning renew a shared operating condition rather than any single component. The matching policy is the second of Barlow and Hunter (1960): a periodic intervention every T time units restores that condition, and failures arriving between interventions are corrected by minimal repair, a restoration that leaves the unit as bad as old without renewing its age. The expected number of failures per cycle is then the cumulative hazard, equal to $(T/\eta)^{\beta}$

for the Weibull, as in Equation 2.12.

$$H(T) = \int_0^T h(t)\,\mathrm{d}t = \left(\frac{T}{\eta}\right)^{\beta} \tag{2.12}$$

The cost per unit of time is then given by Equation 2.13.

$$C(T) = \frac{C_g + C_f H(T)}{T} \tag{2.13}$$

Here, C_g is the cost of one group intervention, and C_f is now the cost of one minimal repair, the failure-response action of this policy, rather than the corrective replacement of the age-based model. A finite optimum again requires an increasing hazard: under a constant hazard, the periodic action buys nothing, and the cost only falls as T grows.

Inspection plans rest on a different premise, formalised in the delay-time concept of Christer and Waller (1984): a failure is preceded by an identifiable defect, and the interval between the moment the defect arises and the failure it would cause, the delay time, is the window in which an inspection can find and remove it. In the standard formulation, defects arise as a Poisson process with rate λ_d, the delay time is exponential with rate λ_h, and inspections are assumed perfect, finding and repairing every defect present at the moment they are carried out (Christer and Waller, 1984). For inspections every T time units, the fraction of defects that become failures before being caught is given by Equation 2.14.

$$b(T) = 1 - \frac{1 - e^{-\lambda_h T}}{\lambda_h T} \tag{2.14}$$

The expected cost per unit of time is then given by Equation 2.15.

$$C(T) = \frac{C_I + C_f \lambda_d T\, b(T)}{T + d_I} \tag{2.15}$$

Here, C_I is the cost of one inspection and d_I the downtime it imposes (Christer and Waller, 1984; Wang, 2012). The two parameters are estimated by MLE from the counts of failures and of defects detected in each inspection interval (Wang, 2012); the model imposes no minimum number of failures, but the delay rate is only identifiable if at least some defects are in fact detected at inspections.

How the cost parameters are obtained, how each preventive plan in the maintenance system is matched to its model, and the statistical safeguards imposed before any interval is recommended are matters of application, addressed in Chapter 4.

The models reviewed so far share one structural assumption: the equipment population is homogeneous, with every unit drawing its failure times from the same distribution and differences between units appearing only as random variation. The subsection that follows reviews the class of models that removes this assumption.

2.3.2 Covariate-based hazard models for repairable systems

A lifetime model such as the Weibull describes how the risk of failure evolves with operating time; it offers no mechanism to explain why some units fail more often than others. In an industrial fleet, units differ in criticality, equipment type, installation area, and maintenance treatment, and these differences are observable in the maintenance records. Regression models for survival data were developed precisely to quantify the effect of such observable factors on the hazard of failure (Cox, 1972), and a dedicated review documents their adoption in reliability engineering (Kumar and Klefsjö, 1994).

The proportional hazards model (PHM) of Cox (1972) expresses the hazard of a unit with covariate vector $\mathbf{x}$ as in Equation 2.16.

$$h(t \mid \mathbf{x}) = h_0(t) \exp(\mathbf{x}^\top \boldsymbol{\theta}) \tag{2.16}$$

Here, $h_0(t)$ is an unspecified baseline hazard common to all units and $\boldsymbol{\theta}$ is a vector of regression coefficients. The model is semi-parametric: the baseline requires no distributional assumption, and the coefficients are estimated by partial likelihood without estimating $h_0(t)$ itself (Cox, 1972). The hazard ratio (HR) of covariate j is the exponentiated coefficient, given by Equation 2.17.

$$\mathrm{HR}_j = \exp(\theta_j) \tag{2.17}$$

A value of 1.5 then means that a one-unit increase in the covariate multiplies the instantaneous risk of failure by 1.5, holding the other covariates fixed. This direct interpretation, accompanied by confidence intervals and significance tests, is the main reason for the model's popularity in applied reliability work (Kumar and Klefsjö, 1994).

Applications in maintenance settings have a long history and, importantly for the present work, many of them rely on failure and work-order records alone. Jardine et al. (1987) fitted a Weibull PHM to aircraft and marine engine failure data; Love and Guo (1991) applied the model to plant maintenance records to support repair and replacement decisions; and Kumar et al. (1992) identified the factors affecting the reliability of power transmission cables of mine loaders. More recently, Ozturk et al. (2018) used survival analysis on the maintenance histories of 109 wind turbines to quantify how scheduled maintenance and drive-train configuration alter survival, an exercise in population-level factor identification rather than unit-level prediction. These studies establish that hazard regression extracts engineering insight from exactly the kind of data a computerised maintenance management system accumulates.

Industrial equipment is repairable, so a single unit contributes several failures over the observation window, and the classical model in Equation 2.16, which assumes one event per unit, must be extended. Andersen and Gill (1982) reformulated the model as a counting process, in which each inter-failure interval of a unit enters the risk set over its own calendar segment $(t_{\text{start}}, t_{\text{stop}}]$.

The intensity of the recurrent-event process is then given by Equation 2.18.

$$\lambda(t \mid \mathbf{x}) = Y(t)\, h_0(t) \exp(\mathbf{x}^\top \theta) \tag{2.18}$$

Here, $Y(t)$ indicates whether the unit is at risk at time t. The Andersen-Gill formulation treats all events as interchangeable and is the most data-efficient extension, but it assumes that the covariates fully capture any dependence between successive events of the same unit. Prentice et al. (1981) proposed a conditional alternative in which a unit is only at risk for its k-th event after experiencing the $(k-1)$-th, and the baseline hazard is stratified by event order. This gap-time model is given by Equation 2.19.

$$h_k(g \mid \mathbf{x}) = h_{0k}(g) \exp(\mathbf{x}^\top \theta) \tag{2.19}$$

Here, g denotes the gap time since the previous event and h_{0k} is the baseline hazard specific to event order k. This gap-time formulation absorbs the dependence of the failure process on its own past into the stratified baselines, so the covariate effects are estimated conditionally on the accumulated event history (Prentice et al., 1981; Amorim and Cai, 2015). Whichever formulation is chosen, the repeated intervals of one unit are not independent, and the recommended practice is to compute robust sandwich variance estimates clustered by unit, an approach covered in detail by Therneau and Grambsch (2000); shared frailty models, which add a multiplicative random effect per unit, are the main alternative (Amorim and Cai, 2015).

The central assumption of Equations 2.16 to 2.19 is proportionality: the hazard ratio of each covariate is constant over time. Schoenfeld (1982) introduced residuals that isolate the contribution of each covariate at each event time, and Grambsch and Therneau (1994) showed that a regression of the scaled Schoenfeld residuals on time yields a formal test of the assumption; a non-zero slope indicates a time-varying effect. When violations are detected, the standard remedies are stratification on the offending covariate or a parametric accelerated failure time specification, and with large event counts the test detects departures too small to alter the substantive reading, in which case the estimates are interpreted as time-averaged effects (Therneau and Grambsch, 2000).

Sample size in hazard regression is governed by the number of events rather than the number of units. Simulation work by Peduzzi et al. (1996) established the guideline of at least ten events per estimated parameter, below which coefficient estimates become unstable; Vittinghoff and McCulloch (2007) later showed that the rule can be relaxed in defined circumstances, and Riley et al. (2019) replaced it with criteria based on expected shrinkage for prediction settings. For recurrent events the relevant count is the total number of observed inter-failure intervals, which in a fleet of repairable units can exceed the unit count several times over, a property that makes covariate-based hazard analysis feasible precisely where per-unit histories are short.

One caution governs the interpretation of any coefficient estimated from observational maintenance data. Maintenance effort is not assigned at random: preventive attention concentrates on the equipment already known to be problematic, so a covariate recording maintenance intensity can show a positive association with failure risk even when the maintenance itself is beneficial, the pattern known in epidemiology as confounding by indication (Walker, 1996). Coefficients of

this kind are read as associations, never as causal effects of maintenance.

Hazard regression also occupies a distinct position with respect to the machine learning models of Section 2.2. The post-hoc attributions used to explain learned models can be unstable, as noted in Section 2.2.3; hazard ratios estimated from Equation 2.19, by contrast, are model parameters carrying confidence intervals and significance tests, and therefore provide a stable inferential complement to predictive models: the latter rank individual units by short-term risk, the former explain which structural factors raise risk across the population. The two roles are complementary rather than competing, a distinction this work exploits.

Taken together, the reliability and hazard-regression literature shows that maintenance records alone can support trend assessment, lifetime modelling, and population-level factor identification; what they cannot do is single out, at any given date, the individual units most likely to fail next, the task the machine learning models of Section 2.2 address. The synthesis that follows positions the two strands against the gap this work addresses.

2.4 Synthesis and research gap

The approaches reviewed in this chapter can be read along a single axis: the trade-off between the depth of information available per unit and the coverage of the equipment population. At one end lies depth. Continuous sensor monitoring and complete run-to-failure histories support rich estimates of RUL, as in the condition-monitoring regime of Section 2.2.1, but they require instrumentation, calibration, and degradation models tuned to each component. That investment is justified for small fleets of high-value critical assets, where the consequences of an unplanned failure are severe; it does not scale to a plant of thousands of heterogeneous locations. At the other end lies coverage. The work-order history of the maintenance management system spans the entire population, because it is generated as a by-product of operation, but per unit it is sparse, categorical, and free of any direct measurement of degradation.

The two analytical strands reviewed above sit on opposite sides of this axis, and each answers only one half of the problem faced by a plant that has coverage but not depth. Reliability analysis and hazard regression (Section 2.3) operate on the maintenance record itself and so achieve full coverage, but they require enough events per unit to be informative and speak to the failure process and the structural drivers of risk, not to which unit to act on this week. Supervised learning on event data (Section 2.2) supplies exactly that near-term, unit-level ranking and tolerates sparse histories through aggregated features, yet it is silent on how failures arise and on whether the preventive policy in force is adequate.

The building blocks therefore already exist, but only separately: tree-based classifiers on work-order data (Allah Bukhsh et al., 2019; Dai and Liu, 2023; AlGanem and Abdallah, 2022), the reliability theory of repairable systems, and covariate-based hazard models. What the reviewed literature does not demonstrate, for this data regime, is their articulation over a single body of maintenance records in the service of day-to-day planning. Closing this gap requires a framework that draws exclusively on CMMS event records and, from them, delivers calibrated short-horizon

alerts for the individual unit, characterises the failure process, assesses the adequacy of the preventive policy and recommends cost-optimal intervals wherever the data allow, identifies the structural risk factors of the population, and consolidates these outputs into a single reproducible decision-support tool.

This is the contribution of the present work. Existing studies typically address these tasks in isolation, whereas this dissertation integrates predictive and reliability analytics under sparse CMMS event data into a single operational framework for a continuous-process plant, turning a maintenance record kept for reporting into day-to-day decision support.

Chapter 3

Problem Description

The general tension between reactive and preventive maintenance, introduced in Chapter 1, takes a concrete form at the Mangualde plant. Continuous MDF production generates a dense and uninterrupted stream of operational events, and more than four years of corrective and preventive activity have been recorded across thousands of equipment locations. Despite this volume, maintenance decisions remain largely reactive and experience-driven, and the historical record is seldom used in a structured way to anticipate failures or to test whether preventive plans are correctly calibrated.

Two questions frame this chapter. The first is whether the records already held by the plant are detailed and reliable enough to support a data-driven maintenance policy. The second is what those same records cannot do, because the limits of the data shape the methodology as strongly as its strengths. The chapter characterises the data environment in which any such policy must operate: the production process and its equipment (Section 3.1), the two source systems and the information each one holds (Section 3.2), the way equipment is coded and located (Section 3.3), and the structural gap between the two systems (Section 3.4). A SWOT analysis then consolidates the diagnosis and motivates the framework developed in Chapter 4 (Section 3.5).

3.1 The Mangualde plant and the MDF production process

The Mangualde plant produces medium-density fibreboard (MDF), an engineered panel formed from refined wood fibre bonded with resin under heat and pressure. Production runs on continuous lines, so the relevant unit for availability is the line as a whole: a stoppage at any single stage halts everything downstream once the intermediate buffers are exhausted. This continuity is what makes unplanned downtime expensive, and it is the operational reason the plant invests in maintenance data.

The main stages of the process are summarised in Figure 3.1. Roundwood and wood chips are first debarked, chipped, and screened to remove oversized material, then washed to remove sand and grit. The cleaned chips are steamed and refined into wood fibre in the defibrator, where rotating refiner discs perform the size reduction. Resin is injected into the fibre along the blowline, the glued fibre is dried, and the dried fibre is distributed into a uniform mat. The mat is finally

consolidated under heat and pressure into a continuous board, which is cooled, sanded, and cut to size.

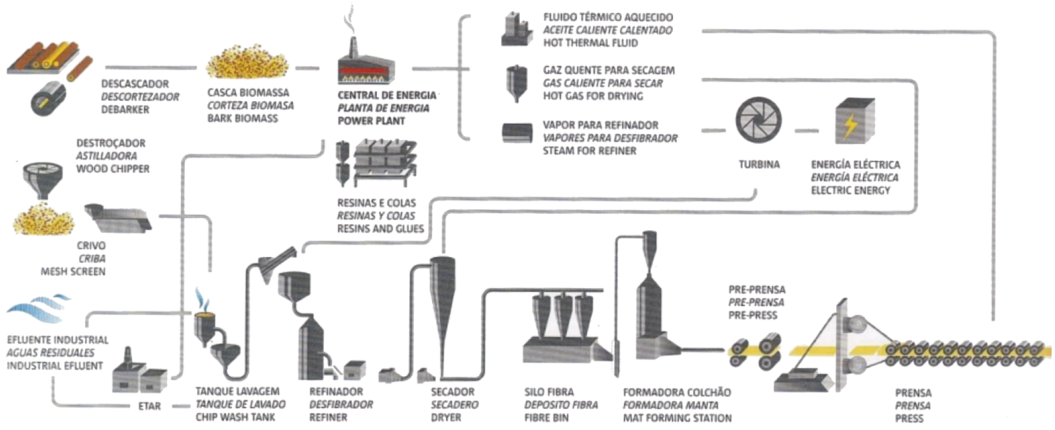

Figure 3.1: Simplified flow of the MDF production process.

Each of these stages is a dense population of rotating and hydraulic equipment: motors, gearboxes, screw conveyors, pumps, hydraulic units, and the press itself. Mechanical wear in this equipment is the dominant source of corrective work, whereas a few interventions are inherently planned, such as the periodic replacement of the refiner discs, whose stoppage is also used to carry out other maintenance. A continuous process combined with a large, wear-prone asset base is precisely what makes the plant's maintenance records both voluminous and operationally consequential, and it is these records that the remainder of the chapter examines.

3.2 Data architecture: two source systems

The plant's maintenance-relevant data live in two corporate systems with complementary scopes. One is Maximo, the plant's CMMS and the system of record for maintenance work; it holds work orders, service requests, and the master data describing assets and their locations. The other is the Shop Floor Control (SFC) system, which records production stoppages, that is, the downtime of each line together with its cause. Both systems feed a Power BI reporting layer, Maintenance Insights, which serves as the routine analysis interface for the maintenance and operations teams and is reached through the corporate Digital Plant portal.[1]

Maintenance Insights is a batch system rather than a real-time one. Data are extracted from Maximo and SFC into a data lake twice per working day, and the report is refreshed each morning with the previous working day's records. A query run directly in Maximo during the day can therefore differ from the figure shown in Maintenance Insights, which reflects the last scheduled

[1] Some figures and data in this dissertation reproduce or derive from Sonae Arauco's information systems and maintenance records. This material is included with the company's permission.

extraction. This cadence is adequate for a daily maintenance meeting, and it also sets the natural rhythm for any automated analysis built on the same data: a single daily update rather than continuous monitoring.

The SFC system classifies every stoppage by a downtime type and a reason. Table 3.1 lists the categories that matter for maintenance. The mechanical and electrical breakdown reasons capture unplanned equipment failures; the preventive and cleaning reasons capture planned stoppages; and the remaining categories cover events outside the scope of maintenance, such as raw-material shortages, product changeovers, and short operator-resolved stops. The plant's reliability indicators, including availability and the mean time between failures (MTBF), are computed only from the mechanical, electrical, preventive, and cleaning categories, since these are the stoppages attributable to equipment condition and maintenance performance.

Table 3.1: Operational downtime categories in the SFC system.

Downtime category	Code	Description
Mechanical breakdown	A02	Stoppage caused by a mechanical failure, for example in bearings, chains, rollers, cylinders, or press belts.
Electrical breakdown	A01	Stoppage caused by the failure of electrical components such as motors, drives, sensors, or programmable controllers.
Production breakdown	A03	Process or operational interruption not directly linked to an equipment failure.
Preventive maintenance	A07	Planned stoppage to execute scheduled preventive tasks.
Cleaning	A08	Planned stoppage to clean equipment and maintain stable operation.
Other stoppages	A06, A09 to A18	Events outside maintenance, including raw-material or utility shortages, product and tool changeovers, and short operator-resolved stops.

In the reporting layer, these categories appear as the downtime ranking of Figure 3.2, where stoppages are ranked and broken down by type, reason, and area for a selected period. In routine use, the team draws on this ranking as its current basis for prioritising maintenance attention, weighing each area's downtime against the reliability indicators computed from it. As a guide it is coarse and retrospective. It is coarse because it sits at a very high level of the hierarchy, no finer than the area and system: a single entry there aggregates the hundreds of positions and components that sit beneath it, so it shows where downtime accumulates but leaves the equipment responsible for it to be found by hand among them. It is retrospective because it reflects the stoppages already suffered rather than anticipating the next ones. It is nonetheless the plant's

established prioritisation practice, and the one against which the framework's own rankings are later set in Section 5.4.

Figure 3.2: Downtime analysis view in the SFC reporting layer. Source: Maintenance Insights.

Within corrective work, Maximo separates unplanned corrective orders (CM-UNP) from planned corrective orders (CM-PLN), a split that corresponds to the immediate and deferred corrective maintenance defined in EN 13306 (Figure 1.1): an unplanned order is the immediate repair that follows an unanticipated failure or a line stoppage, whereas a planned order is deferred corrective work, scheduled once a fault or degradation has been identified that poses no immediate operational risk. Table 3.2 sets out this split within the full taxonomy of Maximo work-order types.

Table 3.2: General taxonomy of maintenance work orders in Maximo.

Work type	Class	Meaning
Corrective (CM)	Unplanned (UNP)	Reactive repair following an unanticipated failure or a line stoppage.
Corrective (CM)	Planned (PLN)	Corrective repair scheduled after a fault or degradation has been identified, without immediate operational risk.
Preventive (PM)	Time-based	Scheduled cleaning, adjustment, lubrication, and overhaul or replacement at fixed intervals.
Preventive (PM)	Condition-based	Inspection, measurement, and condition- or prediction-driven tasks.
Improvements (II)	Improvements (IPV)	Modifications, upgrades, and investments.
Other works (OW)	Other (OWK, PCS)	Tasks outside maintenance and the handling of production consumables.

A corrective work order in Maximo's tracking screen is shown in Figure 3.3, with the fields the analysis draws on: the work type and its CM\UNP classification, the six-level functional

location, the priority, and the status. These are among the fields used by the feature engineering of Chapter 4.

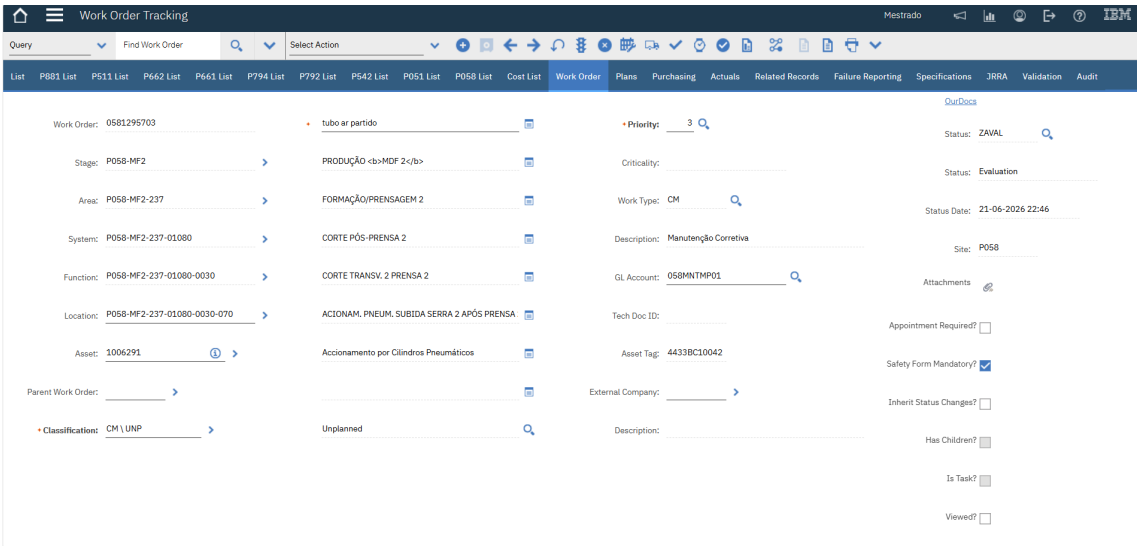

Figure 3.3: A corrective work order in Maximo's Work Order Tracking screen.

The two systems are summarised side by side in Table 3.3. The comparison exposes an asymmetry that is the starting point for the rest of the chapter. Maximo records what was done and to which equipment, but not how long the line was stopped; SFC records the downtime, the most direct measure of a failure's operational impact, but only at a structural level well above the individual equipment, and without identifying that equipment at all. Reconciling the two is therefore neither automatic nor, as the following sections show, always possible.

Table 3.3: Complementary scope of the two source systems.

Dimension	Maximo (CMMS)	Shop Floor Control (SFC)
Primary content	Work orders, service requests, and asset and location master data	Production stoppages (downtime) classified by type and reason
Failure information	What was done, and to which equipment	How long the line was stopped, and why
Structural unit	Functional location, resolved down to the individual equipment position	Plant, area, and system, with no equipment-level location
Time concept	Calendar day, from 00:00 to 23:59	Manufacturing day, defined by production shifts
Extraction to the data lake	Twice per working day	Twice per working day

3.3 Equipment coding and the location hierarchy

Equipment in Maximo is organised by a functional location code, a hierarchical identifier that places every item within the plant according to the function it performs rather than the object it is. The code has six levels, from the plant as a whole down to the individual equipment position, and each level narrows the location by one degree: plant, stage, area, system, function, and position.

One such code is traced through its six levels in Table 3.4, following a single path from the plant down to a specific item on MDF line 2. The complete code at the lowest level, P058-MF2-237-01010-0010-010, identifies fibre bin 2 within the fibre-storage function of the forming and pressing area. A position is the level at which an asset is registered, and the coding does not stop at large machines: the same fibre-storage function also contains the bin's twin feed screw and that screw's left-hand drive motor (tagged 401M1), each occupying its own position.

Table 3.4: Functional location coding levels in Maximo, with an example path from MDF line 2.

Level	Segment	Cumulative code	Description
1. Plant	P058	P058	Mangualde
2. Stage	MF2	P058-MF2	MDF production, line 2
3. Area	237	P058-MF2-237	Forming and pressing
4. System	01010	P058-MF2-237-01010	Storage
5. Function	0010	P058-MF2-237-01010-0010	Fibre storage
6. Position	010	P058-MF2-237-01010-0010-010	Fibre bin 2

The same hierarchy can be browsed interactively in Maximo, as shown in Figure 3.4.

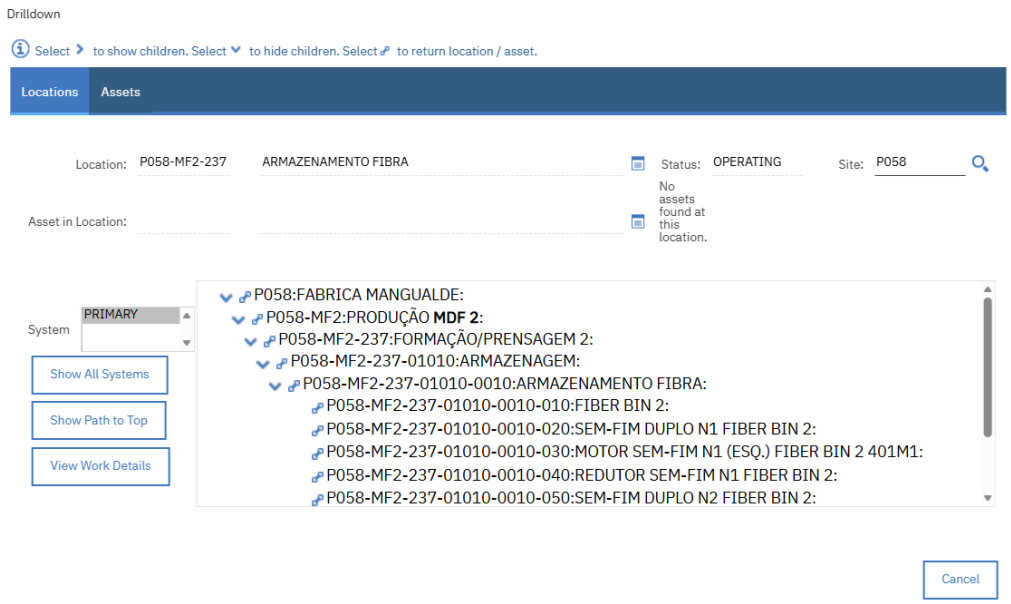

Figure 3.4: Functional location hierarchy in Maximo. Source: screen capture from Maximo.

The tree expands a line into its areas, each area into its systems, and each system into functions and positions, so the depth of the scheme is apparent at a glance. That depth carries a direct

consequence: a single high-level entity such as the press is not one location in Maximo but a system that resolves into dozens of functions and hundreds of positions and components. This fine resolution is what lets failures and preventive tasks be recorded against the part that actually fails; it is also why a machine named as one unit on the production side has no single counterpart in Maximo, a mismatch developed in Section 3.4.

A further property of the scheme governs the unit of analysis adopted throughout this work. The functional location is fixed, but the physical asset occupying it is not: under the plant's rotating-asset process, a motor removed for repair may later be reinstalled in a different position, so the same physical unit may serve a minor door mechanism in one period and a critical drive in the next. Criticality is therefore assigned to the location rather than to the asset. For this reason, the failure history, the reliability estimates, and the recommendations developed in this work are all keyed to the functional location rather than to the asset. Throughout this dissertation, location refers to a functional location at this position level, the unit at which equipment is registered and analysed.

The failure prediction of Section 4.2 and the interval-optimisation of Section 4.4 both work from the records since the most recent asset change, so their outputs reflect the unit currently in place. The plan-adequacy assessment of Section 4.3 offers both views, from the most recent asset change and across the full recorded history. Seeing the full history allows the engineer to assess whether the maintenance plan has been adequate across the life of the location, not just since the last unit was installed. This affects only a minority of the fleet: of the plant's 6,859 locations, only 318 (4.6%) have ever held more than one asset, and of these only 86 belong to the critical and supercritical class.

3.4 The structural gap between SFC and Maximo

The asymmetry introduced in Table 3.3 becomes a concrete obstacle the moment the two systems must be joined. SFC records downtime against the plant, the area, and the system, but no deeper; a mechanical breakdown is logged against a system such as the press, not against the position or component that actually failed. Maximo, as Section 3.3 showed, resolves the same press into hundreds of positions. A single downtime event in SFC therefore points to a region of the plant that, in Maximo terms, contains hundreds of candidate locations, with no field to identify which of them was responsible.

The mismatch is not only one of depth. The reporting documentation states explicitly that the system and function used in SFC are defined differently from those used in Maximo, so the two cannot be aligned even where they appear to share a vocabulary. In practice, downtime, the one quantity that captures the operational cost of a failure and that Maximo does not hold, cannot be attached to the maintenance record of the equipment that caused it.

A conceptual correspondence nonetheless exists. A breakdown logged with a line stoppage in SFC should give rise to a service request in Maximo, which in turn generates an unplanned corrective work order, and the plant tracks this correspondence in aggregate by comparing the number of

mechanical and electrical stoppages reported in SFC with the number of service requests opened in Maximo. This is a relationship between counts over a period, however, not a key between individual events: it confirms that the two systems describe the same failures, without indicating which downtime belongs to which work order. Figure 3.5 summarises the mismatch.

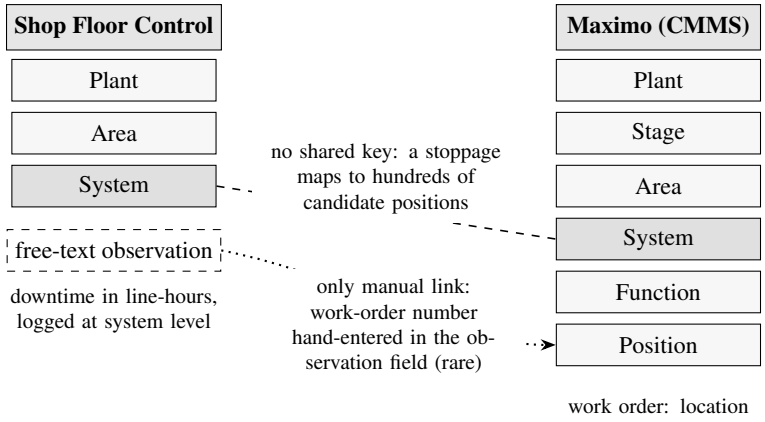

Figure 3.5: The structural gap between SFC and Maximo.

One route does close the gap, but only by hand. Technicians sometimes record the work-order number in the free-text observation field of the SFC stoppage, the field highlighted earlier in Figure 3.2. When that number is present, the downtime can be linked directly to the corresponding work order in Maximo, and through it to the exact location, equipment, and failing component. Figure 3.6 shows one such case.

Start Date	Start Time	End Date	End Time	DWNT (min)	Area	System	Ori System	Func Unit	Ori Func Unit	Type	Reason	Obs	DWNT Index (%)
7/17/2025	10:22:32	7/17/2025	10:25:35	3.05	337 - Prensa III		4000 - Saída Prensa			A - Breakdown	985 - Electric bre...	Substituir cabeça 8 da IMAL na saída da prensa	0.00%
7/17/2025	10:31:49	7/17/2025	10:59:46	27.95	337 - Prensa III		4000 - Saída Prensa			A - Breakdown	985 - Electric bre...	Substituir cabeça 8 da IMAL na saída da prensa	0.00%
12/1/2025	11:45:00	12/1/2025	13:59:58	134.97	337 - Prensa III		2000 - Lavagem			A - Breakdown	986 - Mechanical ...	Substituir bomba de estilha OT - 0581266560	0.02%
5/1/2026	09:00:47	5/1/2026	12:10:56	190.15	337 - Prensa III		5000 - Refinador			A - Breakdown	986 - Mechanical ...	Substituição diverter-valve que estava presa. OT ...	0.03%
5/1/2026	12:10:59	5/1/2026	12:50:06	39.12	337 - Prensa III		5000 - Refinador		508 - Desfibrador - S/ ...	A - Breakdown	986 - Mechanical ...	Substituição correias do sem-fim. OT 0581287309	0.01%
6/4/2025	17:15:00	6/4/2025	18:29:59	74.98	337 - Prensa III					A - Breakdown	986 - Mechanical ...	Subir temperaturas na prensa(Intervenção na bo...	0.01%
11/4/2025	14:01:28	11/4/2025	14:14:00	12.53	337 - Prensa III					XF - Change Overs	ZJ1 - Product cha...	Subir temperaturas da prensa	0.00%

Figure 3.6: Stoppage observation in SFC. Source: screen capture from Maintenance Insights.

The practice is valuable but inconsistent: it depends on the individual operator, it is enforced by neither system, and it is absent from the great majority of stoppages, so it cannot serve as a systematic join.

Given this gap, the present work takes the Maximo maintenance event record as its primary failure signal and treats the cost of downtime as a generic, family-level figure rather than the measured impact of each specific failure (Section 4.4). The gap is thus both a limitation of the present data and a principal opportunity for future integration, taken up in Chapter 6. It is not, however, the only constraint the records impose, and the limitations that recur throughout this work are drawn together below.

Three limitations of the records shape the decisions taken throughout this work, and gathering them here gives a single point of reference for the methodological choices each one drives. The

first is the structural gap just described: a failure can be counted but not weighted by its cost, which is why downtime is treated generically throughout.

The second concerns how failures are described. The intervention text is free-form and entered inconsistently. The failing component, by contrast, is selected from a predefined per-equipment list and recorded together with an ordinal condition grade; but this entry too is optional and not always completed, so a per-component analysis can only reach the components that are consistently logged.

The third is the dependence on manual entry: completeness and accuracy vary with recording discipline, and corrective work that is carried out but never logged leaves equipment looking more reliable than it is, a distortion that resurfaces when the preventive plans are assessed. None of these prevents the analysis, but each narrows what the records can support, and each is referred back to at the point where it shapes a modelling decision.

3.5 SWOT analysis of the data environment

The diagnosis of the preceding sections can be drawn together as a SWOT analysis of the plant's data environment, summarised in Table 3.5.

Table 3.5: SWOT analysis of the plant's maintenance data environment.

Factor	Summary
Strengths	More than four years of structured corrective and preventive history; consistent coding of work orders, downtime, and equipment; failures and interventions recorded down to the position; established reliability indicators and a daily reporting layer.
Weaknesses	Downtime, the measure of failure impact, is held only in SFC and cannot be joined to the equipment or work order; intervention text is free-form, and the failing component, though selected from a structured list and graded on a condition scale, is recorded only optionally and not always completed; data are refreshed twice daily rather than in real time; data quality depends on manual entry; the historical record is, in practice, used reactively.
Opportunities	The existing Maximo history can support failure prediction and reliability analysis without new instrumentation; location-level criticality enables prioritisation by impact; the daily data lake can host automated analysis; a reliable SFC-to-Maximo link would add impact-weighted prioritisation.
Threats	Sparse failure data at the level of the individual location may limit some statistical methods; analysis depends on manual extraction and on continued data-entry discipline; asset rotation and periodic recoding can disrupt longitudinal tracking.

The internal picture is one of genuine richness constrained by a single structural flaw. The plant has accumulated more than four years of maintenance history, coded consistently enough

that every work order carries a type and class, every stoppage a reason, and every item a six-level location. Failures are recorded at the level of the component that failed and graded on an ordinal condition scale, the resolution any predictive or reliability method requires, though that entry is optional and not always completed, which limits how many components can be analysed individually (Section 3.4). The flaw is that the quantity expressing how much each failure cost, the downtime, resides in a separate system at a coarser level and cannot be attached to the equipment record, so the data describe failures fully in every dimension except their operational impact. A milder, operational weakness completes the picture: the records refresh in daily batches and depend on manual entry, and, above all, they have so far been consulted reactively rather than used to anticipate.

The external picture turns these same facts into a direction. Because the history is detailed and already centralised, failure prediction and reliability analysis can be built on the existing records alone, without the cost and delay of new instrumentation, and the criticality attached to each location offers a ready basis for prioritising by consequence rather than by frequency. The principal risk is statistical: failures are sparse at the level of the individual location, which constrains the methods that can be applied there and forces careful judgement about what can be estimated and what cannot. Asset rotation and periodic recoding add a further risk to any analysis that follows a location over several years. Taken together, these strengths, weaknesses, opportunities, and threats define the design space for the framework presented in Chapter 4, which builds its predictive and reliability layers on the Maximo event history, keys every estimate to the functional location, and works within the limits identified here rather than against them.

This chapter fulfils the first research objective, RO1, the diagnosis of the plant's maintenance data environment and the assessment of its fitness to support a data-driven maintenance policy. It characterises the two source systems and the information each holds, the six-level coding that fixes the functional location as the unit of analysis, and the structural gap that leaves downtime unlinkable to the equipment that caused it. Consolidated into a SWOT analysis, this diagnosis establishes what the records can support and what they cannot, and on that basis motivates the framework developed in Chapter 4.

Chapter 4

Methodology

This chapter develops the decision-support framework. It is organised in five parts that follow the path from raw records to deployed guidance. Section 4.1 describes how the maintenance records are turned into a temporally correct, leakage-free dataset. Section 4.2 develops the failure prediction system, from the formulation of the learning problem to the protocol for its out-of-time evaluation. Section 4.3 develops the plan adequacy system, including the population-level hazard analysis that supports it. Section 4.4 develops the interval optimisation for the components whose failures show confirmed wear. Section 4.5 integrates the three systems into the decision-support dashboard delivered to the maintenance team. Figure 4.1 gives an overview of the framework before the parts are developed in turn.

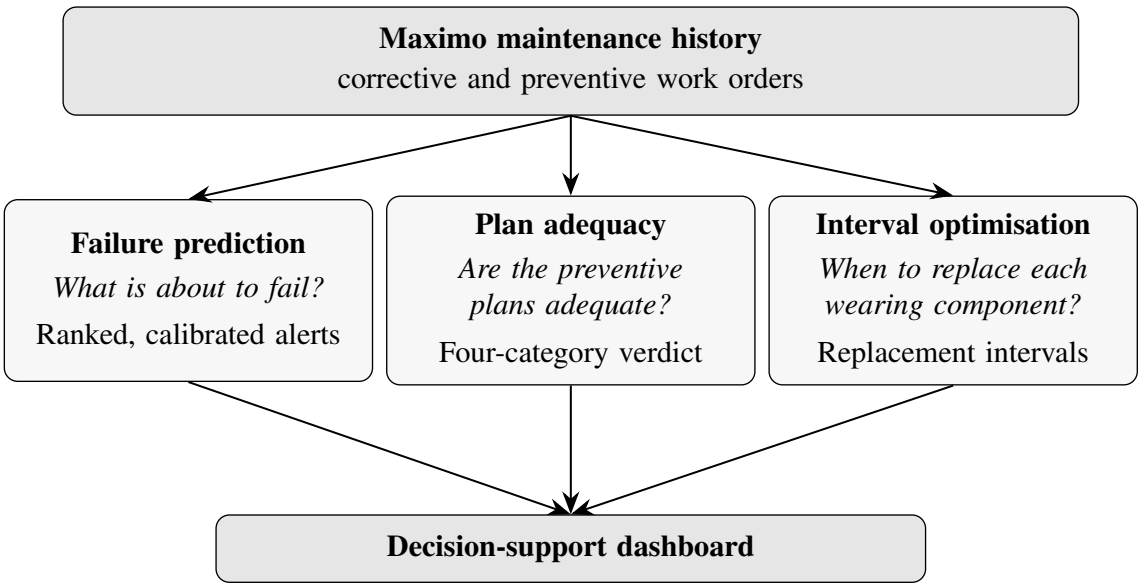

Figure 4.1: Overview of the decision-support framework.

4.1 Data extraction and preparation

This section describes how the maintenance records introduced in Chapter 3 are turned into the dataset on which the three systems of this chapter are built. The earlier chapter set out the two source systems and their general categories; the specific tables, fields, and cleaning steps are given here in full. The extraction draws on both maintenance teams, mechanical and electrical. The records run from 2020 to June 2026. Maximo reached stable and consistent use during 2021, and the way the analysis accounts for the earlier, less reliable period is addressed together with the training protocol in Section 4.2.

4.1.1 Records from Maximo

Maximo contributes four extracts: the location data, the corrective work orders, the preventive work orders, and the condition evaluations recorded against them. Table A.1 of Appendix A lists the fields drawn from each and their meaning.

The four extracts do not form a single flat table; Figure 4.2 shows how they relate.

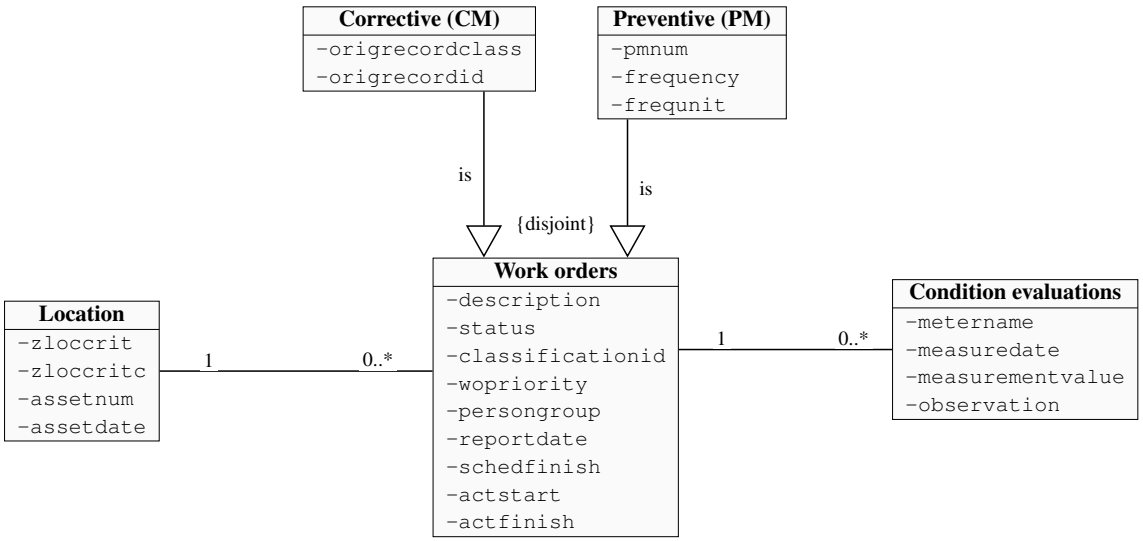

Figure 4.2: UML class model of the Maximo extracts.

A work order, identified by its `wonum`, is the central entity, and the corrective and preventive extracts are two specialisations of it: they share this identifier and the common work-order fields and differ only in the attributes specific to each kind. A work order is of exactly one kind, so a given `wonum` is either corrective or preventive and never both. Both kinds relate to the same two entities: each order sits at a single functional location, while a location groups many orders, and each order may carry condition evaluations, with one order assessing several components on different dates and many orders carrying none at all. Because the evaluations and the location attach to the work order itself rather than to its kind, they apply uniformly to corrective and preventive work.

Two fields carry most of the modelling weight. The classification of a corrective order as unplanned or planned encodes whether the failure stopped the line or was scheduled once identified,

the distinction introduced in Chapter 3; and the priority code separates genuine urgency (immediate to one month) from the four shutdown horizons at which planned work waits for a window. The condition evaluations add a third source of signal. When an evaluation is entered, it records the component concerned together with an ordinal assessment of that component's condition, graded from total failure to no issue noticed on a scale whose levels are anchored in descriptors specific to each component type. These evaluations are entered inconsistently, however, and many orders carry none at all, a limitation drawn together in Section 3.4.

The origin fields make the provenance of each corrective order explicit, and the rule that creates them is what gives the unplanned–planned split its meaning. A corrective order arises in one of two ways. It may follow a service request raised from outside the maintenance team, where a breakdown that stops the line opens an unplanned order and a breakdown without a stoppage a planned one. Or it may be triggered by the condition assessment carried out during preventive work: when a component is graded, the most severe level opens an unplanned corrective order, an intermediate grade a planned one, and a milder grade leaves the item under preventive follow-up. Preventive maintenance therefore acts as a discovery mechanism for corrective work, which is why the unplanned–planned label is informative about the failure process rather than a purely administrative tag.

4.1.2 Records from the SFC system

The SFC system contributes a single extract of production stoppages, with the fields listed in Table 4.1.

Table 4.1: Fields extracted from the Shop Floor Control (SFC) system.

Field	Meaning
`loc`	Plant identifier
`date`	Date of the stoppage
`area`	Production line or area
`type`	Downtime type (categories defined in Table 3.1)
`reason`	Reason code, resolved to a description through the reason dictionary
`dwnt_min`	Stoppage duration, in minutes

4.1.3 Cleaning

The extracted records are cleaned in a single pass before any analysis. Three rules decide what is kept. Cancelled and non-executed orders are set aside, with the scheduled-but-unexecuted ones held separately for the dashboard (Section 4.5); an order counts as executed only when it carries an actual completion date. Orders that carry none of their date fields, reported, scheduled, started, or finished, are removed, since they cannot be placed in time. Finally, only valid six-segment functional-location codes are kept, the level at which equipment is identified; records in any other format are dropped.

What remains is then normalised and assembled. Each plan's defined periodicity is converted to days, so that preventive frequency and failure rate share one time base, which the adequacy ratio of Section 4.3 requires. For each corrective order, the delay between its report and the start of work is computed. It is flagged when negative, which marks a recording error, or when the priority shows the order waiting for a planned shutdown. The production-stoppage records are cleaned in parallel: durations are converted to hours, the type, reason, and area fields are normalised, and rows missing a date, an area, or a duration are discarded. The cleaned orders and their condition evaluations are then joined on the work-order key, and each order to its location (Figure 4.2), yielding a single chronology of failures and executed maintenance keyed to the functional location. This chronology is the common basis for the three systems developed in the rest of the chapter.

4.2 Failure prediction system

The first system answers the operational question the plant raised: at any given date, which functional locations are most likely to suffer a mechanical or electrical failure in the near future, so that the weekly maintenance review can act on them before the failure occurs. This section sets out how that question is posed as a learning problem, how the model is trained and evaluated without leaking information from the future, and how the design was settled through a sequence of controlled experiments.

4.2.1 Planning

The design of the system rests on three decisions taken before any model was trained: what to predict and for which unit, what counts as a failure, and how to estimate performance without leakage.

4.2.1.1 Prediction target and unit of analysis

The clean chronology produced in Section 4.1 records, for each location, the dates of its corrective events. A corrective event here is a mechanical or electrical corrective order: the two maintenance groups are modelled together, a choice made with the plant, since the same location is often served by both and the failures of one frequently accompany the other, so a single alert covering both is more useful to the weekly review than two separate ones. The quantity of interest is the time from a reference date to the next such event. Rather than predict that time directly, which the data regime does not support, the problem is posed as binary classification: whether a location suffers at least one corrective event within the next forty-five days.

The horizon is not arbitrary: forty-five days is close to the median time between corrective events across the population, so it separates the locations whose next failure is imminent from the rest. It also matches the rhythm of maintenance planning. Forty-five days is long enough to act on an alert, by ordering parts and fitting the intervention into a scheduled stop rather than an emergency one, yet short enough that the prediction stays specific to the period ahead rather than

diffuse. In the terms of Section 2.2, the window is the operational stand-in for the P-F interval: the lead time within which detection, decision, and intervention must fit. The reasons for treating failure prediction as classification rather than remaining-useful-life regression, and for reading performance through precision-recall rather than accuracy, were set out in Section 2.2; they apply directly here. The model outputs a single calibrated probability, the estimated chance of a failure within forty-five days, which the dashboard consumes directly.

A prediction is made for each functional location, the level at which the plant assigns both equipment and most of its maintenance plans. Whether prediction is sharper at the level of the individual location or at the coarser level of the function that groups related locations is not assumed; it is tested in Section 4.2.3.

The model is applied only to functional locations with an established corrective history: those recording at least five independent corrective events (Section 4.1). This is a scope decision grounded in the features. Several of them describe the sequence of times between failures, namely its mean and dispersion and the acceleration and overdue ratios derived from them, and none of these is well defined until several intervals exist. With fewer than five events these features are degenerate and the short-window counts mostly zero, so five is the minimum history at which they can be computed meaningfully, and the point below which a short-horizon prediction has no pattern to anticipate. This five-event minimum marks the boundary of the regime in which the method is valid, consistent with predictive accuracy on event data being governed by the depth of failure history (Snider and McBean, 2020).

4.2.1.2 What counts as a failure

Two features of the source data shape the target. First, a corrective order generated from a preventive inspection, identifiable through its originating record, is the maintenance system discovering and scheduling work, not an equipment failure arriving on its own; such orders are therefore excluded from the target and from the failure history, entering the model only as a feature that records how often preventive work is finding problems. Second, the unplanned–planned split introduced in Chapter 3 carries operational meaning: an unplanned corrective order is a breakdown that stopped the line, a planned one a fault scheduled without immediate stoppage. The system predicts the arrival of corrective work and, in a dedicated view, the arrival of the unplanned events that stop the line; the two views are developed below.

4.2.1.3 Training, validation, and test

Performance is estimated with a leakage-free out-of-time backtest, for the reasons given in Section 2.2.3. The dataset is built from monthly snapshots from January 2022 onward. At each snapshot every feature is computed only from the history available up to that date, and the label records whether a qualifying corrective event occurs in the following forty-five days. The records from 2021 serve only as history feeding the twelve-month features, a burn-in period that is never scored.

The data take three roles, kept disjoint in time, all defined relative to one fixed date: a cut-off of 22 March 2026. *Training* is everything the model may learn from, namely the snapshots whose forty-five-day label closes strictly before that cut-off, so that no training label can draw on information at or after it. *Validation* reuses those same training snapshots through the blocked time-series cross-validation described below, which selects the model and its hyperparameters without ever touching the test. *Test* is the single snapshot taken at the cut-off itself, scored against what actually happened in the forty-five days after it. That forty-five-day window runs to 6 May 2026, and the records were extracted later still, on 21 May 2026, so the test outcome is fully observed with two weeks to spare. At the cut-off, 926 equipment locations are scored, of which 187 fail within the window, a base rate of 0.202. This cut-off is specific to the prediction system, which requires its forty-five-day labels fully observed; the interval analysis and dashboard of later sections draw on a more recent extraction of the same records, as they depend on no forward label window. The split is shown in Figure 4.3.

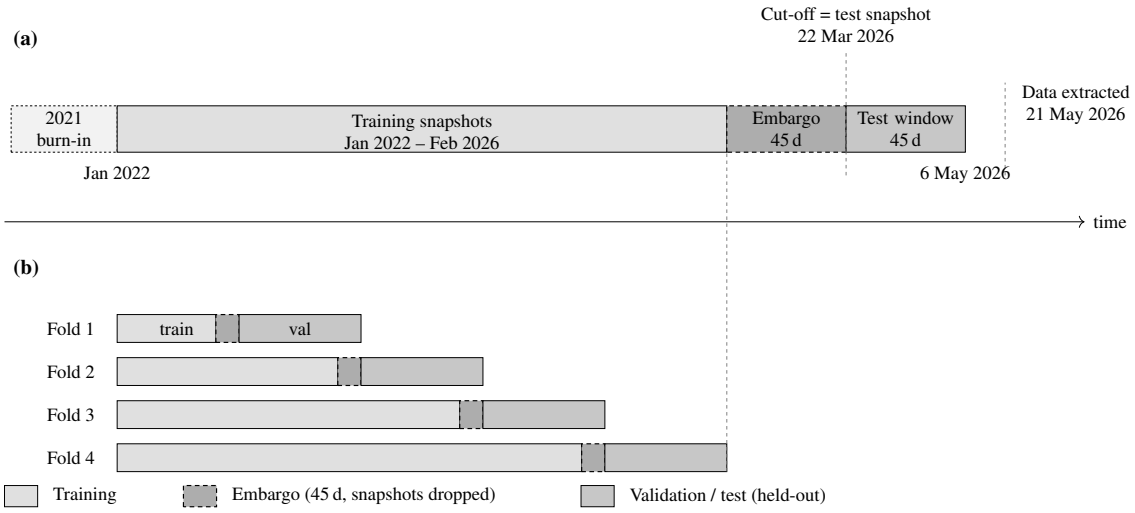

Figure 4.3: The failure-prediction backtest: (a) out-of-time train/test split; (b) blocked time-series cross-validation.

The hyperparameter search is scored by a four-fold blocked time-series cross-validation over the training snapshots. Each fold trains on the earliest blocks and validates on the next, so a validation block always falls strictly after the data used to fit it. A forty-five-day embargo is applied at each fold boundary, dropping any training snapshot whose label window would reach into the validation block. This mirrors the deployment task, predicting a unit's future from its past, rather than the easier question of generalising to unseen equipment that a random or grouped split would pose. The number of folds is lower than the typical five or ten of random cross-validation because of the temporal blocking: folds must be contiguous in time, and each validation block must hold enough failure events for average precision to be estimated reliably. Given how few failure events fall in any short contiguous block, four expanding blocks are the maximum feasible partitioning; additional folds would shorten every validation block below that event count.

The estimate is free of leakage by construction, and it is worth being explicit about every channel through which the future could otherwise reach the model, since each is closed deliberately. A feature could see beyond its snapshot: none does, because every feature reads only the records dated before its own snapshot, the forty-five-day label window included. A preprocessing step could carry future information: it does not, because the normalisation statistics and the categorical encoders are fitted on the training snapshots alone and then applied unchanged to the test snapshot. The class baseline behind the relative-to-class features could average over the future: it does not, being computed only from the history before each snapshot. A training label could overlap the test period: none does, because every training label closes strictly before the cut-off. Model selection could borrow from the future of the cases it scores: it cannot, because the validation blocks are ordered in time behind the forty-five-day embargo. Finally, the calibration map could be learned from scores the model has already fitted: it is not, being learned from the out-of-fold predictions of the cross-validation, as Section 4.2.2.4 sets out. The one structural risk specific to this dataset is that each location contributes several monthly snapshots, so a random split would place a location's later snapshots in training and its earlier ones in test, letting the model learn each case's own future. Splitting on time rather than at random removes exactly that possibility, which is the central reason the protocol is temporal rather than random.

4.2.2 Development

The system is built in five steps: the choice of learning algorithm, the features that describe each location, the interpretation of what the trained model learned, the calibration of its probabilities, and the secondary classifier that separates unplanned from planned failures.

4.2.2.1 Choice of learning algorithm

The data are tabular, heterogeneous, and of moderate size: each location-snapshot is described by features of mixed type and scale, many of them sparse or missing. This is the regime in which gradient-boosted decision trees are the documented method of choice, ahead of both deep networks and linear models, for the reasons reviewed in Section 2.2.2. Three established tree ensembles were compared on the out-of-time test of Section 4.2.1.3, on the same perimeter and the same features, and selected by the blocked time-series cross-validation described there. The few categorical features, namely production area, equipment class, and the two trend flags, are integer-encoded for every model, so the three are compared on identical inputs.

Each ensemble entered the comparison in two forms: with the default hyperparameters of its library, and with the best configuration returned by a hyperparameter search. The search examined thirty configurations per model, drawn at random from a fixed grid. Because the classes are imbalanced, configurations were ranked by average precision rather than by accuracy, which a model that never raises an alert would otherwise maximise. This random search is the sampled form of the exhaustive grid search, sampling configurations from the same grid instead of enumerating every point in it. Doing so is the strategy of Bergstra and Bengio (2012), who show that when only

a few hyperparameters carry most of the effect, random search reaches configurations as good as a full grid at a fraction of the cost; the saving is decisive here, where each grid spans hundreds to thousands of combinations and a single tuned model can take hours to fit. The grid searched and the configuration selected for each model are reported in Appendix C. Performance is read through two ranking measures, the area under the receiver operating characteristic curve (ROC-AUC) and the area under the precision-recall curve, computed as average precision and referred to as PR-AUC below; both are computed on the uncalibrated outputs, since probability calibration leaves the ranking unchanged. Table 4.2 reports the outcome.

Table 4.2: Out-of-time comparison of the three tree ensembles.

	Default		Tuned		
Model	ROC-AUC	PR-AUC	CV PR-AUC	ROC-AUC	PR-AUC
Random forest	0.671	0.392	0.446	0.711	0.425
XGBoost	0.689	0.413	0.437	0.725	0.437
CatBoost	0.714	0.432	0.444	0.719	0.430

Default is the out-of-the-box model; tuned is the best of thirty random-search configurations under blocked time-series cross-validation, selected on the CV PR-AUC, the mean PR-AUC across its validation folds. ROC-AUC and PR-AUC are on the out-of-time test. CatBoost is deployed in its default configuration with probabilities calibrated as in Section 4.2.2.4.

Once tuned, the three fall within a narrow band, with average precision between 0.425 and 0.437 and ROC-AUC between 0.711 and 0.725, and that band lies within the noise of the estimate: the model that leads in cross-validation, the random forest at 0.446, is not the one that leads on test, XGBoost at 0.437, which in turn is barely separated from the other two. A difference of this size, around 0.01 in average precision, does not single out a best model on the metric alone.

What separates them is how they reach that band. The random forest and XGBoost start well below it untuned, at average precisions of 0.392 and 0.413, and need the search to climb into it. CatBoost is already there out of the box, at 0.432, the strongest default of the three, and tuning it moves the ranking not at all, from 0.432 to 0.430. A low-learning-rate candidate from its search even underfits, the high-bias end of the bias-variance trade-off, and harms the calibration of the raw scores, raising the Brier score from 0.144 to 0.183. Strong defaults are a known property of the method, and here they mean the deployed model needs no tuning to perform.

With the three statistically level, the choice rests on operational merit. The system is not an automatic trip that stops a machine; it is a decision aid that ranks a limited maintenance effort. The team cannot inspect every asset each cycle, so the agreed objective is to place at the top of the list the equipment whose failure is both likely and unplanned, which directs scarce maintenance windows to where an unplanned breakdown, far more disruptive than a planned intervention, would otherwise occur. The ranking must do this without overwhelming the team in false alarms, which waste technician hours and erode trust. Under this objective the quantities that matter are the quality of the ordering and the meaning of the probabilities behind the cut-offs, not the labels produced at an arbitrary threshold of one half, which is why the threshold-dependent figures are left out of the comparison above.

Two further properties settle the choice, both operational rather than statistical. First, reaching the band untuned makes the model straightforward to retrain unattended, since the production pipeline can refit it without a hyperparameter search. Second, and decisively, its alert list is the most stable of the three across successive re-trains: the experiment reported in Section 4.2.3 re-trains each candidate at twelve monthly cut-offs and finds CatBoost's top-twenty list the least volatile, on both the overlap of successive lists and the correlation of their order. This matters because the team plans interventions ahead, so a list that does not change composition from month to month sustains their confidence and is auditable, and in a stable list a change is more likely to signal real degradation than re-training noise.

The class imbalance is not handled inside the model: the deployed configuration carries no class weighting, and it is addressed instead by selecting on average precision rather than accuracy, by calibrating the probabilities, and by the operating thresholds that follow, none of which depends on reweighting the training classes.

CatBoost is therefore adopted for the remainder of this chapter, in its default configuration, with probabilities calibrated by the isotonic regression of Section 4.2.2.4. The tuned configuration is not deployed: its difference from the default is within the estimation noise, and a fixed, validated configuration keeps the automatic retraining stable and inexpensive. Tuning would be revisited only if retrospective monitoring were to show the model degrading over time.

4.2.2.2 Feature engineering

Each location-snapshot is described by features computed entirely from the history available at the snapshot date, grouped into families that capture complementary views of the same equipment. None of them looks beyond the snapshot, so the temporal protocol of Section 4.2.1.3 holds at the level of every individual value. For the small minority of locations whose physical asset has been rotated (Section 3.3), this history is taken from the most recent asset change, so the features describe the unit currently in place rather than a succession of different ones.

The first family summarises the *corrective history*: counts of corrective events over the three, six, and twelve months before the snapshot, the mean and standard deviation of the time between them, and the days elapsed since the last one. Two derived signals sharpen this view. Acceleration ratios compare the recent failure rate with the longer-run rate, so that a location failing more often lately stands out from one with the same total but a stable cadence; and an overdue ratio expresses the time since the last failure as a multiple of the mean interval, flagging locations that are overdue relative to their own rhythm. A second family measures *response time*, the delay between a fault being reported and the work starting, together with the share of high-priority orders, both as proxies for how the team has been treating the location.

Three families describe the preventive side and its interaction with failures. *Plan compliance* compares scheduled against executed preventive work over each window, and contrasts the defined frequency of each plan with the frequency actually achieved, yielding flags for plans that are systematically over- or under-executed. The *plan–failure relation* examines each independent

corrective order against the preventive schedule of its location at the time it occurred, assigning it to one of the four situations of Table 4.3.

Table 4.3: The four plan–failure situations into which each past corrective order is classified.

Label	Situation at the time of the corrective order
A	The preventive plan was on schedule and the failure happened anyway
B	The expected preventive task had been executed, but late
C	The expected preventive task had not been executed at all
D	No effective plan covered the location

The comparison uses the expected date of the next preventive execution, with an asymmetric tolerance band, tighter before the expected date than after it, so that small operational delays are not penalised. The share of each category in the location's recent history enters the model as a feature, alongside the count of corrective orders that were themselves generated by preventive inspections, the only place those orders enter the model.

A *degradation* family draws on the ordinal condition grades recorded during inspections, summarising their level, their worst value, and their trend.

The remaining families place the location in its wider context. Shop-floor *line downtime* aggregates the stoppage records of the production line the location belongs to, distinguishing categories such as breakdowns, micro-stops, and changeovers, with short- against long-window acceleration terms. These records enter the model chiefly at the production-line level, as context, since the data do not link a stoppage to the work order that caused it (Section 3.4); they signal how disturbed the location's line has been, not the downtime of the location itself. A small set of matches links specific stoppages to specific corrective orders through the order number cited in the shift log. *Seasonality* encodes the month and the quarter of the snapshot, which capture the summer-shutdown pattern in the failure record. Finally, a set of *relative-to-class* features expresses a location's recent failure count, mean interval, and response time as ratios to the average of the other locations of the same equipment class, so the model can tell whether a value is high in absolute terms or only high for that kind of equipment. The class baseline is itself computed only from history before the snapshot, preserving the temporal discipline.

All features enter the model. No manual pruning is applied, and the contribution of each is read afterwards from the trained model rather than assumed in advance.

4.2.2.3 What the model learns

The trained model is interpreted with SHAP values, which attribute each prediction to its features on an additive scale, as described in Section 2.2.3. The fifteen features with the largest mean absolute contribution are summarised in Figure B.1 of Appendix B, ordered from top to bottom; each point is a location, positioned by its effect on the predicted probability and coloured from low (blue) to high (red) feature value. No single feature dominates, the strongest accounting for

under five per cent of the total attribution, so the prediction is built from many weak signals rather than a few strong ones.

The failure history leads. The dispersion of the time between failures (`mtbf_std`) is the single most influential feature: a location whose intervals are erratic is judged riskier than one that fails on a regular cycle, consistent with wear that has not settled into a stable pattern. Its mean and the recent corrective counts follow, both in absolute form and normalised against the equipment class. The relative-to-class encoding is confirmed as informative: what matters is not how often a location fails outright, but how often it fails for its kind, so that a pump judged against pumps and a motor against motors are read on comparable scales. The class-normalised mean interval, in its two encodings, and the six-month count relative to class are the second, third, and fourth most influential features, behind only the dispersion. Failure acceleration (`cm_accel_3v6`), the recent rate against the longer-run one, pushes the predicted probability in the direction a leading indicator should, and the time since the last failure completes the short-term view.

Context and identity contribute the next tier. The equipment class itself separates a group of locations that rarely fail within the horizon, and whether the location sits on a main production line raises its risk. The shop-floor signal enters through recent line changeover activity, and a calendar-quarter term captures the seasonal pattern in the failure record. A small number of engineered interaction terms also carry moderate weight, the overdue ratio scaled by the line's downtime cost among them, confirming that the combinations the raw features only imply are worth forming explicitly. The condition-evaluation features contribute with smaller weight, reflecting that the plant records condition grades for only part of its equipment.

One reading matters for the consistency of the framework. The equipment class and the criticality indicators appear with modest influence on the predicted probability. This is not a statement that those attributes are irrelevant to failure; it is a statement that they are uninformative given the recency and frequency features, which already encode how a location has been behaving. The same attributes re-emerge as significant population-level structural factors in the hazard analysis of Section 4.3.3, where the effect of criticality is estimated without conditioning on recent history. The two findings are compatible: recent history, when available, screens off the static attributes for prediction, while those attributes remain informative about which kinds of equipment carry higher risk across the population. The plan–failure shares rank low for the same reason, the A/B/C/D composition of recent corrective work being largely screened off, for short-horizon prediction, by how recently and how regularly the location has been failing.

This complementarity was tested directly. The linear predictor of the hazard model of Section 4.3.3 was added to the feature set as a single extra variable and the model retrained under the same protocol; ranking quality did not improve, the change in every metric being negligible. The hazard score ranked among the more influential features only because its ingredients, the previous repair type and the recent preventive intensity, were already present in the feature set, so it acted as a redundant summary rather than a new signal. This confirms that the predictive value of the structural factors is already captured by the history features, and that the hazard model earns its place as a population-level explanation rather than as a predictor, a role developed in Section 4.3.3.

The picture that emerges is coherent: the model draws on the regularity, recency, and intensity of past corrective work, adjusted for what is normal for each class of equipment and for the time of year. It does not rely on any single feature, and the strongest signals are broadly those a maintenance engineer would inspect first.

4.2.2.4 Probability calibration

A boosted tree ensemble ranks cases well but its raw scores are not reliable probabilities: they tend to be pushed towards the extremes, so a score of, say, 0.7 does not mean the event occurs seventy per cent of the time. Because the dashboard does not merely rank locations but applies an absolute probability floor to decide which ones to surface, the scores must correspond to real failure frequencies. They are therefore calibrated, for the reasons set out in Section 2.2.3.

Calibration uses an isotonic regression, which assumes only that the mapping from raw score to probability is monotone and lets the data set its shape, a flexibility that suits the irregular miscalibration of a boosted-tree ensemble. The mapping must be learned on data the model has not memorised. Rather than hold out a single block of the training data, which would withdraw the most recent and most informative months from the model itself, the calibration map is learned from the out-of-fold predictions of the time-series cross-validation of Section 4.2.1.3: every training snapshot is scored by a model that did not see it, the isotonic regression is fitted on those scores, and the deployed model is then refitted on the whole training history. This is the standard route to calibration when data are finite, learning the correction without spending data on a fixed hold-out, and it keeps the map leakage-free, since each score it learns from was produced out of fold. The effect is checked on the out-of-time test set with a reliability diagram, shown in Figure B.2 of Appendix B, and with the Brier score, the mean squared error of the predicted probabilities. The raw scores are already well-calibrated, at a Brier score of 0.144 that the isotonic step holds essentially unchanged at 0.146; calibration preserves the order of the scores and makes the probabilities map faithfully onto observed frequencies, so the cuts that follow can be read as actual failure probabilities.

The calibrated probability is then turned into an operating decision by three fixed cuts. These are operational thresholds, not statistically tuned ones: a single 0.50 cut, the default of binary classification, is in any case inappropriate under the class imbalance of this problem (Section 2.2.3), and the operating point belongs to the task rather than the model. Their purpose is to turn the ranking into priorities the team can act on, the role the literature assigns to a failure early-warning system (Erbiyik, 2022). The plant asked for a graded triage rather than a flat list, so that effort goes first to the most exposed equipment. Response capacity is also finite, varying week to week with the failures logged and the production load, so the cuts organise the ranking into manageable tiers rather than flagging every location of non-negligible risk. The entry floor is kept permissive, since a missed failure costs far more than an unnecessary inspection. A location whose probability falls below 0.30 is not surfaced; at or above it, it is placed in one of three risk tiers, labelled *Monitor – Medium risk* ($0.30 \leq p < 0.50$), *Act – Medium-high risk* ($0.50 \leq p < 0.70$), and *Act – High risk* ($p \geq 0.70$), the verb naming the action and the level the urgency. Within each tier the alerts

are ordered further and can be filtered by criticality or restricted to the line-stopping events, as the dashboard of Section 4.5.2 details. The cuts are round, interpretable probabilities rather than score percentiles, so they keep their meaning as the record grows and remain design parameters open to adjustment as capacity changes; their realised hit rates are reported in Section 5.1.1.

4.2.2.5 Separating unplanned from planned failures

The model so far predicts whether a corrective event will occur, but not what kind. As Chapter 3 set out, only unplanned events stop the line, and these are the ones the plant most needs to anticipate. A location can give rise to both kinds over time, so the type of the next event cannot be read off the location; it has to be predicted. A second CatBoost classifier is therefore trained, under the same out-of-time protocol of Section 4.2.1.3, on the training snapshots that end in a failure, labelled by whether that failure is unplanned. It estimates, for a location about to fail, the probability that the failure is unplanned rather than planned. It uses the same default CatBoost configuration as the primary model, but with balanced class weights to account for the imbalance between unplanned and planned failures.

The two models answer different questions, and combining them gives a sharper view of the events that matter most. The failure probability $P(\text{fail})$ ranks locations by whether they will fail at all; the secondary probability $P(\text{unplanned})$ ranks them by whether that failure would stop the line. The two are combined into a single unplanned-risk score, given by Equation 4.1.

$$\text{score}_{\text{unplanned}} = P(\text{fail}) \times P(\text{unplanned}) \tag{4.1}$$

The product is a ranking score, not a calibrated probability: it multiplies a calibrated $P(\text{fail})$ by an uncalibrated conditional estimate, and serves to order locations by the risk of an unplanned failure rather than to quantify it. The score was tested against the failure probability alone and against a model trained directly on unplanned events; as quantified in Chapter 5, it ranks unplanned failures clearly above the failure probability and on par with the directly trained model, so the unplanned ranking remains a transparent re-weighting of the general one rather than a separate model.

Both probabilities feed the dashboard, which offers two complementary views. A *general* view ranks the equipment by $P(\text{fail})$, for the broadest coverage of upcoming failures; an *unplanned* view keeps that ranking but restricts it to the equipment the secondary classifier marks as likely to fail in an unplanned way ($P(\text{unplanned}) > 0.5$), a ranking cut on the secondary score that is distinct from the probability cuts of Section 4.2.2.4. The two are kept distinct so the team can choose between coverage and selectivity, and their operating behaviour is quantified in Chapter 5.

4.2.3 Design experiments

Three design choices were settled empirically rather than assumed, each under the same out-of-time protocol: the level of analysis, the definition of the target event, and the choice of learning algorithm by the stability of its alert list. The first asks whether prediction should be made for the individual equipment item or for the coarser function that groups related items; the second,

whether the alert should fire on any corrective event or only on the unplanned ones; the third, which of the three ensembles produces the steadiest ranking when re-trained over time.

4.2.3.1 Level of analysis

The model was trained and evaluated at both levels under the same protocol, features, and cut-off, differing only in the unit of prediction: at the function level each equipment item is aggregated into the function it belongs to, one level up the hierarchy of Section 3.3. On the same uncalibrated out-of-time scores reported in Table 4.2, the function level reaches a ROC-AUC of 0.844 and an average precision of 0.649, against 0.714 and 0.432 at the equipment level. This comparison is misleading on its own, however: aggregating items into functions raises the base rate of failure from 0.202 to 0.297, since a function fails whenever any of its items fails, which makes the task mechanically easier and inflates both measures. The decisive consideration is operational. An alert at the function level tells the team that something within a group of items is likely to fail, without saying which; an alert at the equipment level identifies the specific item to inspect. The plant assigns its equipment and reviews its maintenance at the item level, so an actionable alert must be made there. The equipment level was therefore adopted, accepting a lower aggregate metric in exchange for an alert the team can act on directly. The function-level result is retained only as evidence that the signal strengthens under aggregation, the expected behaviour of a sound model.

4.2.3.2 Target event

The choice between predicting any corrective event and predicting only the unplanned ones was also tested directly. A model trained to predict unplanned events alone attains a higher ROC-AUC than one trained on all corrective events, but against a much lower base rate, and it surfaces very few locations: the rare positive class leaves little to learn from and few alerts to act on. Predicting all corrective events covers more of the failures and supports the separate unplanned view through the secondary classifier of Section 4.2.2.5, which recovers the unplanned ranking without discarding the broader signal. The all-events target was therefore adopted as the basis of the system, with unplanned risk handled by the combined score rather than by a dedicated, data-starved model. This is the configuration evaluated in Chapter 5.

4.2.3.3 Stability of the alert list

Average precision does not separate the three ensembles (Section 4.2.2.1), so the choice between them was settled on a second axis: how stable the alert list is when the model is re-trained on the plant's data cycle. Each candidate was re-trained at twelve consecutive monthly cut-offs, from March 2025 to February 2026, giving eleven comparisons between successive lists. At each cut-off the model was trained on the snapshots whose forty-five-day label had already closed, the class-relative features were recomputed as of that cut-off, and the eligible population was held

fixed at the locations with at least five corrective events, so that any movement in the list reflects re-ordering rather than locations entering or leaving the eligible set.

Each comparison was read through two measures, defined here at first use: Jaccard@20, the overlap between the two successive top-twenty sets, which captures *who* is on the list; and the Spearman rank correlation of the predicted scores on the locations common to both, which captures *in what order*. A higher value means a steadier list on both. Each model was assessed under the configuration it would deploy in production: CatBoost in its default, XGBoost and the random forest in their tuned configurations. Table 4.4 reports the averages over the eleven comparisons.

Table 4.4: Stability of the top-twenty alert list across twelve monthly re-trains.

Model	Jaccard@20	Spearman
CatBoost (default)	0.70	0.92
Random forest (tuned)	0.62	0.89
XGBoost (tuned)	0.59	0.86

CatBoost is the most stable on both measures, by a clear margin and with no crossing. The Spearman exceeds the Jaccard for every model, as expected, since the global order of the scores moves less than the boundary of the top twenty. The result separates cleanly from predictive performance: XGBoost, which edges CatBoost on average precision by 0.437 to 0.430 (Table 4.2), is here the least stable of the three. For a list the team plans against, that marginal difference in average precision is outweighed by the steadier ranking, which is why CatBoost is the deployed model.

4.3 Maintenance plan adequacy system

The prediction system ranks equipment by imminent risk; this system asks a different question of the preventive plans already in place: whether they are adequate. A plan can fail to fit the equipment in three ways. It can be executed too rarely to keep pace with failures, leaving the equipment under-maintained; it can be executed far more often than the failure rate warrants, wasting effort on over-maintenance; or it can be executed diligently and still not prevent the failures, the sign of a plan aimed at the wrong thing. This system measures all three against the maintenance record itself and consolidates the verdict into a single view. It draws on the maintenance strategy of Section 2.1 and the reliability and hazard theory of Section 2.3.

4.3.1 Planning

The aim is to classify every location by the adequacy of its preventive plan and, where the plan is inadequate, to say in which direction and what to do about it. The assessment combines two complementary analyses. The first compares, for each location, how often preventive work is actually performed against how often the equipment fails, which separates under-maintenance from over-maintenance. The second looks across the whole population to find the equipment

where preventive effort is high yet failures remain frequent, the equipment whose plan is not delivering. Together they cover the three ways a plan can be inadequate, named above.

This system is distinct from the interval optimisation of Section 4.4. Optimisation asks, for a component that wears out, what the cost-minimising replacement age is, and so requires an increasing hazard and a fitted lifetime distribution. Adequacy asks the prior and broader question of whether the existing plan is even of the right frequency and effectiveness, a comparison of rates and outcomes that applies to every location, whether or not its failures show wear-out. The two are sequential in use: a plan judged inadequate here is a candidate for the optimisation there, where the data support it.

4.3.2 Development

The frequency assessment is built in two steps: a ratio that compares the realised frequency of preventive work with the rate of failure, measured on a common time base, and the translation of that ratio, together with schedule compliance, into a per-location recommendation.

4.3.2.1 A common time base and the adequacy ratio

Comparing the frequency of preventive work with the rate of failure requires the two to be measured over the same window. For each location the observation window runs from the equipment's entry into service to the extraction date, and both rates are computed over it. The dashboard additionally offers the same comparison since the most recent asset change, so the life of the location and the life of the current asset can be read side by side.

The mean time between corrective events, the MTBC, is this window divided by the number of corrective orders. The mean time between preventive executions, the MTBP, is defined the same way, the window divided by the number of preventive executions, with all of a location's preventive plans treated as a single stream. Because the two share the same window, their ratio reduces to the number of preventive executions performed per corrective event, a dimensionless quantity that places the location on the adequacy scale. The MTBC counts every corrective event alike; an unplanned-only rate is computed alongside it for the cases where that distinction matters.

Both boundaries of the acceptable band are anchored on the plant's own target corrective-to-preventive ratio of 25 per cent, that is, four preventive executions per corrective event. The lower boundary is set at this target: below a ratio of four, corrective work is outpacing prevention relative to the plant's own objective, and the location is flagged as under-maintained. The upper boundary is set well above it, at ten, and the gap is deliberate: the preventive count pools every kind of planned work, and routine inspection and lubrication are by design performed many times per failure, so a tighter bound would flag equipment whose frequent low-cost servicing is entirely legitimate. A location is therefore flagged as over-maintained only when its preventive work runs out of all proportion to its failure rate. Between four and ten the plan is retained.

The under-maintenance test is applied only to locations with at least five corrective events, the same eligibility floor used by the prediction system (Section 4.2.1); below this the MTBC rests on

one or two intervals and is too unstable to support a verdict (Snider and McBean, 2020). The over-maintenance test carries no such floor, since a location receiving regular preventive work while recording few or no failures is one of the clearest signals of over-maintenance in the record.

Both boundaries, together with the minimum corrective count, are exposed as editable parameters rather than fixed for the whole plant. The plant's preventive plans were largely inherited and time-based, set without a population-wide view of how each stood against its equipment's failure history; with that view now supplied, an engineer can tune the partition to the resources available, tightening the bands to the most critical equipment when review time is scarce or widening them when there is capacity. The exact value of a boundary is not, in any case, the operative safeguard: the classification surfaces locations for examination rather than for unexamined action. Expanding a flagged location sets its corrective occurrences beside the executions and the descriptions of its preventive plans, from which the engineer confirms or revises the category before acting. Moving a threshold changes which locations enter that examination, not the engineer's ability to judge each one.

4.3.2.2 Compliance and recommendations

The ratio measures whether the realised preventive frequency fits the failure rate; a separate measure asks whether the plans are executed as written. Schedule compliance is the number of preventive executions divided by the number scheduled over the plan's own window, capped at completion. A plan can be of the right frequency on paper yet poorly executed, or executed faithfully yet too infrequent by design, and the two measures together separate a compliance problem from a design problem.

Each category maps to an action, set out with its signature in Table 4.5. For an under-maintained location, the two measures decide which action applies: if the plan is frequent enough on paper but is simply not being executed, the remedy is to run the existing plan (a compliance problem); if the plan is too sparse even when fully executed, a new or shorter plan is needed (a design problem). The over-maintained recommendation, by contrast, is a prompt to review rather than an instruction to cut: on critical equipment, servicing above the bare failure rate can be a deliberate and defensible safety margin (Section 5.2.1).

The same comparison is offered at two levels. At the equipment level each location is assessed on its own plans and failures. At the function level, which groups the equipment that share a function, the plans and failures of the group are pooled, which is the right view when plans are defined for the function rather than for each item. The dashboard switches between the two, so a plan defined functionally is judged as it is managed.

Table 4.5: The four adequacy categories, their signature in the record, and the action each prompts.

Category	Signature in the record	Recommended action
Under-maintained	Preventive work falls below the plant's target rate, with corrective work outpacing it (ratio below four)	Execute the existing plan if it is frequent enough on paper; otherwise create a new plan or shorten an existing one
Acceptable	Preventive executions commensurate with corrective events (ratio between four and ten)	Retain the current plan
Over-maintained	Preventive work far in excess of the failure rate (ratio above ten)	Review for a lower frequency or for consolidation of redundant plans, not an automatic cut
Ineffective	Frequent preventive work that does not suppress recurring failures (drawn from the hazard quadrant, not the ratio)	Investigate whether the plan targets the correct failure mode

4.3.3 Population-level hazard analysis

The frequency comparison above asks, location by location, whether preventive work keeps pace with failures. It cannot say whether the preventive effort is actually working: a location may receive frequent attention and still fail often. Answering that requires looking across the whole population at once, to ask which structural factors raise the hazard of failure and where preventive intensity is not matched by a lower failure rate. This is the role of the covariate-based hazard models reviewed in Section 2.3.2, applied here to the same corrective work-order data. The analysis serves the plan-adequacy assessment in two ways: it estimates the effect of each structural factor on the hazard, and it produces the failure-rate-against-preventive-effort crossing that defines the ineffective-plan category of the dashboard.

4.3.3.1 Analysis dataset and covariates

The data were arranged in counting-process form. Each functional location at the position level of the hierarchy, the level at which physical equipment resides and at which criticality is assigned, was treated as one subject, and its corrective orders were converted into the intervals between successive failures. This level concentrates the overwhelming majority of all mechanical and electrical corrective events, which makes it the natural analysis universe. The time origin of each subject was set at the equipment's entry into service so that exposure is measured from installation, and the interval still open at the end of the extraction window was treated as censored at that date. The construction yields several thousand locations contributing roughly ten thousand inter-failure

intervals, a count of events that, set against the handful of covariates estimated, places the analysis far above the ten-events-per-parameter floor reviewed in Section 2.3.2.

The covariates were specified in four groups, all read directly from the maintenance records. Criticality entered as two indicators, supercritical and critical, against a reference of the locations without an assigned criticality, which form the majority of the universe. Equipment class and production stage entered as grouped categorical factors, with individual levels retained only above a minimum event count and the rest pooled into the reference category, so that no contrast rests on too few events. The state left by the previous repair entered as an indicator of whether the order opening the interval was unplanned, testing whether an unplanned repair leaves the equipment in a less stable condition than a planned one. Preventive intensity, the maintenance-intensity covariate discussed in Section 2.3.2, entered as the logarithm of one plus the number of preventive orders executed in the one hundred and eighty days before the interval began. Because a location typically carries several plans of different periodicities at once, from weekly to yearly, the window is set to capture the recent aggregate preventive effort rather than any single plan: at one hundred and eighty days it spans multiple passes of the frequent plans and typically at least one of the semi-annual ones, so the count discriminates across the range of plan frequencies, while staying recent enough not to reflect long-past work. One further covariate, the logarithm of one plus the count of prior corrective orders, was added in a single specification only, to separate structural effects from the effect of recent history.

4.3.3.2 Model specifications and assumption checks

Three specifications were estimated, mapping onto the formulations of Section 2.3.2. The reference model (M3) is the conditional gap-time model of Equation 2.19, with the baseline hazard stratified by event order, grouped as the first four events and the fifth onward; the stratification conditions the estimates on each location's accumulated event history without spending covariate parameters on it. Two counting-process models following Equation 2.18 were estimated alongside it for triangulation: one with the structural covariates alone (M2), which gives the raw population contrasts, and one that adds the prior corrective-order count (M1), which shows how far those contrasts survive once recent history is modelled explicitly. Reading the three together separates heterogeneity between units from dependence between the successive events of one unit. All models were fitted by partial likelihood, as defined for Equation 2.16, with cluster-robust standard errors grouped by location to reflect the dependence between intervals of the same subject, using the `lifelines` implementation (Davidson-Pilon, 2019).

The proportionality assumption of the reference model was examined with the scaled Schoenfeld residual test of Grambsch and Therneau (1994) introduced in Section 2.3.2, with stratification or an accelerated failure time specification held in reserve for any material violation. Two sensitivity analyses were planned in advance. The first refits the reference model on the subset of locations with at least five corrective orders, the eligibility universe of the prediction system of Section 4.2, to check that the findings hold where the two analyses overlap; because that subset conditions on having failed repeatedly, the between-group contrasts are expected to attenuate. The

second rebuilds the dataset one level higher in the hierarchy, at the function level, and refits the reference model to confirm that the conclusions do not depend on the chosen level of aggregation.

4.3.3.3 The ineffective-plan quadrant

The same dataset supplies the operational by-product that the dashboard uses directly. Among the locations with a stable enough history, taken here as at least three corrective events, each location's observed failure rate (its corrective events per unit of exposure) is crossed with its preventive intensity (the recent preventive-order count defined above), both split at the median of that population. This floor is lower than the five-event minimum used for the ratio, because at least three corrective events are the minimum needed to see how a location's failures evolve and so place it above or below the population split. The quadrant combining an above-median failure rate with above-median preventive intensity is the one of interest: equipment that is maintained heavily and still fails often, where the preventive plan is, on the evidence, not delivering. Locations in this quadrant are flagged as carrying a potentially ineffective plan. The flag is explicitly non-causal and is read as a shortlist for audit, not as proof of inefficiency, for the reason set out in Section 2.3.2: preventive effort is directed at equipment already known to be troublesome, so a positive association between maintenance and failure is expected even where the maintenance helps. The results of the hazard analysis and of this quadrant are reported in Chapter 5.

4.4 Maintenance interval optimisation system

The prediction system of Section 4.2 identifies which locations are about to fail; it says nothing about how often each component should be serviced. That is the question this system answers: for the components whose failures follow a pattern of genuine wear, what preventive replacement interval minimises the long-run cost of maintenance. The system rests on the reliability theory of Section 2.3 and applies it to the cleaned work-order data of Section 4.1.

This system optimises a replacement interval by balancing the cost of a failure against the cost of a preventive replacement, so the failure cost C_f is central to it. That cost includes the downtime a failure imposes, and here the records fall short: downtime is logged only in the SFC system, at the level of the production line, with no key tying it to the work order or the equipment that caused the stoppage (Section 3.4). A per-failure downtime cost therefore cannot be recovered. The model instead uses a generic, family-level downtime value within C_f, and exposes the failure-to-replacement ratio C_f/C_p as an adjustable parameter, whose effect on the recommendations is examined in the sensitivity analysis of Chapter 5.

4.4.1 Planning

Three matters are settled before any fitting: the objective and scope of the analysis, the statistical prerequisites a component must meet to be analysed at all, and the choice of a single replacement model.

4.4.1.1 Objective, scope, and eligibility

The aim is to recommend, for a component subject to wear, the replacement age that balances the cost of renewing it early against the cost of letting it fail. The unit of analysis is the pair of a functional location and one of its listed components: each such component is treated as a distinct failure mode, a distinct competing cause of the equipment's failure, and a separate lifetime distribution is fitted to each. This is the competing-risks construction of Section 2.3 applied at the component level, and it is what lets the analysis reach below the equipment to the part that actually wears out.

The universe of this analysis is the population of critical and supercritical equipment, analysed at the level of its individual components, one level below the equipment. This focus rests on two grounds. First, these are the equipment whose failure carries the greatest operational consequence, where a change of replacement interval is worth acting on. Second, it is for this equipment that the per-component condition grade is recorded most completely and carefully, which yields the more reliable fits and the more solid results. Within that universe a component qualifies for the cost model on statistical grounds: it must carry a failure history dense enough to estimate a lifetime distribution, and that history must show the signature of wear.

4.4.1.2 Statistical prerequisites

The model is applied behind a three-stage eligibility gate, applied in sequence for each candidate component. First, the component must have at least four recorded failures, the floor below which neither the trend test nor the lifetime fit is informative. Second, the Laplace test of Equation 2.1 must not reject its null hypothesis H_0 that the inter-failure times are IID, that is, that they carry no trend; this is the condition under which they admit a renewal representation and a fitted lifetime distribution is meaningful, and a component whose failures are accelerating or decelerating shows a trend, fails this test, and is excluded. Third, the lower confidence bound on the Weibull shape parameter of Equation 2.5 must exceed one, confirming an increasing hazard with statistical confidence, since preventive replacement buys nothing on a component whose hazard is flat or falling.

The confidence level adopted for this bound is 90 per cent. With the short component histories typical of computerised maintenance records, where many components carry only a handful of failures, a one-sided test at the conventional 95 per cent level has little power to detect a genuinely increasing hazard, rejecting almost every legitimate wear-out case through type-II error; the 90 per cent level is a documented compromise between the two error types, retaining statistical evidence of an increasing hazard while not discarding real signal (Guimarães and Leitão, 2025). Only components that clear all three stages, and for which the failure cost exceeds the preventive replacement cost, are carried to the cost model. Figure 4.4 summarises this gate.

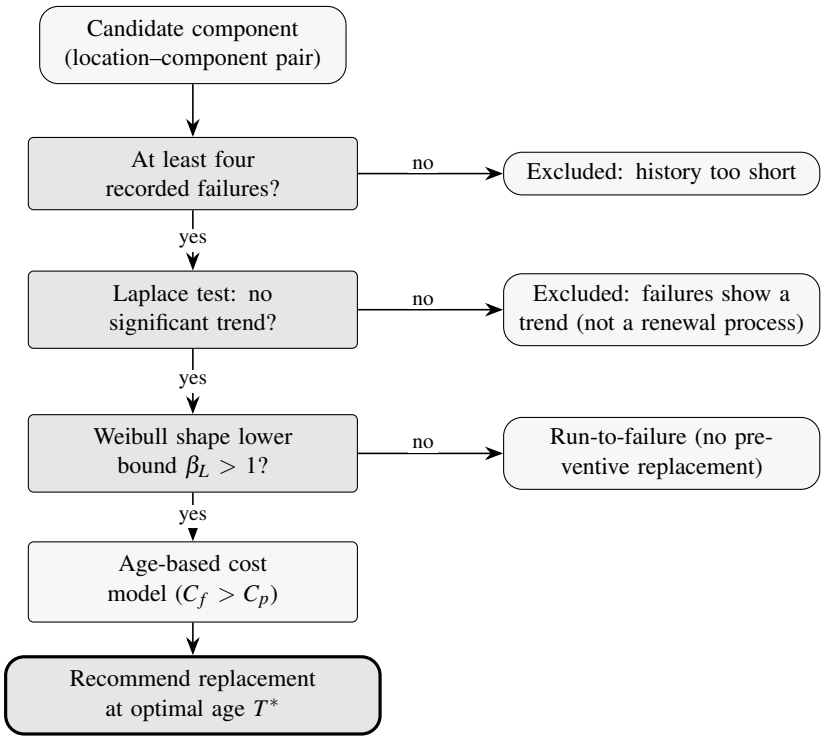

Figure 4.4: The three-stage eligibility gate of the interval-optimisation system, and the outcome at each stage.

4.4.1.3 Why a single model

Section 2.3.1 reviewed three interval models, one for each kind of preventive action: age-based replacement for individual components, a periodic policy for group actions such as lubrication, and the delay-time model for inspections. The present system uses only the first, and the other two were set aside on the evidence of the data rather than by assumption. The periodic group-action model (Equation 2.13) returns a finite optimum only under an increasing hazard pooled across the whole location; in the cases examined the pooled process showed no such trend, so the model collapsed to the trivial recommendation of never intervening, exactly the constant-hazard behaviour anticipated in Section 2.3.1. The delay-time model (Equations 2.14 and 2.15) requires the arrival of defects and the delay between defect and failure to be identifiable from the records, which in turn requires inspections to log when a defect is found rather than only when a replacement is made; the condition records do not separate these reliably, leaving the delay rate unidentifiable, the identifiability condition already noted in Section 2.3.1. The age-based replacement model of Equation 2.11 survived: it matches the systematic time-planned replacements the plant already performs, and its inputs are the component lifetimes the records do support (Guimarães and Leitão, 2025). A single model with firm support was preferred to three whose applicability the data did not equally justify.

4.4.2 Development

The analysis proceeds from the raw work orders to a fitted cost model in three steps: defining what constitutes a failure and a preventive replacement of a component, constructing the inter-failure intervals from those events, and attaching the cost parameters that drive the optimisation.

4.4.2.1 Failure and replacement events

The lifetime distribution is fitted to the times between successive failures of a component, so everything depends on reading the records for when a component actually failed. The criterion adopted is the condition grade, not the type of work order. Each condition evaluation grades a component selected from the equipment's predefined list on Maximo's ordinal scale, and a grade in the most severe band, one to four, records a failure of that component: a loss of function, a fracture, a leak, a seizure, or any of the conditions the scale places at this level. This reading is applied uniformly, whatever the work order that carried the evaluation. A corrective order with a severe grade is a failure; equally, a preventive inspection that grades a component in this band has detected a failure, since the degradation is real whether or not the order was planned. The converse matters just as much: a corrective order whose component was graded in the benign band, recorded as showing no issue, is not counted as a failure, even though it is a corrective order, because no degradation of that component was in fact observed. This correction, treating the observation grade rather than the order type as the definition of failure, is what keeps the lifetime data faithful to the physical events. Table 4.6 gives representative condition descriptions recorded at each grade, drawn from the maintenance records, showing how the severe band corresponds to genuine loss of function while the benign band records components still serviceable.

Table 4.6: The condition grade scale.

Grade	Band	Representative recorded conditions
1	Failure (counted)	Collapsed; broken; out of order; loss of function; seized; structural failure; shaft broken; rupture
2		Not working; missing or disintegrated; burnt; torn; blocked; inoperative; chain broken
3		Leaking; cracked; worn out about to fail; scorched contacts; loss of magnetism
4		Worn; damaged; loose; affected tightness; degraded but functioning
5	Serviceable (not counted)	Minor wear within tolerance
6		Slight signs noticed, no action required
7		No issues noticed

The records further classify corrective orders as unplanned or planned, but this distinction was found to mark only whether the failure stopped the production line, not whether a failure occurred. Both classes are therefore counted as failures, and the unplanned/planned split is not used in the lifetime fit.

The preventive replacements that the age-based model optimises are the revision-and-replacement orders, which Maximo defines as the systematic time-based replacement of a component on a fixed schedule. Each such order renews the component and resets its age, and it enters the lifetime fit as a right-censored observation. The censoring here does not mean the date is unknown: the replacement date is recorded and exact. It means that the event the fit measures, the failure of the component, did not occur at that age, because the component was replaced while still working; its true time to failure would have been longer, so the observed running time is a lower bound on a lifetime never realised. This is precisely the right-censoring of Section 2.3. Lubrication and cleaning plans renew a shared operating condition rather than a component and do not enter the replacement model; inspection plans perform no renewal unless they find a degraded component, in which case the finding is already counted on the failure side through the condition grade. The replacement model is thus matched only to the revision-and-replacement plans, the operational reading of the distinction drawn theoretically in Section 2.3.1.

4.4.2.2 Constructing the inter-failure intervals

For each component the events are ordered in time and converted into intervals, measured from the current asset's installation. Because this installation date is recorded, each component is observed from new, so the time from installation to its first failure is a complete lifetime and forms the first interval. Each subsequent failure closes an interval measured from the previous event, whether that previous event was a failure or a preventive replacement, and contributes a failure observation to the fit. Each preventive replacement resets the clock without contributing a failure. The interval still open at the extraction date is closed as a single right-censored observation, the time the component has run since its last event without yet failing. The Weibull distribution of Equations 2.2–2.6 is then fitted by MLE over these intervals (Equations 2.7 and 2.8), with the censored observations entering the likelihood through the survival function (Equation 2.4); this is the component-level failure-mode decomposition of Section 2.3, in the repairable-system form in which a failure of one mode leaves the exposure of the others running.

4.4.2.3 Cost parameters and the cost ratio

The cost model of Equation 2.11 requires two costs per component: the cost of a preventive replacement, C_p, and the cost of an unplanned replacement upon failure, C_f. The maintenance records hold neither, so both are built from family-level estimates supplied by the plant. The preventive replacement cost C_p comprises the part and the labour of the intervention; the failure cost C_f adds the production downtime that the failure imposes. Labour is valued at 40 €/h, and downtime at 3000 €/h, the average line-stoppage cost recorded by the plant's cost-control department. Table 4.7 reports the resulting C_p and C_f for each component family, together with their ratio.

Table 4.7: Cost parameters by component family, the starting point for the cost ratio.

Code	Family	C_p (€)	C_f (€)	C_f/C_p
ET001	Bearings	230	1810	7.9
ET005	Screw conveyor	920	4000	4.3
ET006	Conveyor belt	2660	8740	3.3
ET010	Chassis / structure	720	2300	3.2
ET011	Gears / pulleys	520	3560	6.8
ET012	Drive unit	1920	6500	3.4
ET013	Structure / supports	620	2200	3.5
ET016	Rollers	330	1910	5.8
ET019	Fan / turbine	1360	5940	4.4
ET023	Filter element	380	1320	3.5
ET040	Seals / gaskets	200	1780	8.9
ET045	Flow-control valve	820	3900	4.8
ET053	Pump / disperser	1660	6240	3.8
ET078	Power supply / battery	240	280	1.2

What governs the optimal interval is not these costs in absolute terms but their ratio C_f/C_p, so the system works in terms of this ratio. The downtime term is the least certain of the inputs: it depends on production buffers, on the volume being run, and on whether a given stoppage halts the line at all, since the buffers between stages can absorb a fault that would otherwise propagate. For the reason set out at the opening of this section, it is therefore a generic, family-level figure, and the ratio is exposed as an adjustable input, with the optimal interval recomputed as it changes; its robustness to this choice is examined in Chapter 5.

The annual saving reported for each component in Section 5.3 follows directly from these costs. It is the reduction in expected cost per unit of time, here per day, the common time base of Section 4.1, that the optimal age delivers over the policy currently in force, scaled to one year, and is given by Equation 4.2.

$$\text{Saving per Year} = \left[C_{\text{current}} - C(T^*)\right] \times 365 \tag{4.2}$$

Here, $C(T^*)$ is Equation 2.11 evaluated at the optimal replacement age. The baseline C_{current} depends on the existing practice. For a component already on a preventive cycle, it is Equation 2.11 evaluated at the current replacement interval. For a component currently run to failure, it is the cost of corrective replacement alone, given by Equation 4.3.

$$C_{\text{current}} = \frac{C_f}{\text{MTBF}} \tag{4.3}$$

In this expression, the mean time between failures is that of Equation 2.6. This baseline is the limiting case of Equation 2.11 as the replacement age grows without bound, so the current and optimised policies are costed on the same renewal-reward footing.

4.5 Decision-support dashboard

The three preceding systems each answer one question, and each produces its output as a table or a model. On their own they are analyses; for the plant to use them daily they must be consolidated into a single interface that a maintenance engineer can open and act on. That is the role of the dashboard: it integrates the failure prediction of Section 4.2, the plan adequacy assessment of Section 4.3, and the interval optimisation of Section 4.4 into one decision-support tool, built entirely on the maintenance records of Section 4.1 and requiring no new instrumentation. This section describes what the tool presents and how it is delivered; the methods behind each view are those already set out, and are referred to rather than repeated.

4.5.1 Design

The dashboard is organised as three views, one per system, reachable from a single navigation bar, so the three questions are kept distinct yet are explored in one place. A common set of filters runs across all three: the user can restrict any view to a production area, to a single location by name, or to the critical and supercritical equipment alone.

The design favours the shop-floor reader over the analyst: each view leads with the actionable list and keeps the statistical detail one click away, behind an expandable row or a tooltip. The result is a tool that puts information the records already held into a form the team can act on.

4.5.2 The three views

The three views are presented in the order an engineer would use them: what is about to fail, whether the plans in place are adequate, and when each wearing component should be replaced.

4.5.2.1 Failure prediction

The first view is the operational front page. It lists the locations most likely to fail within the next forty-five days, each tagged with its risk tier, High risk, Medium-high risk, or Medium risk, set by the absolute probability cuts described in Section 4.2, and sorted so that the most urgent rise to the top. The list can be shown under either of the two views of Section 4.2.2.5: the general view, for the broadest coverage of upcoming failures, or the unplanned view, which restricts the list to the equipment predicted to fail in an unplanned way, to focus on the breakdowns that stop the line. A set of summary indicators reports how the alerts distribute across the windows and how many fall on critical equipment. Expanding a row opens three panels for that equipment: an *action plan*, listing the preventive tasks still outstanding, whether overdue or scheduled ahead, each with its planned date and the historical compliance of its plan; a set of *recommendations*; and a *reliability* panel that links across to the second view. For any equipment the view also opens a component-history chart, which plots each component's condition grade over time on the 1–7 scale of Table 4.6 (Figure E.2): it lets the engineer see which components are degrading and at a glance which are closest to failure, the diagnostic complement to the ranked probability. The

view closes with a risk-against-criticality matrix that places each alert by its failure-probability tier and its criticality, and orders the alerts within each tier by the combined unplanned-risk score, so attention can be focused on the critical equipment that is both close to failure and most likely to stop the line. The view is reproduced in Figure E.1 of Appendix E.

4.5.2.2 Plan adequacy

The second view presents the plan-adequacy assessment of Section 4.3. Its central element is the table that sets each location's mean time between corrective events against the realised frequency of its preventive work, classifying every location as under-maintained, over-maintained, ineffective, or acceptable. The table is filterable by category, by the observation window, and by maintenance group, and it can be viewed at the equipment level or aggregated to the function level, which is the right view when plans are defined functionally. Alongside the classification, the view opens, for a selected location, the per-equipment timeline of corrective events against preventive executions, which makes plain whether prevention has kept pace, together with the descriptions of the preventive plans in force and of the corrective orders recorded against the location. Reading the assigned category against this record, the engineer confirms whether the plan addresses the failures observed, the verification on which the adjustable thresholds rely (Section 4.3.2). The view is reproduced in Appendix E: the general table in Figure E.3, and the per-equipment timeline for one location of each flagged category in Figures E.4, E.5 and E.6.

4.5.2.3 Interval optimisation

The third view presents the interval optimisation of Section 4.4, for the components that clear the eligibility gate. For each, it reports the fitted Weibull shape and scale, the mean time between failures, and the recommended optimal replacement interval, and it draws the cost-per-unit-time curve so the user sees the cost being minimised and where the optimum falls. The cost assumption that drives the recommendation is exposed rather than hidden: the ratio of failure cost to replacement cost, initialised from the family estimates of Table 4.7, is selectable over its plausible range, and changing it recomputes the interval and the cost curve immediately, so the engineer sees directly how the recommendation responds to the costs assumed. The view is reproduced in Figure E.7 of Appendix E.

4.5.3 Implementation and deployment

The dashboard is the end point of a single scripted pipeline that runs from the records to the published interface with no manual step in between, shown in Figure 4.5: extraction, cleaning, feature construction, the three models, and the rendering of the views execute in sequence and produce a self-contained static web application.

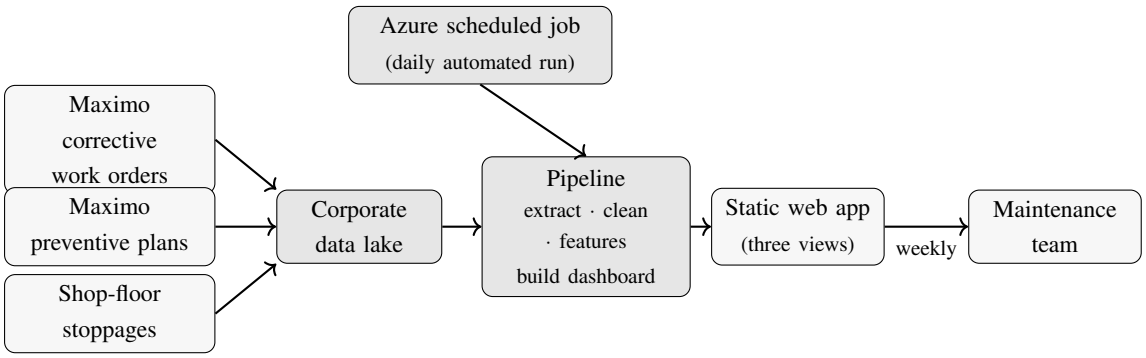

Figure 4.5: The dashboard's data and refresh pipeline as designed to run unattended.

Because the pipeline is fully scripted and runs in minutes, it is designed to rebuild on the daily data cycle, keeping the operational view in step with the work orders as they are logged at negligible cost; the reliability and interval layers evolve over months and would not need daily recomputation, but are rebuilt with the rest, since one daily run is simpler and more consistent than scheduling each layer apart. This refresh cadence is distinct from the review cadence: the team reviews the surfaced alerts weekly, working down the risk tiers of Section 4.2, so each review rests on the latest records.

This cycle is designed to run unattended: a scheduled job on the company's Azure environment draws the records through the system's API into the corporate data lake, runs the pipeline, and publishes the regenerated application. This automation is only partially in place. The corrective work orders are already exposed to the data lake; the preventive-plan tables and the scheduled run in Azure await provisioning by the company's IT. The dashboard is therefore regenerated on a manually triggered basis at present. The constraint is organisational rather than methodological: the bridge already works for part of the data, so the workflow is feasible once the remaining access is granted. Once running on the daily cycle, the same re-training that refreshes the alerts also makes the model monitorable: the stability of the top-twenty list between consecutive runs (Section 4.2.3.3) is itself a diagnostic. A sharp fall in it would signal that the model or the underlying data had shifted, the condition under which the fixed configuration would be revised.

This chapter fulfils the second, third, and fourth research objectives, building on the data foundation diagnosed in Chapter 3. It addresses RO2 by developing a predictive layer that anticipates short-term failures: a calibrated CatBoost classifier trained on monthly snapshots under a leakage-free out-of-time protocol, its scope delimited by the distinction between unplanned and planned corrective work and the level at which prediction is reliable. It addresses RO3 by developing a reliability layer that assesses preventive-plan adequacy, combining the frequency-ratio classification with a population-level hazard analysis, and recommends replacement intervals where confirmed wear-out justifies them. It addresses RO4 by consolidating both layers into a single decision-support dashboard, built entirely on the maintenance records and designed to refresh on the plant's daily data cycle. The performance of each layer and its validation against the plant's history are reported in Chapter 5.

Chapter 5

Results and Discussion

This chapter presents the results of the framework developed in Chapter 4 and so addresses the fifth research objective. Where the previous chapter set out how each system is built, this one reports what each produces when applied to the plant's maintenance record, how the outputs stand up to validation, and what they mean taken together.

The three systems are reported in turn (the failure prediction in Section 5.1, the plan adequacy in Section 5.2, and the interval optimisation in Section 5.3), after which Section 5.4 reports their validation in operation and Section 5.5 draws the three together with their managerial implications and limits. Each system's results are reported first and then placed in the context that makes them meaningful, since a figure such as a predictive metric or a hazard ratio means little without the baseline and the data regime against which it is measured.

5.1 Failure prediction results

This section reports the performance of the failure prediction system of Section 4.2, evaluated under the leakage-free out-of-time protocol defined there: the model is trained on snapshots whose label closes before the cut-off, applied once at the cut-off of 22 March 2026, and scored against the corrective events of the following forty-five days. At that date 926 equipment locations are evaluated, of which 187 fail within the window, a base rate of 0.202. The results are read at the operating point first, where the system is actually used, and then as a global ranking, before being placed in the context of comparable work.

5.1.1 Predictive performance

The dashboard does not act on the full ranking but on the locations whose calibrated probability clears the entry floor of 0.30, which it sorts into the three risk tiers of Section 4.2.2.4. At the cut-off, 57 of the 926 locations clear the floor, and 31 of them fail within the window: a precision of 0.54, more than twice the base rate of 0.202. The full breakdown at the floor is given in Table 5.1.

Table 5.1: Confusion matrix at the operating floor ($p \geq 0.30$).

	Fails	Does not fail
Surfaced ($p \geq 0.30$)	31	26
Not surfaced	156	713

Precision 0.54, recall 0.17, F1 0.25, false-positive rate 0.035. The floor favours precision: the surfaced locations fail at 2.7 times the base rate.

This is a high-precision instrument by design, not a high-coverage one. The 57 surfaced locations capture 31 of the 187 failures, a recall of 0.17, and surfacing few reliable alerts rather than many noisy ones is a deliberate choice that follows the plant's own requirement. Maintenance capacity is finite: the team cannot act on every flagged location, and a short list it can trust is more useful than a long one diluted by false alarms it would only have to filter out. The floor is an adjustable control rather than a fixed property, as Table 5.2 shows: lowering it to 0.20 recovers most failures, a recall of 0.59, but surfaces 299 locations at a precision of 0.37, more than the team can act on, while raising it to 0.40 lifts precision to 0.79 at only 28 locations. The floor at 0.30 is the operating point the plant chose, a list of 57 it can work through.

Table 5.2: Precision and recall at candidate operating floors.

Floor	Surfaced	Precision	Recall
0.20	299	0.37	0.59
0.30	57	0.54	0.17
0.40	28	0.79	0.12
0.50	19	0.79	0.08

The deployed floor of 0.30 trades coverage for a list the team can act on, at a precision well above the 0.202 base rate.

Accuracy is uninformative here. A rule that never raises an alert is correct on the 79.8 per cent of locations that do not fail, barely below the model's 80.3 per cent at the floor; a metric that cannot separate the model from doing nothing cannot guide it. The system exists to surface failures, so it is read through precision, recall, and lift.

The hit rate is the share of a tier's flagged locations that fail within the forty-five-day window, of any corrective event in the general views and of an unplanned one in the unplanned views; the lift is that hit rate over the view's base rate. The same ranking takes different shapes under the dashboard's other views, formed by crossing the general and unplanned rankings of Section 4.2.2.5 with the full and critical-only equipment sets, in the remaining blocks of Table 5.3.

The signal concentrates sharply towards the top, as the first block of Table 5.3 shows. The High-risk tier holds three locations, all three of which fail, a hit rate of 1.00 at a lift of 4.9; the Medium-high tier twelve of sixteen, at 0.75; and the Medium tier sixteen of thirty-eight, at 0.42. Read cumulatively, precision climbs monotonically with the cut: 1.00 above 0.70, 0.79 above 0.50, and 0.54 above 0.30, confirming that the operational cuts of Section 4.2.2.4 separate risk in the order intended.

Table 5.3: Risk tiers at the cut-off across the dashboard's four views.

View	Risk tier	N	Failures	Hit rate	Lift
General, all	Act – High ($p \geq 0.70$)	3	3	1.00	4.9
	Act – Medium-high ($0.50 \leq p < 0.70$)	16	12	0.75	3.7
	Monitor – Medium ($0.30 \leq p < 0.50$)	38	16	0.42	2.1
General, critical	Act – High	3	3	1.00	3.5
	Act – Medium-high	5	3	0.60	2.1
	Monitor – Medium	11	3	0.27	0.9
Unplanned, all	Act – High	3	3	1.00	6.1
	Act – Medium-high	14	10	0.71	4.4
	Monitor – Medium	22	10	0.45	2.8
Unplanned, critical	Act – High	3	3	1.00	3.0
	Act – Medium-high	5	3	0.60	1.8
	Monitor – Medium	3	1	0.33	1.0

Base rates: 0.202 (general, all), 0.287 (general, critical), 0.164 (unplanned, all), and 0.338 (unplanned, critical). The unplanned views cover only the locations the secondary classifier marks as likely to fail unplanned ($P(\text{unplanned}) > 0.5$), a filtered subset in which unplanned events are commoner than across all 926 locations, so their base rate sits above the 0.119 of the full population. In the critical-only views the Medium tier sits near the elevated critical base rate, so its lift is close to one; discrimination is carried by the upper tiers.

The three top-tier locations recur across all four views: the highest-probability locations are also critical and the ones most likely to fail unplanned. The pattern of the first block extends to the other three, with one qualification: in the critical-only views, which cover 164 and 74 locations, the lowest tier sits near the already elevated base rate of critical equipment and adds little, while the High and Medium-high tiers hold hit rates of 1.00 and around 0.60. In routine use the general view is the operating default, covering the widest span of upcoming failures, while the unplanned view is the one the team turns to when the priority is to pre-empt the breakdowns that stop the line.

These counts are anchored to the evaluation cut-off. The ranking is stable across refreshes, but the number of locations in each tier moves as the calibrated probabilities and the failure history evolve; the live snapshot reproduced in Appendix E (Figure E.1), taken later, surfaces a different count for this reason.

Across all 926 locations, on the forty-five-day label, the model attains a ROC-AUC of 0.713 (95% CI [0.667, 0.754]) and an average precision of 0.418 (95% CI [0.354, 0.486]), the intervals from a 2,000-sample bootstrap. These global summaries sit below the operating-point figures because the dashboard acts only on the top of the ranking. Against the no-skill baseline of 0.202, an average precision of 0.418 is roughly twice chance. The 0.432 of Table 4.2 and the deployed 0.418 are the same model, the default CatBoost trained on the whole training history and scored on the out-of-time test, and the only thing that separates them is the isotonic calibration. The calibration is monotone, so it preserves the order of the scores, but it is a step function, so it ties cases the raw scores had separated, and average precision is sensitive to such ties; the ROC-AUC slips marginally from 0.714 to 0.713 for the same reason. As Section 4.2.2.4 reported, the isotonic step leaves the Brier score essentially unchanged (0.144 raw, 0.146 calibrated): its role is to make the scores map faithfully onto observed frequencies, so the cuts that follow can be read as actual

failure probabilities. That the model generalises rather than memorising is shown by the tuned CatBoost, the configuration selected by the cross-validation search: its 0.444 on the validation folds sits at the level of the 0.430 it reaches on the out-of-time test, the small gap in the direction expected of honest generalisation, where a boosted ensemble scores far higher on the data it was trained on than on either. The calibration and precision-recall diagnostics are shown in Figure B.2 of Appendix B.

A natural question is whether reweighting the ranking by the chance a failure is unplanned sharpens the identification of the line-stopping events, or merely adds noise: multiplying the failure probability by an imperfect conditional estimate could in principle have degraded the ordering. It does not. Evaluated against the unplanned events alone, whose base rate among the 926 locations is 0.119, the combined score of Section 4.2.2.5 raises average precision from 0.329 to 0.378 over the failure probability on its own, and the gain is statistically reliable: the paired bootstrap 95% CI on the average-precision difference is $[+0.009, +0.092]$, excluding zero. The conditional layer therefore adds genuine signal about which failures stop the line.

A dedicated model, trained directly on the unplanned events, was tested as the alternative and proved statistically indistinguishable from the combined score: its average precision of 0.368 falls within sampling uncertainty of the 0.378 of the combined score (95% CI on the difference $[-0.049, +0.030]$, including zero). With no performance to gain, the combined score is preferred for two reasons beyond the data-starvation of a dedicated unplanned model already noted in Section 4.2.3. It is transparent, decomposing each alert into the chance of a failure and the chance that the failure stops the line, where a single direct model would return one opaque number. It is also reusable: the conditional probability that forms it drives the unplanned filter, while the combined score itself orders the alerts within each tier of the risk-against-criticality matrix.

5.1.2 Reading the results in context

A ROC-AUC of 0.713 must be read against the evaluation protocol and the data regime that produce it, and three points fix its interpretation. First, the evaluation is a leakage-free out-of-time backtest, not a random cross-validation: the model is only ever scored on a date strictly after its training labels close. This is the more stringent of the two protocols and it lowers the aggregate figure relative to a random split, so the value reflects the difficulty of the test rather than a weakness of the model; an out-of-time 0.713 and a figure from a leaky random split are not comparable. Second, the data regime is the most constrained: sparse corrective-event records with no condition-monitoring sensors. The comparable studies reviewed in Chapter 2 operate under the same difficulty, and the ranking quality reported here, a ROC-AUC of 0.713 and an average precision at roughly twice the base rate, sits in the range Dai and Liu (2023) and Snider and McBean (2020) report as a good result on event-log data of this kind. Those studies read their models at a higher-coverage operating point than the floor adopted here; the same ranking, read at a lower floor, recovers a comparable share of failures, as Table 5.2 shows. Third, the system's misses are understood rather than hidden: as Section 5.1.3 sets out, the model fails on equipment whose corrective history gives little warning, a limit of the available data rather than of the method.

5.1.3 What drives the alerts, and how they are used

The factors behind each alert matter to the team as much as the score, and the attribution of Section 4.2.2.3, summarised in Figure B.1, makes them legible. The strongest drivers are the regularity of a location's failures together with the recent frequency and acceleration of corrective work and the time since the last intervention; the periodic condition grades, such as *worn*, *loose*, or *broken* (Table 4.6), add to this where the records hold them. An alert is therefore not a bare probability but a short explanation a planner can act on: an item flagged on an accelerating count of recent corrective work is examined differently from one flagged on a long gap relative to its usual failure interval. The same attribution explains where the model is blind. Its misses are not equipment without history but equipment whose history gives little warning: items that had been quiet for a long time and then failed without recent build-up, or whose failure was a one-off the prior record did not foreshadow, leaving no recent trace for the model to read. Richer condition data is the natural route to the failures the corrective record cannot anticipate, a direction returned to in Chapter 6.

In use, the system produces a ranked alert list on each refresh of the dashboard. The maintenance planner reviews the surfaced items weekly, working down the risk tiers, reading each alert together with its drivers and the component's maintenance timeline to decide whether it warrants an inspection, a brought-forward intervention, or no action, the human-in-the-loop judgement the system is designed for, since it orders attention rather than asserting that each flagged item will fail.

5.2 Preventive-plan adequacy results

This section reports the verdicts of the plan adequacy system of Section 4.3, applied to the maintenance record at the extraction date. It first reports the four-category verdict over the whole maintainable population, then turns to the population-level hazard analysis that underlies the ineffective category and identifies the structural drivers of risk, and closes with how the verdict is meant to be used.

5.2.1 The four-category classification

Under the default thresholds, and over the whole maintainable population, the classification returns the four-way verdict the framework is built to deliver: 374 locations are under-maintained, their preventive work falling below the plant's target rate; 2,072 are over-maintained, carrying preventive work out of all proportion to their failure rate; 192 are ineffective, absorbing frequent preventive work that does not suppress recurring failures; and the remaining 3,193 fall in the acceptable band. Three of these categories follow from the frequency ratio; the ineffective one is drawn from the hazard quadrant and is returned to in Section 5.2.2.

The four counts partition the 5,831 locations that carry enough history to be assessed; the remaining locations record too few corrective events to evaluate and carry no preventive plan to flag,

and are left unclassified. The ineffective category is resolved first, on the locations with a stable enough history for the hazard quadrant (Section 5.2.2); those it flags are fixed as ineffective, and the frequency ratio then sorts every remaining location into under-maintained, over-maintained, or acceptable.

Two of these verdicts stand out, and they differ in kind. The most actionable is the ineffective-plan category: because it marks equipment that already fails at an above-median rate and continues to fail despite frequent preventive work, it is not an artefact of equipment that is merely serviced often, and it points directly at plans not addressing the failure mode that matters. The most striking in scale is over-maintenance. At the deliberately conservative threshold of ten preventive executions per corrective event, 2,072 locations, over a third of the population, carry preventive work far in excess of their recorded failure rate. This is evidence the plant did not previously hold: a large share of preventive effort is directed at equipment that, on the record, rarely or never fails (how this verdict is meant to be used is taken up in Section 5.2.3).

Restricting the view to the critical and supercritical equipment, the most operationally important, the same four-way split applies: 18 under-maintained, 294 over-maintained, 68 in the ineffective-plan quadrant, and 290 in the acceptable band. The over-maintenance count here invites a more careful reading than in the general population. On critical equipment, where the consequence of a failure is severe, maintaining more often than the bare failure rate would dictate is frequently a deliberate and defensible choice, buying a safety margin that the cost of failure justifies. The ineffective-plan flag, by contrast, is the more pointed signal even here, marking critical equipment where heavy preventive effort is not translating into fewer failures. The per-equipment timeline of the dashboard supports each verdict against the visual record, and is reproduced for one location of each category in Appendix E (Figures E.4 to E.6).

5.2.2 The hazard analysis

The reference hazard model identifies the structural factors that raise the hazard of failure across the population; the significant effects are reported in Table 5.4.

Table 5.4: Significant structural effects from the reference hazard model.

Factor	HR (M3)	95% CI	p	M2 → M3 → M1
Supercritical	1.43	1.23–1.65	<0.001	2.16 → 1.43 → 1.11 n.s.
Previous repair unplanned	1.19	1.11–1.27	<0.001	1.47 → 1.19 → 1.04 n.s.
Preventive intensity (log)	1.16	1.12–1.21	<0.001	1.55 → 1.16 → 1.17
Sanding line (stage)	1.37	1.17–1.60	<0.001	1.65 → 1.37 → 1.10 n.s.
Motor (class)	0.71	0.59–0.87	0.001	0.60 → 0.71 → 0.72

Intervals are cluster-robust at the 95% level; n.s. denotes $p \geq 0.05$. The last column traces each effect across the three specifications of Section 4.3.3: the raw structural model (M2), the reference gap-time model (M3), and the model adding recent corrective history (M1).

Locations classified as supercritical carry a hazard 43 per cent higher than those without an assigned criticality (HR 1.43, 95% CI 1.23–1.65), conditional on event order and the other covariates, an empirical validation of the criticality classification held in Maximo. An interval opened by

an unplanned repair carries a 19 per cent higher hazard than one opened by a planned intervention (HR 1.19, 1.11–1.27), confirming that the unplanned character of a repair is informative about the stability of the period that follows. Recent preventive intensity is associated with a higher hazard (HR 1.16 per log unit, 1.12–1.21); this is read as confounding by indication rather than as a causal effect of maintenance, in keeping with the caution of Section 2.3.2. Among the equipment classes, motors show a 29 per cent lower hazard than the pooled reference (HR 0.71, 0.59–0.87), and among the production stages the sanding line shows the highest structural hazard (HR 1.37, 1.17–1.60).

The last column of Table 5.4 reports each effect across the three specifications estimated in Section 4.3.3. The triangulation across them carries a finding of its own and reconciles this system with the prediction system. The raw structural contrasts shrink substantially once recent failure history is added as a covariate: the supercritical effect falls from 2.16 to a non-significant 1.11, and the unplanned-repair effect from 1.47 to 1.04. Recent history therefore absorbs the static factors almost entirely, confirming empirically the reading anticipated in Section 4.2.2.3: criticality and equipment class are population-level structural factors, not unit-level predictors, once recent history is available. The proportionality assumption was checked with the scaled Schoenfeld test, which the supercritical, motor, and sanding-line effects satisfy; where it flags a departure the event count is large enough that the estimate is read as a time-averaged effect. Both pre-planned sensitivity analyses, on the high-frequency subset and at the function level, return the same direction and similar magnitude for the principal effects, the motor contrast the one exception, losing significance on the high-frequency subset as the conditioning on repeated failure compresses the between-class differences.

The 192 ineffective plans counted above are those the quadrant of Section 4.3.3 flags: among the 654 locations with a stable history of at least three events, they represent 29.4 per cent. Across the population, failure rate and preventive intensity are positively correlated (Spearman's $\rho = 0.25$, $p < 10^{-10}$), the rank-correlation counterpart of the confounding by indication already seen in the regression. The quadrant complements the frequency classification: the latter measures whether plans are executed and at the right frequency, the former whether the outcome responds.

5.2.3 Reading the verdict

The four-way classification is a decision-support indicator, not a prescription, and each category points to a different managerial action rather than an automatic one. An under-maintained plan is a candidate for more frequent prevention, an over-maintained one for redirecting effort towards equipment that needs it, and an ineffective one for investigating whether the plan targets the right failure mode at all. The over-maintenance verdict in particular is best read as the lever this reallocation turns on: a list of where servicing runs far ahead of the failure rate is what lets the team move inspection and replacement effort off equipment that rarely fails and onto the under-maintained and ineffective cases, the equipment that does. It is a prompt to redirect, not to cut: as the critical-equipment split showed (Section 5.2.1), on a high-consequence asset frequent prevention can be the sound choice, which is why the classification surfaces it for a judgement the planner

is best placed to make rather than acting on it. The value to the plant is that this judgement now rests on the full record (the failure history, the executed frequency, and the outcome) assembled in one place, where before it rested on recollection.

5.3 Interval optimisation results

This section reports the output of the interval optimisation system of Section 4.4, moving from the particular to the general: a worked example traces the full calculation on a single equipment item, the results are then read across the population of eligible components, the sensitivity of the recommendations to the cost assumptions is examined, and the section closes with how a planner uses the recommendations and the limit that bounds them.

5.3.1 A worked example

The procedure is shown first on a single equipment item. Table 5.5 presents the complete event history of one sanding-line unit, from which the worked example below is drawn.

Table 5.5: Event history of the example equipment.

Equip.	Days to occurrence and mode (F = failure, P = preventive)	Op. days
1	5,3F; 50,4F; 86,2F; 89,4F; 95,1F; 95,4F; 99,4F; 99,5F; 139,1F; 144,6F; 187,5F; 300,6F; 357,7F; 421,8F; 428,9F; 437,10F; 487,1F; 510,11P; 543,8F; 581,5F; 587,6F; 603,3F; 679,10F; 684,12F; 690,8F; 725,13F; 756,10F; 775,10F; 788,4F; 788,13F; 816,13F; 819,2F; 910,6P; 915,10F; 932,2F; 946,1F; 947,12F; 988,4F; 1015,10F; 1157,10F; 1194,13F; 1197,2F; 1267,10F; 1331,10F; 1349,1F; 1461,4F; 1475,10F; 1490,10F; 1490,4F; 1546,10F	1552

Failure modes: 1 = Drive unit; 2 = Transmission set; 3 = Shaft / piston; 4 = Tensioning/alignment devices; 5 = Piping/ducts/hoses; 6 = Bearings; 7 = Accumulator element; 8 = Conveyor rollers; 9 = Chassis; 10 = Structure; 11 = Gears/pulleys/screws; 12 = Acoustic/thermal insulation; 13 = Instrumentation

The records follow the format used in maintenance practice: each occurrence is given as the number of days since the equipment entered service, tagged with its failure mode and with whether it was a failure (F) or a preventive replacement (P). This unit was installed on 9 March 2022 and has 1,552 operating days at the extraction date. A single equipment item accumulates failures across many distinct components over its life, here thirteen failure modes across more than four years of service, and the preventive replacements (the two events marked P) sit interleaved with the failures. This is the raw material the analysis works from, and the next step is to separate it by failure mode and fit each one in turn.

Two failure modes of this equipment are worked through in full, chosen to show how the eligibility gate of Section 4.4.1 separates a component in confirmed wear-out from one left to run to failure. Table 5.6 reports, for the drive unit (mode 1) and the transmission set (mode 2), the

time t_i of each failure from installation, the inter-failure interval x_i since the previous event, and the interval still open at the extraction date, censored at $T_0 = 1552$ days. The failure times feed the Laplace test of Equation 2.1, and the intervals the Weibull fit of Equations 2.7 and 2.8.

Table 5.6: Worked example: two failure modes of one equipment item.

Mode 1: Drive unit (ET012)

i	Date	t_i (days)	x_i (days)	Type
0	2022-03-09	–	–	installation (origin)
1	2022-06-12	95	95	failure
2	2022-07-26	139	44	failure
3	2023-07-09	487	348	failure
4	2024-10-10	946	459	failure
5	2025-11-17	1349	403	failure
–	2026-06-08	1552	203	censored (T_0)

Laplace ($N = 5$): TS $= -0.86$, $p = 0.39 > \alpha = 0.05 \Rightarrow H_0$ not rejected (no trend)
Weibull: $\hat{\beta} = 1.60$, $\beta_L = 1.10 > 1$; $\eta = 321.6$ days, MTBF ≈ 9.6 mo.
Recommendation: replace at $T^* = 9.3$ months ($C_f/C_p = 3.4$)

Mode 2: Transmission set (ET015)

i	Date	t_i (days)	x_i (days)	Type
0	2022-03-09	–	–	installation (origin)
1	2022-06-03	86	86	failure
2	2024-06-05	819	733	failure
3	2024-09-26	932	113	failure
4	2025-06-18	1197	265	failure
–	2026-06-08	1552	355	censored (T_0)

Laplace ($N = 4$): TS $= -0.08$, $p = 0.94 > \alpha = 0.05 \Rightarrow H_0$ not rejected (no trend)
Weibull: $\hat{\beta} = 1.28$, $\beta_L = 0.80 \leq 1$ (increasing hazard not confirmed)
Outcome: run-to-failure

The drive unit fails five times. Its Laplace statistic is TS $= -0.86$ ($p = 0.39$); with p well above $\alpha = 0.05$, H_0 is not rejected, so the failure process shows no trend and the inter-failure times are fitted as a renewal sequence. The Weibull shape is $\hat{\beta} = 1.60$, with a lower 90 per cent bound of $\beta_L = 1.10$ above one: the hazard is confirmed increasing and the component is in wear-out. At its family cost ratio ($C_f/C_p = 3.4$) the cost model places the optimal replacement at $T^* = 9.3$ months, just below its MTBF (Equation 2.6) of about 9.6 months.

The transmission set fails four times and the Laplace statistic again does not reject H_0 (TS $= -0.08$, $p = 0.94$), so its times also show no trend and are equally admissible for a fit. The shape parameter, however, does not clear the gate: its lower 90 per cent bound stays below one ($\beta_L = 0.80$), so without a confirmed increasing hazard the key condition for preventive replacement is not met, and the component is left to run to failure.

The two modes thus diverge not at the trend test, which both pass, but at the shape gate: only the drive unit shows the statistically confirmed increasing hazard that, as Section 2.3 established, justifies a planned replacement, while the transmission set, on a comparable failure count, does

not.

5.3.2 Results across the population

The eligibility gate was applied across the whole cleaned corpus of critical and supercritical equipment (the scope fixed in Section 4.4.1). Of the 1,195 location–component pairs carrying at least one event, 137 had the four or more failures required to attempt a fit. Of these, 113 did not clear the two statistical stages that follow, taken together: they either showed a significant trend under the Laplace test or lacked a confirmed increasing hazard at the shape gate, and in both cases the renewal-plus-wear-out basis of a preventive replacement is absent, so no interval was recommended for them. This is the expected outcome for the majority of industrial components, and it identifies where preventive replacement would waste effort. The remaining 24 components cleared the full gate and received a recommended replacement interval. Table 5.7 reports these, with their shape parameter, the lower confidence bound, the replacement periodicity currently practised where one exists, the recommended interval at the family cost ratio, and the estimated annual saving against the current policy.

Table 5.7: Components clearing the full eligibility gate.

Location	Component	n_F	β	β_L	Current (mo.)	T^* (mo.)	Saving (€/yr)
SH1-139-01040-0040-060	Chassis	16	1.25	1.12	–	7.5	15
MF3-600-00030-0010-020	Fan/impeller	14	1.55	1.37	–	2.8	2 099
MF2-237-01620-0010-030	Pump	13	1.17	1.02	44.1	9.6	9
SH1-139-01020-0030-010	Structure	12	1.63	1.41	–	3.1	636
EZ1-155-10950-0090-050	Battery/power	7	3.31	2.55	–	0.4	25
SH1-139-01020-0100-010	Structure	6	1.39	1.03	–	10.3	81
SH1-142-01010-0010-020	Drive unit	6	1.76	1.31	–	8.1	898
SH1-142-01020-0010-020	Drive unit	6	1.45	1.07	–	9.5	335
SH2-239-01050-0050-020	Gears/pulleys	6	1.62	1.20	–	4.2	1 209
SH2-239-01050-0060-290	Chassis	6	2.30	1.71	–	5.4	766
WC3-334-10010-0030-060	Drive unit	6	2.77	2.06	12.7	3.1	5 142
SH1-139-01020-0030-010	Drive unit	5	1.60	1.10	–	9.3	567
SH1-139-01020-0100-210	Flow control	5	1.55	1.09	–	4.7	906
SH2-239-01050-0050-020	Structure	5	6.83	4.78	–	7.3	1 393
WB3-333-10030-0020-010	Screw conveyor	5	1.63	1.14	–	5.4	848
WB3-333-10030-0020-020	Pump	5	1.52	1.06	–	1.7	2 853
WC2-234-10040-0010-010	Chassis	5	3.60	2.52	19.5	5.0	1 384
IG1-138-01010-0020-010	Rollers	4	4.01	2.59	–	4.8	1 634
MF2-237-01030-0060-020	Drive unit	4	1.87	1.20	–	9.7	870
MF2-237-01060-0010-010	Conveyor belt	4	3.56	2.49	37.8	5.5	4 822
MF2-237-01110-0070-010	Structure	4	2.86	1.85	–	8.5	647
MF3-337-01620-0050-020	Bearings	4	2.34	1.33	–	7.4	637
SH1-142-01010-0070-020	Filter element	4	3.01	1.94	16.8	4.4	811
WB3-333-10030-0040-070	Seals/gaskets	4	1.77	1.00	–	6.1	518

For the five components that already carry a recorded replacement practice, the recommended interval is in every case much shorter than the one practised: the current cycles run from roughly thirteen to forty-four months, while the optimal ages fall between three and ten, indicating that these components enter wear-out long before they are changed, which is consistent with the failures observed alongside the late replacements. The remaining components run to failure, and the recommendation is to introduce a preventive replacement where none exists. In both cases the message is the same: act earlier than the current regime does. Summed across the twenty-four

components, the recommended intervals yield an annual saving of close to thirty thousand euro against the current policies, with the largest individual contributions coming from the components that combine a strong wear-out shape with a currently long or absent replacement cycle.

Two kinds of case warrant comment, both showing where a recommendation, though statistically sound, carries little economic weight. The first is the battery component (EZ1-155): its very high shape parameter and MTBF of only a few days mark a part in accelerated end of life, which the model correctly reads as severe wear, yet in its family a failure costs barely more than a preventive replacement, so the saving is only a few tens of euro. The case for acting is operational, replacing a unit that fails every few days, rather than economic. The second kind is a few components, such as the chassis and pump of lines SH1 and MF2, whose annual saving falls below a hundred euro: an optimal interval exists, but the gain over the current regime is immaterial, so keeping the present policy is defensible. In both kinds the gate behaves as intended, admitting genuine wear while leaving the operational decision to the engineer.

This shapes how the total should be read. It is concentrated: a few components account for most of the saving, while the cases above contribute a few tens of euro and are retained for completeness rather than for any material gain. It is also a point estimate at the family cost ratios, so its absolute level moves with the cost assumptions, whose effect on both the intervals and the saving is examined next.

5.3.3 Sensitivity and robustness

Because the optimal interval depends on the assumed ratio of failure cost to replacement cost, the recommendations were recomputed across ratios from two to twenty, shown in Figure 5.1.

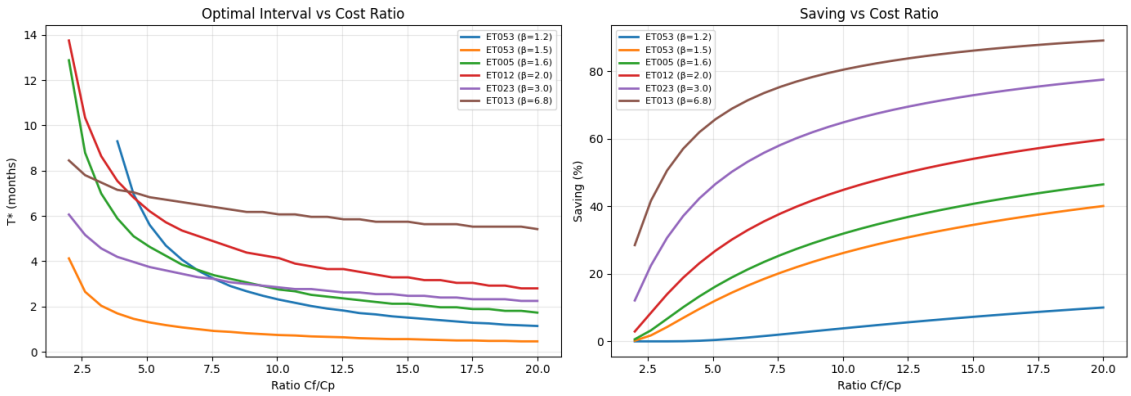

Figure 5.1: Sensitivity of the replacement recommendations to the cost ratio C_f/C_p.

The effect on the optimal interval, in the left panel, is smooth and monotonic: a higher ratio, meaning a more expensive failure, shifts every interval earlier, since the rising cost of failure makes it worth pre-empting sooner. The curves for different components cross in the mid-range of the ratio, so the relative ordering of the intervals themselves is not fixed; what is fixed, and what matters for the decision, is the membership of the set of components that receive a recommendation at all. A component with a Weibull shape close to one acquires an interior optimum only once the

ratio is high enough, the curve for the lowest-shape component entering the plot later than the others; this is the correct behaviour, since when failure is barely more expensive than prevention, running to failure is optimal. The right panel shows the saving rising with the ratio throughout, steeply for the high-shape components and shallowly for those near the threshold. The practical conclusion is that the recommendations are robust to the uncertainty in the cost estimates: which components are worth acting on does not change across the plausible range of ratios, only how tightly each interval is set.

A residual question is whether replacements recorded against an equipment item but not against a listed component might leave some wear unaccounted for, since the censored side of every fit leans on the replacement record being soundly attributed. Of all the revision-and-replacement orders in the corpus, 91 per cent carried a component identifier and entered the analysis directly. Of the 9 per cent that did not, an inspection of their descriptions showed that the large majority were inspections and general revisions rather than component replacements, so their omission does not bias the lifetime fits in any material way. The small residue of genuinely component-changing orders without an identifier is too sparse, and too diluted by non-replacement work, to alter the recommendations.

5.3.4 Using the recommendations and their limits

In use, these recommendations reach the planner through the substitution view of the dashboard. For a component with an existing cycle, such as the drive unit at WC3-334-10010-0030-060, the view sets the current 12.7-month replacement against the model's optimal age of 3.1 months, shows the saving the change would yield, and places both beside the failure timeline that drives them. For a component with no current plan, such as the pump at WB3-333-10030-0020-020, it instead proposes a replacement at 1.7 months where the equipment currently runs to failure. The figure is a starting proposal, not an instruction: the planner weighs it against the component's role and against the cost ratio, which the view lets them vary, and because the recommendation set is robust to that ratio, the decision to plan a replacement rarely turns on the exact figure used. The recommendation says which components are worth a plan and roughly when; the planner sets the final cadence.

These recommendations cover the critical and supercritical equipment (the scope set in Section 4.4.1), where component-level recording is reliable, both in identifying the failing component and in grading its condition. The eligibility gate then keeps the components with a dense enough graded history and confirmed wear-out, so the recommended set is well documented by construction rather than narrowed by weak data. How far the analysis can widen beyond this depends on the completeness of component-level grading across the rest of the fleet, taken up with the framework's other limits in Section 5.5.3.

5.4 Operational validation and team reception

The framework was validated not by a single test but by three lines of evidence that converge on the same conclusion. Two are quantitative and already reported. The out-of-time backtest of Section 5.1 scores the ranking prospectively, on dates strictly after its training labels close, and concentrates the real failures, including the line-stopping ones, at the top of the list. The hazard analysis of Section 5.2 tests the adequacy classification against the failures that followed; and through its M2→M3→M1 triangulation a second, quite different family of models converges on the prediction system's own conclusion: recent corrective history is the operative signal, the structural factors population-level.

The third line is operational. The dashboard was reviewed in use by the plant's maintenance engineers, who set its outputs against their own established method of judging maintenance priority. Their reading corroborated the analyses, extended them below the level at which that method operates, and, in the course of the review, surfaced a set of findings about the records themselves. The two threads are taken in turn: what the review established about the tool's outputs (Section 5.4.1), and what it established about the data those outputs rest on (Section 5.4.2).

5.4.1 The verdicts in use

The natural benchmark for the framework is the method the team already relies on. To set maintenance priority, the engineers work from a matrix built on the plant's SFC downtime records, crossing the MTBF against the time to repair and singling out the areas that fail most often and take longest to restore. That matrix, like the records it draws on, resolves only to the system level (the level at which SFC logs downtime, as Chapter 3 sets out): it identifies a system or an area as critical, but cannot show which of the equipment within it drives that criticality. The first result of the review is that, read at this shared system level, the framework's account of where the problems lie agreed with the engineers' own: the areas their matrix marks as critical are the areas the tool ranks highest. Two independent readings of the plant, one manual and long established, the other automated and developed here, coincide.

Where the framework goes beyond the benchmark is in resolution. Because every estimate is keyed to the functional location rather than to the coarser system record, it descends below the level at which both the matrix and the engineers' own memory operate, isolating the specific equipment that makes a critical system critical, information the team could not previously recover from the SFC records. Agreement at the system level establishes that the tool reads the plant correctly; the equipment-level resolution is what it adds on top, and that is the level at which maintenance is actually planned and carried out.

Within that resolution, each of the three views was examined in turn. The adequacy view was read first. Several of the plans it flags as over-maintained proved, on the engineers' reading, not to be genuine over-maintenance but corrective work that is carried out and not entered into the system: where the repairs that justify a plan go unrecorded, the equipment appears to receive

heavy prevention against an almost empty failure history, which is exactly the signature the over-maintenance flag detects. The flag was therefore correct in pointing to those plans as worth examining; what the examination revealed was not waste but a gap in the recording of corrective work, taken up with the wider recording issues in Section 5.4.2. This refines the population result of Section 5.2 rather than contradicting it. That result classifies each plan on the evidence the records hold, so a plan whose corrective work went unrecorded is correctly read as over-maintained given the record as it stands; what the engineers add is not a correction of the count but an explanation, for a subset of cases, that its cause is a gap in the record rather than true excess.

The prediction view was judged directly actionable, and it was the output the team adopted most readily. It lets them prepare the equipment it ranks as high-risk (ordering spares, scheduling attention, bringing forward inspection) rather than reacting once a stoppage has happened, and the engineers identified it as the layer around which they would organise much of their forward planning.

Most of the flagged equipment they recognised as items they already knew to be failure-prone, which corroborates the ranking against their judgement; a smaller number they had not been treating as critical, and on reviewing the history behind the alert they endorsed those too. A few flagged locations they traced to the records rather than to the equipment: entries logged as corrective work that were in fact automatic alarms, inflating the apparent risk, a recording practice taken up in Section 5.4.2. They read the tool's reach as set by the records behind it, expecting it to sharpen as the recording of failures becomes more complete, the direction developed in Chapter 6. Beyond that corroboration, the tool provides a systematic reading of the entire corrective record, revealing insights that no individual engineer had previously assembled.

The interval view drew a concrete commitment: the engineers confirmed the components it singles out for a planned replacement as being among their most critical, and undertook to revise those components' plans accordingly. This endorses the selection the analysis makes; the precise cadence remains subject to the field confirmation set out in Section 5.5.3.

5.4.2 What the review revealed about the records

Read together, the engineers' responses did more than validate outputs: they turned the tool into an instrument for examining the records themselves, and this reaches the foundation on which the whole framework rests. The gaps it surfaced were of three kinds. The first is missing corrective work, already met above in the over-maintenance flags: repairs performed but not entered, so that a plan appears to guard equipment with no failure history. The second is its mirror image, spurious corrective entries: automatic alarms logged as corrective work, met above in the prediction review, which inflate a location's apparent failure history. The third is structural miscoding, of which the clearest case is a parent location that is not recorded as one: it carries a maintenance plan but is entered, and therefore analysed, as a single isolated piece of equipment within its function rather than as the parent of the equipment beneath it, which is inconsistent with how the rest of the hierarchy is coded. None of the three is an error in the analysis; all are properties of the source data that the analysis, by reading the whole history in one place, brought into view.

What matters most is the significance the engineers themselves drew from this: they knew the records needed to improve, but not where the improvement should begin; seeing what the tool extracts from the records, and where its account departs from what they know to be true, gave them a concrete target for that effort rather than a general aspiration. The point is not confined to the adequacy layer. Because every one of the three systems reads the same records, the same gaps and inconsistencies bear on all of them, and part of the engineer's role in using the tool is to separate a real signal from an artefact of recording and, in doing so, to correct the record. The tool does not only consume the data foundation; it exposes where that foundation is weak and directs the work of strengthening it.

This has an effect that compounds. Until the records were turned into a visible account of what they yield, the maintenance history had been consulted reactively rather than analysed, as observed in Chapter 3, so the discipline of complete and direct recording had shown the team little return and the records were kept unevenly. Seeing what the tool extracts from those records (the rankings, the adequacy verdicts, the replacement proposals) made the return concrete, and the team reports that recording practice has since begun to improve. The framework's value thus extends beyond its own outputs to the quality of the data any such analysis depends on: more complete records make each of the three systems more accurate, and the more the systems yield, the stronger the case for the recording discipline that feeds them.

5.5 Discussion

The preceding sections evaluated each system on its own terms; this closing section reads them as one body of results. It considers the framework as a whole and the shift it brings to the maintenance function, records the approaches that were examined and set aside during development, and states the limits within which the results should be interpreted.

5.5.1 The framework as a whole

Taken together, the three systems answer the three questions the maintenance function faces, and they do so on one foundation. What is about to fail, whether the plans in place fit the equipment, and when each wearing component should be replaced are distinct questions, and each is addressed by the system suited to it, drawing only on the corrective and preventive history the plant already records.

For the maintenance function, this amounts to a shift in posture from reactive to anticipatory. The same records that were previously consulted only after a failure now support three forward-looking decisions (which equipment to watch, which plans to revise, and which components to place on a preventive replacement) and the effort this moves is not new effort but existing effort better aimed: inspection directed towards the equipment most likely to fail, preventive work shifted off equipment that does not need it and onto equipment that does. The systems are positioned throughout as decision-support rather than automation, ordering, classifying, and proposing

while the engineer supplies the judgement, which is both the safeguard against acting on a single imperfect figure and, as Section 5.4 showed, the reason the tool was accepted in use.

What the framework is worth differs by system, and only one part of it is quantifiable in cost terms. The interval analysis yields a monetary figure, the annual saving of Section 5.3, resting on the plant's own cost rates and confined to the components where wear-out is statistically confirmed. The saving therefore rests on two grounds of unequal strength. Which components warrant a preventive replacement barely changes across the plausible range of cost ratios, because that decision turns on confirmed wear-out rather than on the exact costs; the size of the saving, by contrast, is driven directly by the cost ratio, an uncertain input, and so should be taken as indicative of scale rather than a precise amount. The prediction and adequacy layers, by contrast, carry no monetary figure, since quantifying one would require a counterfactual the records do not support; their value is the better-directed effort described above.

The scale of the burden this effort addresses is nonetheless visible in the record: on the two production lines where stoppage is logged, breakdowns accounted for some 534 line-hours in the most recent full year. That figure fixes the order of magnitude of the problem, but no more: as Chapter 3 sets out, the shop-floor downtime it draws on mixes genuine equipment failures with stoppages that are not maintenance at all, such as production interruptions, changeovers, and operator-resolved stops, and it sits above the equipment level, so it cannot be attributed to the equipment that caused it.

5.5.2 Approaches considered and set aside

Several directions were considered and not pursued, and the reasons are themselves results.

The first was a separate analysis of predictability by failure mode, which specific component will fail, a finer view that would let intervention target the component most at risk. It proved unworkable because the failing component is recorded too inconsistently to support a reliable per-mode split: as Section 3.4 sets out, the field is entered for only a minority of corrective events, and inconsistently across much of the equipment. This limitation is partially mitigated in the dashboard: for a given piece of equipment, it plots each component's condition grade over time (Figure E.2), allowing the engineer to track degradation and judge which components are most at risk.

The second was the use of reliability-derived features as model inputs: adding the hazard model of Section 4.3.3 to the prediction pipeline did not improve the ranking, for the reason given in Section 4.2.2.3, the structural factors it captures being population-level effects already absorbed by recent corrective history at the unit level. It is therefore retained in the reliability layer alone.

The third was heavier model tuning. A comparison of candidate algorithms and configurations returned only marginal differences in ranking quality; the extended search added hours of computation for no reliable gain, and one tuned candidate even degraded the calibration of the raw scores. A simpler, well-calibrated model was retained in preference, since a tool the team must trust is better served by stability than by a fractional gain in a single metric. The detail of that comparison is in Section 4.2.2.1.

5.5.3 Limitations of interpretation

Three limits bound how the results should be read. The first is causal: the systems are built on observational records, not designed experiments, so the positive link between preventive intensity and hazard, visible in both the regression and the rank correlation of Section 5.2.2, is confounding by indication (more preventive work goes where more failures occur) rather than evidence that preventive maintenance causes failures.

The second is evidential: the replacement intervals rest on the convention that a severe condition grade marks a component replacement, sound on average but unverified case by case, so they warrant field confirmation before a cadence is fixed as policy.

The third is intrinsic to the data regime: the framework sees what the records hold, so without condition monitoring it does not anticipate the failures that leave no recent trace. This is less a flaw in the method than the natural boundary of sensorless event data, and it points to exactly where added condition data would extend the work, the direction set out in Chapter 6.

With that ceiling acknowledged, the framework stands as a coherent decision-support system built entirely on the records the plant already keeps: a prediction layer that concentrates failures at the top of a reviewable list, an adequacy layer whose verdicts the hazard analysis confirms and the engineers validated in use, and an interval layer that turns confirmed wear-out into a quantified saving. This chapter thereby meets the fifth research objective, evaluating each system of the framework against the real failure record and in operation, and establishing both what it delivers and the data limit that bounds it.

Chapter 6

Conclusions and Future Work

This dissertation set out to convert the maintenance records the Mangualde plant already keeps into guidance its maintenance team can act on. Two problems motivated the work: a four-year computerised record that served retrospective reporting but gave no early warning of failures, and a preventive policy, much of it inherited from another site, that had never been calibrated against the plant's own failure history. Five objectives followed, and this chapter draws the work together: it summarises what each delivered, states what the framework contributes, sets out the limits that bound it, and identifies the directions that would extend it.

6.1 Summary of the work

The first objective, a diagnosis of the data environment, shaped everything built afterwards. The plant's records proved rich enough to support analysis (four years deep, consistently coded, and resolved to the equipment position that failed), but they carried one structural flaw: the downtime that measures a failure's cost lives in a separate system, at a coarser level, and cannot be joined to the equipment that caused it. This fixed two choices carried through the work: every estimate is keyed to the functional location rather than to the rotating asset, and the cost of downtime is treated generically rather than per event.

On that foundation, the framework answers the three questions the maintenance function faces. The second objective supplied the early warning the team lacked: a predictive layer that estimates, for each location, a calibrated probability of corrective work within forty-five days, presented as a ranked and explained list of alerts. Under a leakage-free out-of-time test, it concentrates the real failures at the top of the ranking, every location in the top tier failing and four in five failing above the medium-high cut, at a level in line with comparable studies on event-log data without condition sensors.

The third objective addressed the preventive policy itself. A reliability layer characterises each location's failure process and classifies whether its plan is adequate. Its most pointed verdict is the ineffective-plan category, marking equipment that absorbs frequent preventive work and still fails at an above-median rate. The most striking in scale is over-maintenance: even under a deliberately

conservative threshold, set above the ratio that routine servicing alone would produce, over a third of the maintainable population (2,072 locations) carries preventive work far beyond its recorded failure rate, a pattern the plant had not previously quantified and one the framework surfaces for review and reallocation. For the components where wear-out is statistically confirmed, the same layer recommends a replacement interval, amounting in aggregate to a saving of the order of thirty thousand euro a year against current practice.

The fourth objective consolidated these outputs into a single dashboard, designed to refresh on the plant's daily data cycle, so the guidance reaches the team in a form they can open and act on.

The fifth objective, validation, is where the separate pieces are shown to hold together, confirmed by three lines of evidence that agree: a prospective backtest, a survival model that independently reproduces the predictive layer's reading of which factors matter, and the judgement of the engineers who plan and carry out the work. The review also refined a result rather than merely confirming it: several plans flagged as over-maintained proved to mark corrective work that is carried out but not recorded, so the flag correctly identified an anomaly whose cause was a gap in the records rather than true over-maintenance.

The five research objectives are therefore all fulfilled: the data environment is diagnosed (RO1), the predictive and reliability layers are developed (RO2 and RO3), both are consolidated into a working dashboard (RO4), and the framework is validated on the plant's history (RO5).

6.2 Contributions

For the plant, the work delivers a working decision-support tool built entirely on existing records, with no new instrumentation, that moves the maintenance function from reacting to failures towards pre-empting them and re-aims existing effort rather than adding to it. The scale of the burden it addresses is visible in the record: on the two lines where stoppage is logged, breakdowns accounted for some 534 line-hours in the most recent full year. A further, indirect contribution followed from putting the tool in front of the team: by making visible what the records yield, it gave the discipline of complete recording a concrete return, and recording practice has begun to improve (Section 5.4), which compounds the value of every analysis built on those records.

Beyond the plant, the dissertation makes three contributions to research and to comparable industrial projects. First, it demonstrates that sparse CMMS event-log data, even without condition-monitoring sensors, can support an integrated decision-support framework combining short-term failure prediction, preventive-plan adequacy assessment, reliability analysis, and interval optimisation on a single data foundation. This closes the gap identified in Section 2.4, where the reviewed studies typically address these tasks in isolation. Second, it specifies an explicit and reproducible protocol for leakage-free failure prediction on temporal maintenance records: snapshots whose labels close before the training cut-off, blocked time-series cross-validation with an embargo matching the label horizon, out-of-fold calibration, and a strictly out-of-time test (Section 4.2.1.3). The protocol is independent of this plant and can be adopted directly by other studies of maintenance

event data. Third, it documents the modelling paths that were tested and rejected, with the evidence behind each rejection (Section 5.5.2), which delimits what can be extracted from industrial maintenance records of this kind and spares comparable projects the same exploration.

Specific methodological lessons accompany these contributions. On sparse, sensorless event-log data, a simple, well-calibrated model performs as well as heavier alternatives; an unplanned-failure ranking can be sharpened by a transparent re-weighting of the general model rather than a separate opaque one; and recent corrective history absorbs the static structural factors of criticality and equipment class, which reconciles the predictive and reliability layers.

6.3 Limitations

The framework's reach is set by the records it reads, and its limits are those set out in full in Section 5.5.3. Because it is built on observational history rather than designed experiment, its associations are not causal, so its recommendations are proposals offered for the engineer's judgement rather than prescriptions. It is bounded, too, by the quality of the records themselves (the missing link between downtime and maintenance, the uneven component-level recording, and the dependence on manual entry, diagnosed together in Section 3.4), each of which limits what it can establish.

Above all, the framework can see only what the records hold. With corrective history but no condition monitoring, it cannot anticipate the failures that leave no recent trace, and this single limit caps the predictive layer's reach and confines the interval analysis to the small set of well-documented components. It is the same constraint diagnosed at the outset, and it is the natural starting point for the work that would follow.

6.4 Future work

The most direct extension addresses that ceiling: complementing the corrective record with condition data. Vibration, temperature, or current signals on the most critical equipment would reach the failures that have no warning in the event history, and would let the predictive layer move from anticipating corrective work to tracking physical degradation. It is the single change that would most widen the framework's reach.

A second direction is to close the structural gap between the two source systems. A reliable link from each downtime event to the work order that caused it would let every failure be weighted by the production it cost rather than merely counted, and that single addition would propagate through the whole framework: predictions validated against the hours of output lost, maintenance prioritised by real operational impact, and the value of the predictive and adequacy layers, left here in terms of better-directed effort, expressed in the monetary terms that only the interval layer currently reaches.

A third direction concerns the reliability layer itself. The recommended intervals should be confirmed in the field, by following the components placed on a planned replacement and comparing their realised life against the prediction; and the analysis would widen well beyond its present set if the recording of component-level condition grades were made more complete. Where the plant can vary a preventive frequency deliberately and observe the result, a move from observational association towards controlled, causal evaluation would put the adequacy findings on firmer ground.

Two further directions are operational rather than analytical, concerned with putting the framework into routine use rather than with extending what it computes. The first is on the company's IT side: completing its automation. The refresh pipeline of Section 4.5.3 runs end to end but is not yet fully scheduled; once IT exposes the remaining preventive-plan tables and provisions the scheduled run, the dashboard would rebuild and issue its alerts each day without manual intervention.

The second is to extend the framework to the company's other plants. Because those sites run on the same maintenance system and the same underlying table structure, it is portable by construction, the data foundation it assumes being already in place elsewhere, and work to deploy it at further plants is under way. One framework serving several sites from the records each already keeps is itself an argument for the approach: the effort invested here is not confined to the site where it was developed, and each additional site both benefits from the method and, by the same recording discipline described in Section 5.4, contributes to improving the data on which it runs.

The recurring lesson across these directions is consistent with what the framework already shows: the largest remaining gains come not from a more elaborate model but from richer and better-connected data. The records the plant keeps have proved able to support a great deal; extending what they support is, above all, a matter of completing and connecting them.

Bibliography

AlGanem, H. S. and Abdallah, S. (2022). Exploring the hidden patterns in maintenance data to predict failures of heavy vehicles. In Al-Emran, M. and Shaalan, K., editors, *Recent Innovations in Artificial Intelligence and Smart Applications*, volume 1061 of *Studies in Computational Intelligence*, pages 171–187. Springer.

Allah Bukhsh, Z., Saeed, A., Stipanovic, I., and Doree, A. G. (2019). Predictive maintenance using tree-based classification techniques: A case of railway switches. *Transportation Research Part C: Emerging Technologies*, 101:35–54.

Amorim, L. D. A. F. and Cai, J. (2015). Modelling recurrent events: a tutorial for analysis in epidemiology. *International Journal of Epidemiology*, 44(1):324–333.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 10(4):1100–1120.

Bain, L. J. (1972). Inferences based on censored sampling from the Weibull or extreme-value distribution. *Technometrics*, 14(3):693–702.

Barlow, R. E. and Hunter, L. C. (1960). Optimum preventive maintenance policies. *Operations Research*, 8(1):90–100.

Bergmeir, C., Hyndman, R. J., and Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305.

Bradbury, S., Carpizo, B., Gentzel, M., Horah, D., and Thibert, J. (2018). Digitally enabled reliability: Beyond predictive maintenance. Technical report, McKinsey & Company. Available at `https://www.mckinsey.com/capabilities/operations/our-insights/digitally-enabled-reliability-beyond-predictive-maintenance`, accessed June 2026.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

Christer, A. H. and Waller, W. M. (1984). Delay time models of industrial inspection maintenance problems. *Journal of the Operational Research Society*, 35(5):401–406.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–220.

Crowder, M. J. (2001). *Classical Competing Risks*. Chapman & Hall/CRC, Boca Raton.

Dai, J. and Liu, X. (2023). Machine learning based prediction of rail transit signal failure: A case study in the United States. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 237(5):680–689.

Davidson-Pilon, C. (2019). lifelines: survival analysis in Python. *Journal of Open Source Software*, 4(40):1317.

Deloitte (2025). Asset optimization: Predictive maintenance. `https://www.deloitte.com/us/en/services/consulting/services/predictive-maintenance-and-the-smart-factory.html`. Accessed June 2026.

EN 13306 (2017). *Maintenance — Maintenance Terminology*. European Committee for Standardization (CEN), Brussels. European Standard EN 13306:2017.

Erbiyik, H. (2022). Definition of maintenance and maintenance types with due care on preventive maintenance. In *Maintenance Management — Current Challenges, New Developments, and Future Directions*. IntechOpen, London.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems 35, Datasets and Benchmarks Track*.

Guimarães, L. and Leitão, A. (2025). Maintenance management and failure data analysis. Course lecture notes, FEUP & INESC TEC.

IoT Analytics (2019). The top 10 industrial AI use cases. `https://iot-analytics.com/the-top-10-industrial-ai-use-cases/`. Accessed June 2026.

Jardine, A. K. S., Anderson, P. M., and Mann, D. S. (1987). Application of the Weibull proportional hazards model to aircraft and marine engine failure data. *Quality and Reliability Engineering International*, 3:77–82.

Jardine, A. K. S. and Tsang, A. H. C. (2013). *Maintenance, Replacement, and Reliability: Theory and Applications*. CRC Press, Boca Raton, 2nd edition.

Kumar, D. and Klefsjö, B. (1994). Proportional hazards model: a review. *Reliability Engineering & System Safety*, 44(2):177–188.

Kumar, D., Klefsjö, B., and Kumar, U. (1992). Reliability analysis of power transmission cables of electric mine loaders using the proportional hazards model. *Reliability Engineering & System Safety*, 37(3):217–222.

Lei, Y., Li, N., Guo, L., Li, N., Yan, T., and Lin, J. (2018). Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mechanical Systems and Signal Processing*, 104:799–834.

Louit, D. M., Pascual, R., and Jardine, A. K. S. (2009). A practical procedure for the selection of time-to-failure models based on the assessment of trends in maintenance data. *Reliability Engineering & System Safety*, 94(10):1618–1628.

Love, C. E. and Guo, R. (1991). Using proportional hazard modelling in plant maintenance. *Quality and Reliability Engineering International*, 7(1):7–17.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774.

Mobley, R. K. (2002). *An Introduction to Predictive Maintenance*. Butterworth-Heinemann, Amsterdam, 2 edition.

Molęda, M., Małysiak-Mrozek, B., Ding, W., Sunderam, V., and Mrozek, D. (2023). From corrective to predictive maintenance—a review of maintenance approaches for the power industry. *Sensors*, 23(13):5970.

Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 625–632.

Ochella, S., Shafiee, M., and Sansom, C. (2021). Adopting machine learning and condition monitoring P-F curves in determining and prioritizing high-value assets for life extension. *Expert Systems with Applications*, 176:114897.

Ozturk, S., Fthenakis, V., and Faulstich, S. (2018). Assessing the factors impacting on the reliability of wind turbines via survival analysis – a case study. *Energies*, 11(11):3034.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., and Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12):1373–1379.

Prentice, R. L., Williams, B. J., and Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika*, 68(2):373–379.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems 31*, pages 6639–6649.

Riley, R. D., Snell, K. I. E., Ensor, J., Burke, D. L., Harrell, F. E., Moons, K. G. M., and Collins, G. S. (2019). Minimum sample size for developing a multivariable prediction model: PART II – binary and time-to-event outcomes. *Statistics in Medicine*, 38(7):1276–1296.

Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432.

Salinas-Camus, M., Goebel, K., and Eleftheroglou, N. (2025). A comprehensive review and evaluation framework for data-driven prognostics: uncertainty, robustness, interpretability, and feasibility. *Mechanical Systems and Signal Processing*, 237:113015.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239–241.

Snider, B. and McBean, E. A. (2020). Watermain breaks and data: the intricate relationship between data availability and accuracy of predictions. *Urban Water Journal*, 17(2):163–176.

Susto, G. A., Schirru, A., Pampuri, S., McLoone, S., and Beghi, A. (2015). Machine learning for predictive maintenance: A multiple classifier approach. *IEEE Transactions on Industrial Informatics*, 11(3):812–820.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer, New York.

Tiddens, W., Braaksma, J., and Tinga, T. (2022). Exploring predictive maintenance applications in industry. *Journal of Quality in Maintenance Engineering*, 28(1):68–85.

Vittinghoff, E. and McCulloch, C. E. (2007). Relaxing the rule of ten events per variable in logistic and Cox regression. *American Journal of Epidemiology*, 165(6):710–718.

Walker, A. M. (1996). Confounding by indication. *Epidemiology*, 7(4):335–336.

Wang, W. (2012). An overview of the recent advances in delay-time-based maintenance modelling. *Reliability Engineering & System Safety*, 106:165–178.

Appendix A

Maximo Extracts

This appendix summarises the fields extracted from the Maximo CMMS and used throughout the analysis. Table A.1 lists the retained fields and their meaning.

Table A.1: Fields extracted from Maximo (CMMS).

Entity	Field	Meaning
Location	`location`	Six-level functional-location code locating the asset in the plant (key)
	`zloccrit`	Criticality of the location: supercritical, critical, or unclassified
	`assetnum`	Asset currently installed at the location
	`assetdate`	Date the current asset entered the location (date of the most recent asset change)
Work orders	`wonum`	Work-order identifier (key, shared by corrective and preventive orders)
	`location`	Functional location of the order
	`description`	Free-text account of the order
	`status`	Order status, including scheduled-but-unexecuted
	`classificationid`	Classification: for corrective work, unplanned or planned; for preventive work, the plan class (inspection, cleaning–lubrication, or compliance variants)
	`wopriority`	Priority: urgency levels (immediate to one month) or shutdown horizon (next, monthly, quarterly, yearly)
	`persongroup`	Maintenance team (mechanical, electrical, lubrication, predictive)
	`reportdate`	Date the order was reported
	`schedfinish`	Scheduled completion date
	`actstart`	Date work actually began
	`actfinish`	Actual completion date
Corrective (CM)	`wonum`	Work order the corrective record belongs to (key)
	`origrecordclass`	Type of the originating record (service request or work order)
	`origrecordid`	Identifier of the originating record
Preventive (PM)	`wonum`	Work order the preventive record belongs to (key)
	`pmnum`	Preventive plan the order belongs to
	`frequency`	Defined periodicity of the plan (value)
	`frequnit`	Unit of the defined periodicity
Condition evaluations	`wonum`	Corrective or preventive order the assessment belongs to
	`metername`	Component assessed, by code and name
	`measurementvalue`	Ordinal condition grade, 1 (total failure) to 7 (no issue), on a component-specific scale
	`observation`	Description attached to the grade
	`measuredate`	Date of the assessment

Appendix B

Failure Prediction Model Diagnostics

This appendix reproduces two diagnostics of the failure prediction model of Chapter 4.

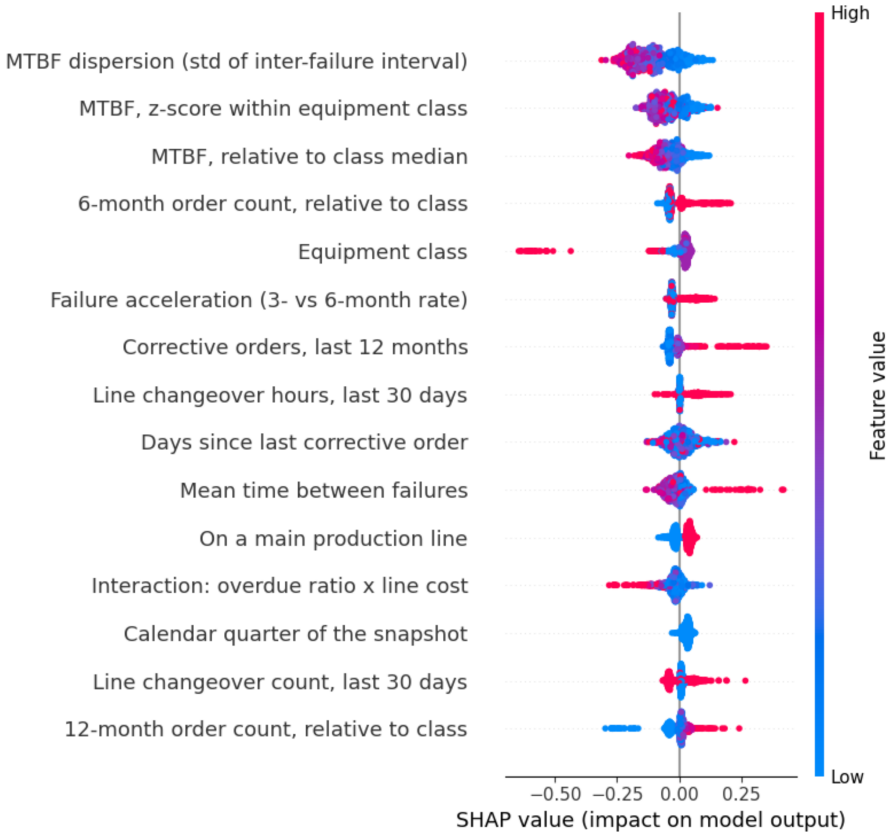

Figure B.1: SHAP summary of the failure prediction model.

The summary in Figure B.1 ranks the fifteen features of highest mean absolute contribution. Each point is one location; its horizontal position is the feature's contribution to the predicted score, and its colour the feature value, red high and blue low. The dispersion of the inter-failure interval leads, followed by the class-relative failure-history features, with no single feature dominating the decision.

The diagnostics of Figure B.2 read in two parts. On the left, the calibrated probabilities track the observed frequencies along the diagonal, sitting slightly below it, so the model is faithful

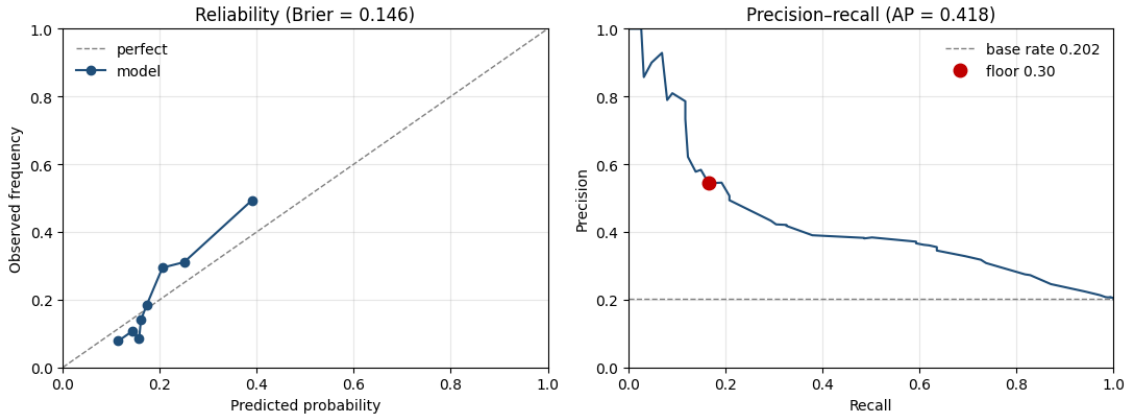

Figure B.2: Calibration and precision-recall of the deployed model on the out-of-time test.

and mildly conservative, at a Brier score of 0.146. On the right, the precision-recall curve runs well above the base rate of 0.202 across its range, and the deployed operating floor ($p \geq 0.30$) is marked: it sits at the knee of the curve, where precision is still high before coverage rises and precision falls away.

Appendix C

Hyperparameter Grid and Selected Configurations

This appendix records the hyperparameter search behind the algorithm comparison of Section 4.2.2.1. For each of the three ensembles, thirty configurations were drawn at random from the grid of Table C.1 and scored by the blocked time-series cross-validation of Section 4.2.1.3, ranked by average precision. Random sampling rather than exhaustive enumeration follows Bergstra and Bengio (2012). The right-hand column reports the configuration selected for each model. The deployed model is CatBoost in its library-default configuration, for the reasons given in Section 4.2.2.1; the selected CatBoost column is shown for completeness, as the tuned variant that was evaluated but not deployed.

Table C.1: Hyperparameter grid searched and the configuration selected for each model.

Model	Hyperparameter	Grid values	Selected
Random forest	`n_estimators`	300, 600, 1000	1000
	`max_depth`	none, 10, 20, 30	10
	`min_samples_leaf`	1, 2, 5, 10	5
	`max_features`	sqrt, log2, 0.3	log2
	`class_weight`	none, balanced	none
XGBoost	`n_estimators`	300, 600, 1000	300
	`max_depth`	3, 5, 7, 9	7
	`learning_rate`	0.01, 0.03, 0.05, 0.1	0.01
	`subsample`	0.7, 0.85, 1.0	0.85
	`colsample_bytree`	0.7, 0.85, 1.0	0.85
	`min_child_weight`	1, 3, 5	5
	`reg_lambda`	1, 3, 5	5
	`scale_pos_weight`	1, 3, 5	1
CatBoost	`iterations`	300, 600, 1000	600
	`depth`	4, 6, 8, 10	4
	`learning_rate`	0.01, 0.03, 0.05, 0.1	0.01
	`l2_leaf_reg`	1, 3, 5, 7, 9	5
	`auto_class_weights`	none, balanced	balanced

Appendix D

Critical Values for the Weibull Shape Parameter

This appendix reproduces the confidence limits for the Weibull shape parameter from which the critical values F_β of the eligibility gate of the interval-optimisation system are read (Section 4.4.1).

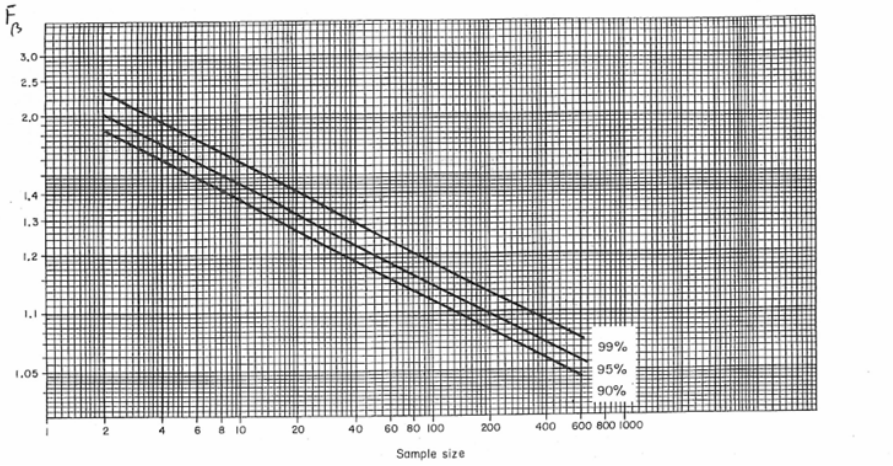

Figure D.1: Confidence limits for the shape parameter β for different confidence values (Bain, 1972).

The chart in Figure D.1 gives the confidence limits F_β for the shape parameter as a function of sample size, at the 90, 95, and 99 per cent confidence levels. The lower bound of Equation 2.9 divides the estimated shape parameter by the critical value F_β read at the sample size of the fit, and wear-out is declared only when that bound exceeds one; the analysis adopts the 90 per cent one-sided level, for the reasons given in Section 4.4.1.

Appendix E

Dashboard Views

This appendix reproduces the three views of the decision-support dashboard described in Chapter 4: the failure-prediction view together with its component-history chart, the plan-adequacy view together with the per-equipment timeline for one location of each flagged category, and the interval-optimisation view.

Figure E.1: Failure prediction view of the dashboard.

The view in Figure E.1 is a live snapshot taken on 14 June 2026, later than the evaluation cut-off, so its tier counts differ from those of Table 5.3. The risk tiers are shown in Portuguese, the plant's working language, and map to the labels of Section 4.2.2.4 as follows: *Monitorizar – Risco Médio* is Monitor – Medium risk; *Intervir – Risco Alto* is Act – Medium-high risk; and *Intervir – Risco Elevado* is Act – High risk.

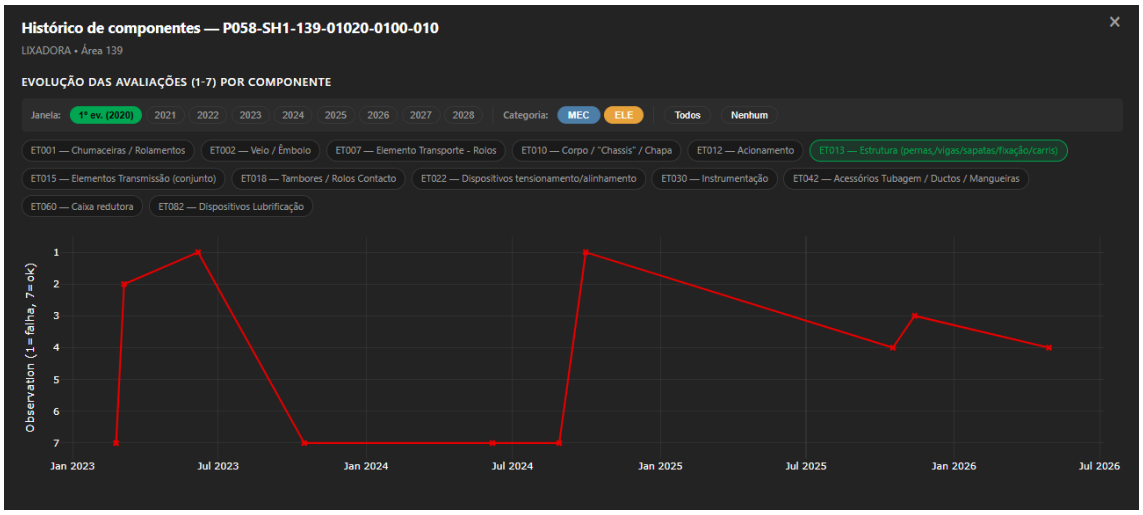

Figure E.2: Component-history chart of the failure-prediction view.

The chart of Figure E.2 plots, for one equipment item, each component's condition grade over time on the 1–7 scale of Table 4.6, so the engineer can read by eye which components are degrading and which are closest to failure.

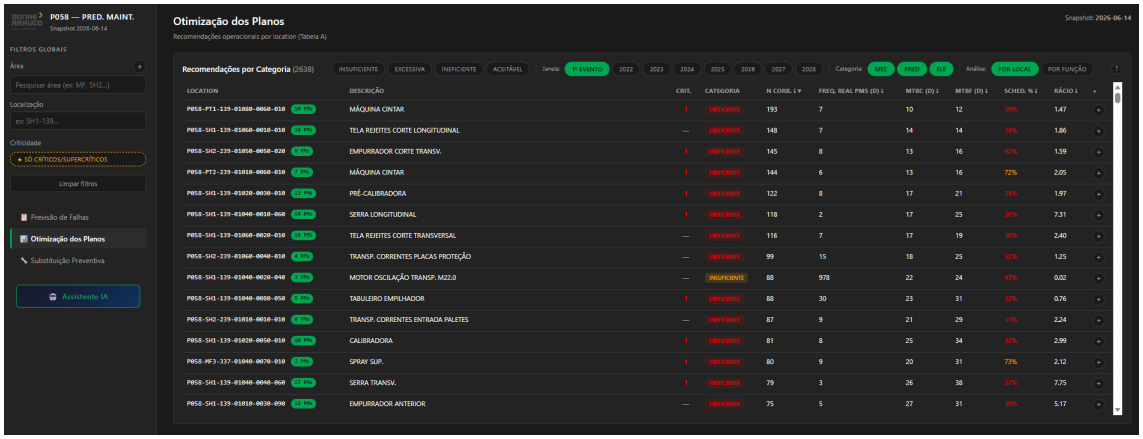

Figure E.3: Plan-adequacy view of the dashboard.

The view in Figure E.3 is the general table of the plan-adequacy assessment, taken from the same 14 June 2026 snapshot as Figure E.1. Each location is listed with its corrective count, realised preventive frequency, MTBC, schedule compliance, and adequacy ratio, and carries its assigned category. The categories are shown in Portuguese, the plant's working language: *Insuficiente* is under-maintained, *Excessiva* over-maintained, *Ineficiente* ineffective, and *Aceitável* acceptable. Under the default thresholds the view surfaces 2,638 recommendations, the sum of the under-maintained, over-maintained, and ineffective locations of Section 5.2.1; expanding a row opens the per-equipment detail shown in the figures that follow.

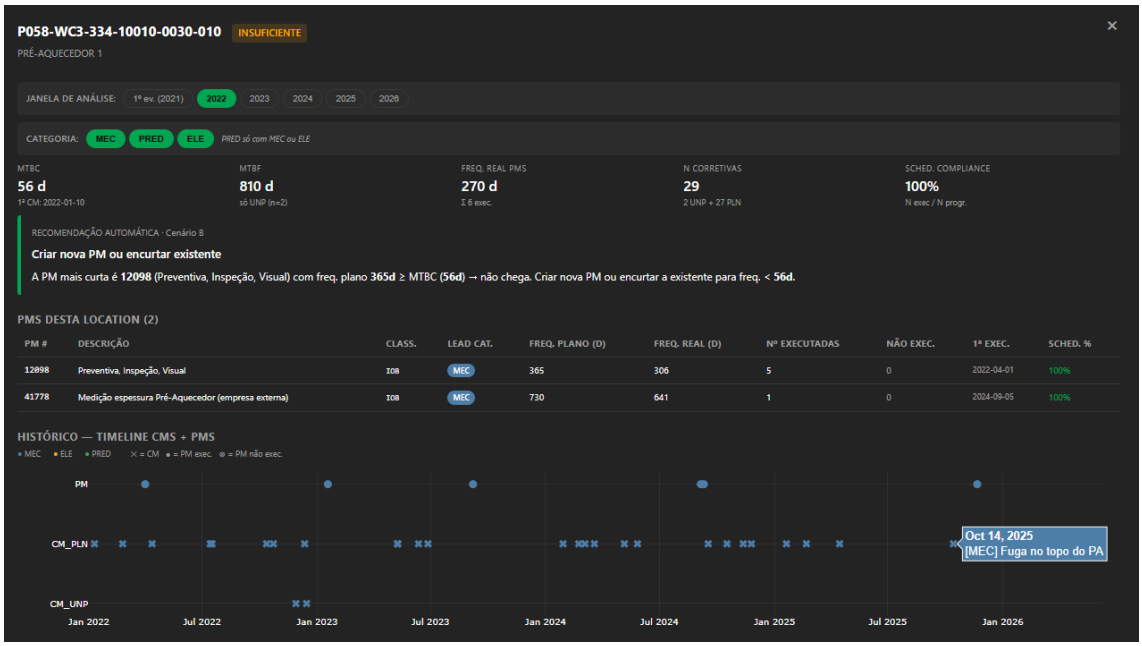

Figure E.4: Plan-adequacy view of an under-maintained location.

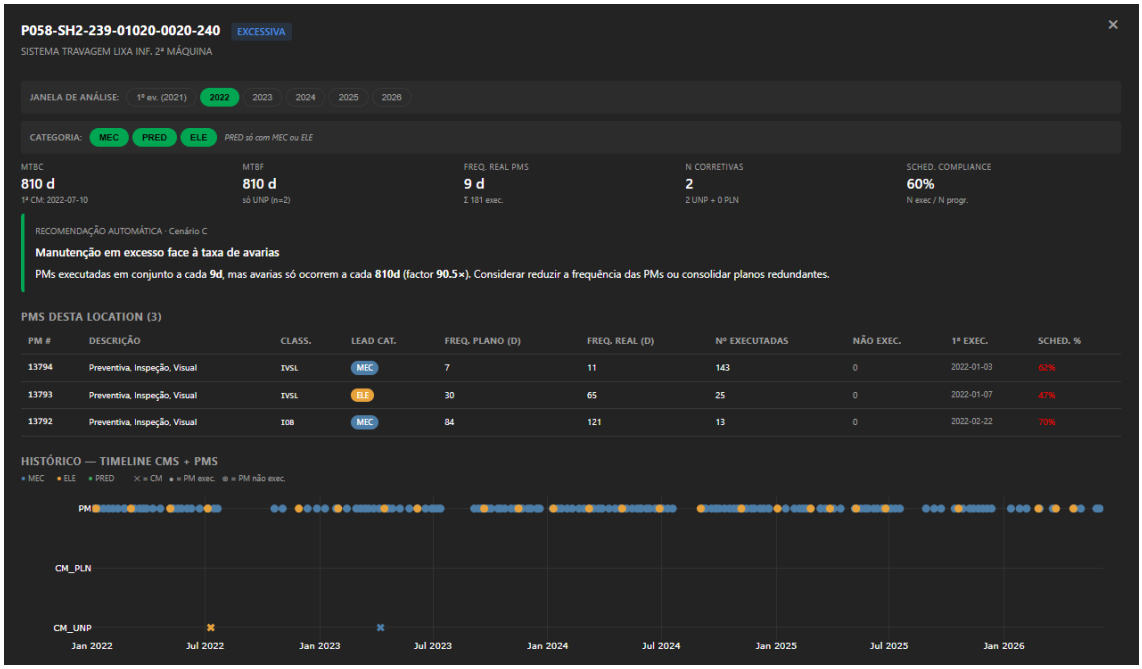

Figure E.5: Plan-adequacy view of an over-maintained location.

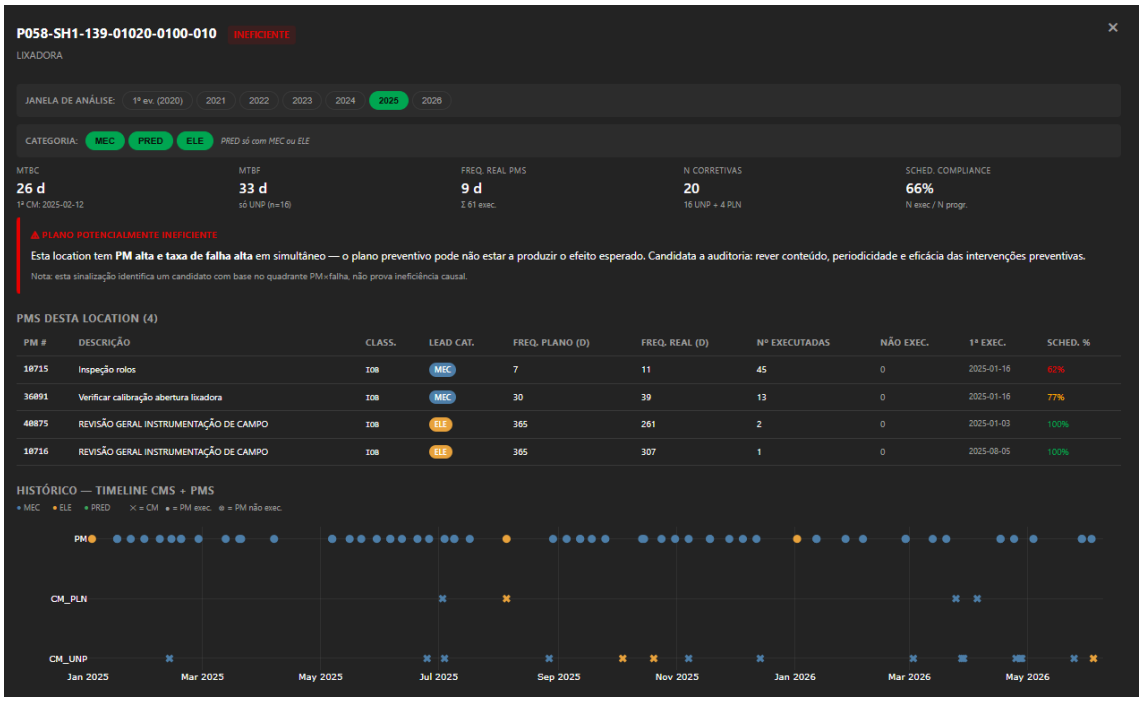

Figure E.6: Plan-adequacy view of a location flagged with an ineffective plan.

The three views of Figures E.4 to E.6 show the per-equipment timeline for one location of each category: failures accumulating against sparse prevention in the under-maintained case, dense prevention against a near-empty failure timeline in the over-maintained case, and failures persisting through frequent prevention in the ineffective one.

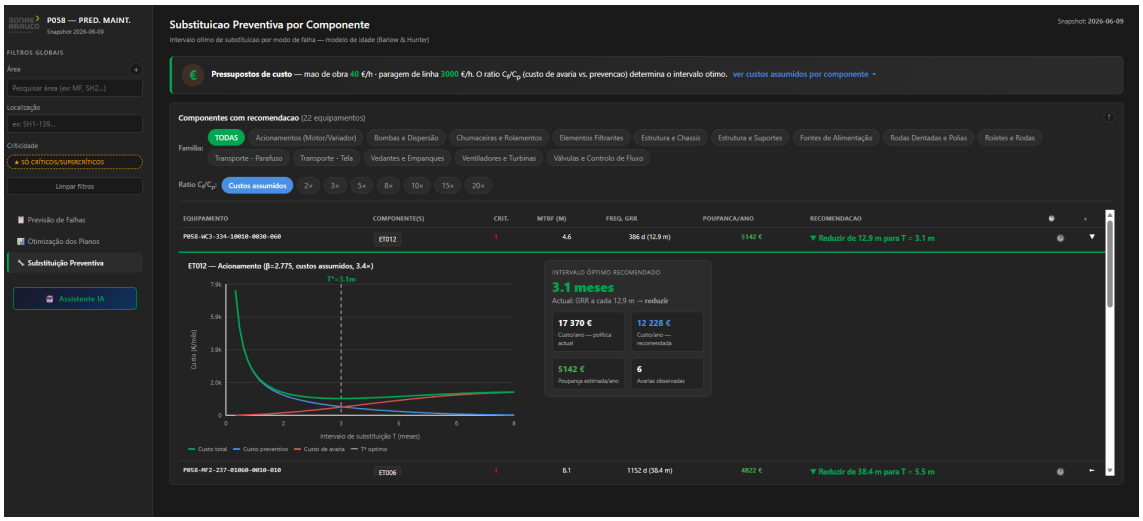

Figure E.7: Interval optimisation view of the dashboard.

The view in Figure E.7 shows, for one component that clears the eligibility gate, the fitted Weibull parameters, the recommended replacement interval, and the cost-per-unit-time curve, with the cost ratio selectable so the engineer sees how the recommendation responds to it.