MASTER

Development of a condition based maintenance decision model by data mining

Çakir, G.S.

Award date:
2011

Eindhoven, June 2011

Development of a Condition Based Maintenance Decision Model by Data Mining

by

Goknur Seyma CAKIR

BSc. Mechanical Engineering-Middle East Technical University, 2008

Student identity number 0728439

in partial fulfilment of the requirements for the degree of

Master of Science

in Operations Management and Logistics

Supervisors:

Prof. dr. ir. G.J.J.A.N van Houtum, TU/e, OPAC

dr.ir. A.J.M.M Weijters , TU/e, IS

ir. G. Streutker, ASML, Customer Support- Operational Services

TUE. School of Industrial Engineering.

Series Master Theses Operations Management and Logistics

Subject headings: condition based maintenance, predictive maintenance, data mining, failure prediction, remaining useful life, neural network

Abstract

This master thesis describes a research project conducted within the Customer Service Operational Services department at ASML. A decision support model is developed for condition based predictive maintenance of a critical machine component by implementing data mining methods. Several data mining techniques are used to predict the upcoming failures and they are compared in terms of their prediction accuracy in order to find out the best model. Furthermore the proposed model is compared with the physical model which has already been developed by ASML by using the system knowledge. The thesis concludes with a discussion on main findings, limitations and possible future extensions.

Preface

This master thesis document presents the result of my graduation project for the Master of Science Program in Operations Management and Logistics at Eindhoven University of Technology. This project was carried out from January 2011 to June 2011 at ASML Netherlands B.V.

I would like to use this opportunity to express my gratitude to all the people who have supported me throughout my project.

First of all I would like to thank my first supervisor from TU/e, Prof. Geert-Jan van Houtum, for his support, helpful comments and advices. His suggestions were very valuable for me and assisted me throughout the project. Furthermore I would like to thank Ton Weijters, my second supervisor from TU/e, for his critical and constructive comments on my project.

Within ASML, firstly I would like to thank Gert Streutker, my supervisor for giving me the opportunity to work on this project in their company. His feedback, extensive knowledge and experience enabled me to include many aspects in the project. Besides, I would like to thank all the members of the Customer Support-Operational Services Department at ASML for sharing their knowledge and experience with me.

I would like to thank my friends for their support during my study. You made sure that I enjoyed this time!

I owe many thanks to Semih Yildiz, my dear fiancée. I am so lucky that you are in my life. Thanks a lot for being with me through all the good and bad moments and for supporting me all the time.

Last but not the least; my parent and my brother deserve the biggest thanks: Thanks a lot for your confidence and encouragement throughout my life.

G. Seyma CAKIR

Eindhoven, June 2011

Executive Summary

To preserve its market share and to satisfy its customers, ASML provides high quality customized support services with technology. Maintenance support service is an essential service provided with the technology. ASML implements periodic and corrective maintenance according to the customer demand. However an innovative maintenance policy, Condition Based Maintenance, commits to increase availability and to reduce scheduled and unscheduled downtime by predicting the failure time. Therefore in order to remain the best, ASML is committed to provide best service and it focuses on the improvement of the predictive tools to implement condition based maintenance.

This master thesis is a study of an exemplary implementation of condition based maintenance policy in ASML. The research assignment has been defined as “to develop a data driven decision support model which alerts the user before failures occur, and indicates the remaining useful life of the critical component by using condition-based data”.

In order to accomplish the assignment, the following research questions have been formulated and answered during this research.

1. How can a condition based maintenance decision support model be designed technically?
    1.1. How to perform the Data Acquisition step?
    1.2. How to perform the Data Processing step?
    1.3. How to perform the Maintenance Decision Making step?
2. What is the difference between the proposed model and the physical model that has already been developed by ASML?

Historical data which include the condition and event data (failure cases) were acquired from the global database. It was aimed to find the relation between condition data and failure cases, and to find a method to predict upcoming failures based on that relation.

Data understanding and data preparation are significant steps. Since data includes errors and missing cases, it is required to obtain the qualified representative data. As a result of the data preparation step, model inputs were defined. Although success of the local monitoring data is indisputable, more complete and accurate data is required to develop a better model.

For the prediction model, three distinct approaches have been investigated. Firstly, several machine learning techniques were used to classify samples in three nominal groups which show the component remaining life time as an interval. As a second approach, machine learning techniques were used to predict the remaining time of the component in days. As a third approach, a mathematical model was developed to explain the relation between condition parameters and failure cases.

The given failure cases imply not only the component being out of service, but also the decreased performance of the component which could have a customer (i.e. process related) dependent impact on wafer quality. The customer expectation on the machine performance depends on machine type, exigent circumstances etc. So, the threshold level of failure depends on many external factors. Thus, the third approach was enhanced with the given threshold level. It provides a significant improvement in the failure prediction. As a result a final model which predicts 82% of the upcoming failures was proposed for ASML use. Therefore prediction of the upcoming failure can contribute to eliminate over-maintenance and decrease unscheduled down time. Moreover since the model diagnoses all faulty sub-modules, it enables a service engineer to specify the scope of the maintenance. This leads to a decrease in maintenance expenditures. The remaining useful life (RUL) indication helps the service engineer decide about when to plan maintenance, when to arrange labor and when to order spare parts cost-effectively.

As a next step, this proposed model was compared with the physical model which has already been developed by the domain experts in ASML. Physical models predict the upcoming machine failure by using physical theories whereas data driven models predict failures according to the relation between given inputs (condition parameters) and outputs (failure events). The prediction accuracy of these models was assessed for 11 failure cases. Consequently, the data driven model which predicts 82% of the failure cases outperforms the physical model which does not produce any warning signal for 55% of the upcoming failures. Thus, the success and feasibility of the data driven model proved the predictability of failures without using the system knowledge.

Last but not least, in the deployment phase, the decision support model has been built in order to integrate the outputs of the prediction model with the maintenance activities. Moreover, in the deployment phase, the user interface has been developed to support ASML to use the output of the proposed prediction model. A user friendly interface has been built in MS Excel. For entered condition parameters and threshold values, the status of the module and the remaining useful life of the module are shown. Therefore the field service engineer is informed about the upcoming failure.

Table of Contents

List of Figures

List of Tables

Abbreviations

BPS	Business Problem Solving
CART	Classification and Regression Trees
CBM	Condition Based Maintenance
CRISP-DM	Cross Industry Standard Process for Data Mining
CS	Customer Support
DDM	Data Driven Model
FN	False Negatives
FP	False Positives
LHM	Labor hour per machine
LL	Lower Limit
LR	Learning Rate
M	Momentum
MLP	Multi Layer Perceptron
N/A	Not available
NHL	Number of Neuron per Hidden Layer
NN	Neural Network
OS	Operational Service
PM	Physical Model
ProSelo	Proactive Maintenance and Service Logistics for Advanced Capital Goods
RBF	Radial Basis Function
RUL	Remaining Useful Life
SD	Scheduled Down
TL	Threshold Level
TN	True Negatives
TP	True Positives
UL	Upper Limit
USD	Unscheduled Down
XLD	Extreme long down
WW	World Wide

The major difference between a thing that might go wrong and a thing that cannot possibly go wrong is that when a thing that cannot possibly go wrong goes wrong it usually turns out to be impossible to get at or repair.

Douglas Adams

Introduction

The maintenance concept for capital goods has gained more importance as availability and reliability has become a significant issue for manufacturing companies and service organizations. Among maintenance policies, Condition Based Maintenance has become prominent by supporting **right-on-time maintenance based on tangible reasons.** Condition Based Maintenance is a developed proactive maintenance strategy which increases availability of capital goods while eliminating over-maintenance cost.

Aiming to increase availability and reduce scheduled and unscheduled downtime, ASML started to develop and use predictive tools. Local offices initiated condition monitoring and these initiatives resulted in local improvements on down time. ASML's objective is to develop a sustainable solution by bringing the locally developed monitoring tooling knowledge into sustainable toolset of means and methods.

This master thesis aims to develop data driven decision support model **which alerts the user before the failure occurs, and indicates the remaining useful life (RUL) of capital goods by using condition-based data.**

Chapter 1 provides information about the research setting, ASML, and gives background information about maintenance policies and condition based maintenance. Then the research assignment is explained in detail. This report is organized in 8 chapters and the report outline is presented at the end of the first chapter.

1. Company Description and Research Assignment

This chapter consists of three sections. In the first section, we present brief introduction about the company ASML and the Customer Support Operational Services Department which constitutes the research settings of this master thesis study. In the second section, background information about the maintenance policies and condition based maintenance policy is provided. Finally, the design of the research assignment is explained.

1.1. Company Description

ASML is the world's leading provider of lithography systems for semiconductor industry. It designs, develops, integrates, markets and services advanced systems used by the semiconductor industry to manufacture complex integrated circuits (ICs or chips). ASML's customers include most of the world's major chip manufacturers such as Intel, Toshiba, Samsung, Texas Instrument, IBM, Micron and TSMC.

In the semiconductor industry, technology is provided with support services. An integrated customer solution is a key for semiconductor manufacturers to remain competitive. To preserve its market share and to satisfy the customers, ASML provides high quality customized support services with technology. Every fab is different and requires a different support coverage package. Therefore ASML's service contract portfolio is designed to be flexible to meet any Customer's need. ASML offers an extensive portfolio of Labor, Applications, Parts and Parts Inventory Management contracts. The contract form depends on the number and type of systems in a fab. Moreover ASML also offers equipment relocation, fab start-up, training and advanced application notes.

Support Packages

ASML offers support packages which may consist of a part contract, a labor contract, an application contract and a logistic contract.

Labor Contracts

ASML's Field Service Engineers are armed with the most up-to-date technical information to assure the highest levels of system performance. ASML's fully qualified technical experts facilitate fast troubleshooting and repair, minimizing downtime and securing maximum performance of your systems.

Applications Contract

ASML offers application support contract that can be customized to specific customer requirements. It aims to optimize process efficiency.

Parts Contracts

In addition to labor contract, ASML also offers Parts Contracts per machine or for a Fab. Owing to this contract, fixed, yearly fee for all relevant spare parts (excluding consumables) will be made available with a guaranteed service level. Planning, shipment, customs clearance and installation of spares are cared by ASML. Yearly expenditures can be budgeted in advance.

Logistics Service Contract

ASML also supports parts inventory management. Logistic service contracts can be designed according to customer specifications in terms of:

- The required service level
- Guaranteed availability of spare parts whenever they are needed
- Minimum unexpected downtime and related costs
- Contract price depends on the agreed service level

1.1.1. Organization of ASML

There are 4 main divisions under the ASML organization, namely: Support, Product, Market and Operations. Figure 1 shows the organizational chart.

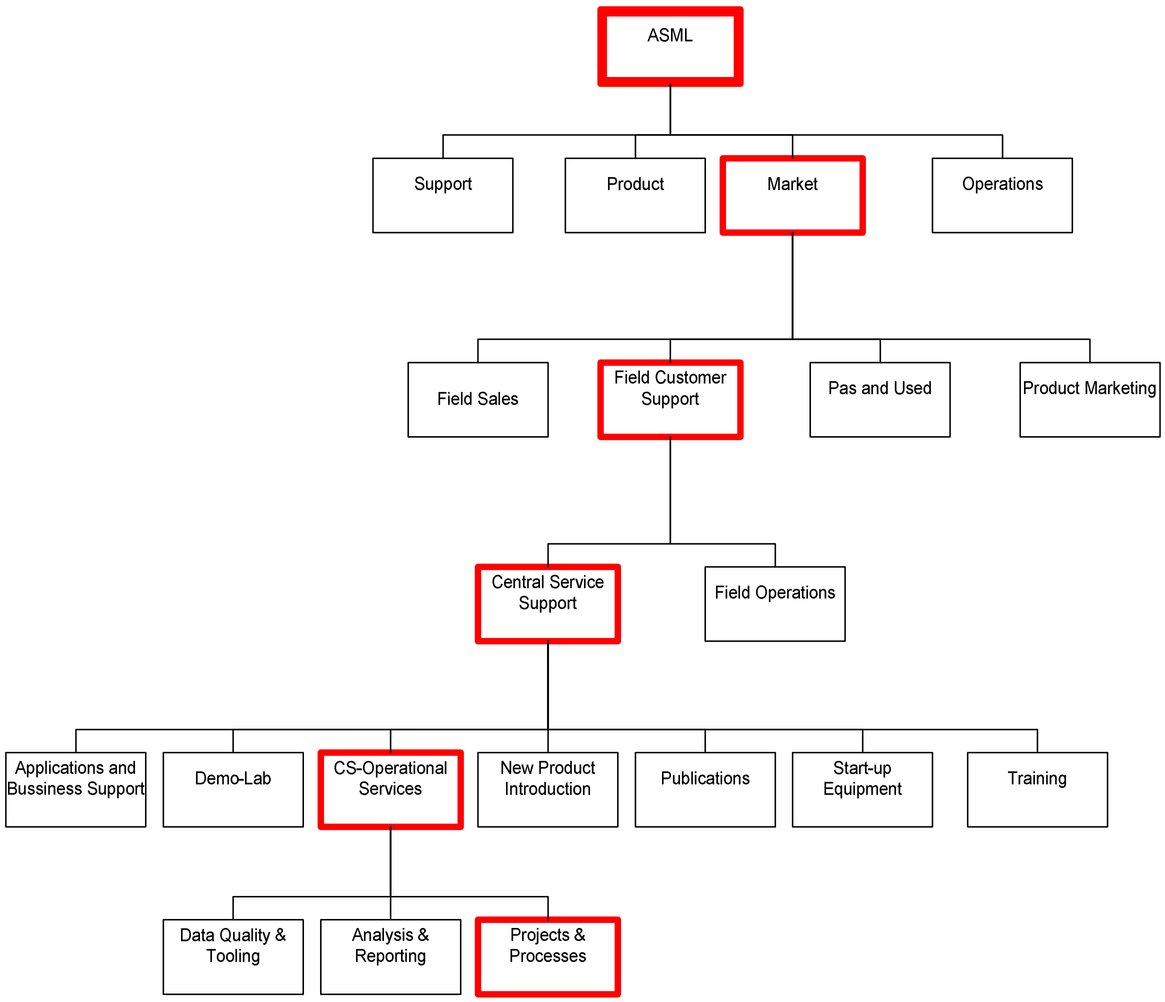


Figure 1: ASML Organizational Chart

The master thesis was carried out with the collaboration of the Customer Support and Operational Services department under Market division of ASML. This department supports the whole market (all customers), in terms of field operations by providing generic solutions. These solutions are implemented by the field operations department. The following part describes the Customer Support and Operational Services department.

1.1.2. Customer Support - Operational Services Department

Customer Support Department aims to provide operational excellence by reducing service cost while improving product performance. Escalation Management, System Performance Management, Maintenance Planning, Service Execution are Customer Support (CS) processes.

Customer Support-Operational Services (CS-OS) supports customers by means of ASML’s field service engineers in terms of (1) data and analysis, (2) tooling and automation, (3) continuous improvement of services and support and (4) standard and reliable ways of working. This department’s main responsibilities are maintenance engineering, business process development and equipment performance monitoring. It consists of three teams: Analysis & Reporting, Data Quality & Tooling and Projects & Processes. Analysis & Reporting team is responsible for providing on time, accurate, complete analysis and reports regularly. Data Quality & Tooling team provides automation tools that meet customer requirements. Projects & Processes team initiates and manages aligned, effective and efficient projects & processes to support customer.

1.2. Background about Maintenance Policies and Condition Based Maintenance Policy

No matter how good capital goods are designed, to keep them operating at desired reliability level, maintenance is required. Tsang et al. (1999) define maintenance as to repair broken items. However as opposed to this traditional perception, maintenance concept has been evolved throughout the years and distinct definitions have been given for maintenance. According to British Standards (1984); maintenance is defined as the combination of all technical and associated administrative actions intended to retain an item or system in, or restore it to, a state in which it can perform its required function.

Zhao et al. (2010) state that the annual cost of maintenance goes up to 15% for manufacturing companies, 20%–30% for chemical industries, and 40% for iron and steel industries. Therefore, importance of maintenance increases significantly and there is a continuous search for a better maintenance policy which provides economic efficiency with higher system reliability, availability and safety.

Under these circumstances maintenance applications have changed from corrective maintenance to proactive maintenance. Whereas users had performed maintenance after failure occurrence, nowadays they try to eliminate failure by performing proactive maintenance. In other words they are moving from reactive to a proactive maintenance policy. One of such proactive maintenance policies is condition based maintenance which aims to predict failure through condition monitoring

Maintenance Policies

In the literature different classifications and denomination exist for maintenance techniques. By taking the definition of maintenance into account, maintenance policies are figured out in two categories in this study:

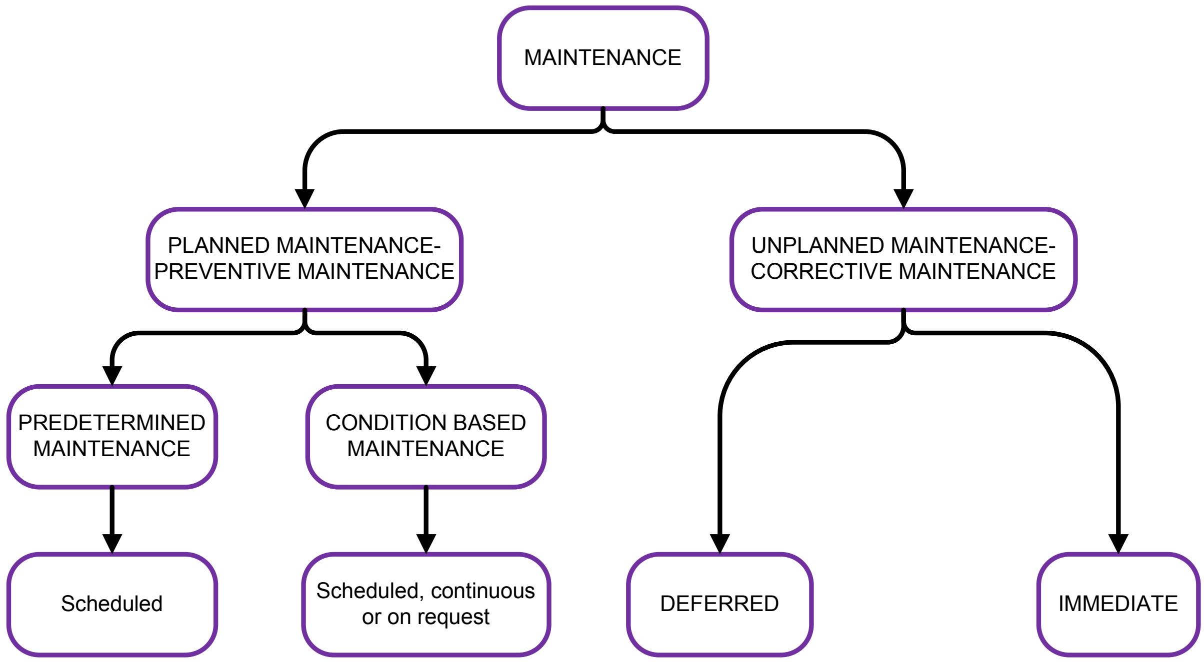


Figure 2: Maintenance Techniques (Niu et al. 2010)

(1)Planned maintenance which aims to **retain** the capital goods **in**, to prevent failures

(2)Unplanned maintenance which aims to **restore** the capital goods after failure

Figure 2 shows maintenance techniques. Three most common maintenance techniques are corrective maintenance, predetermined (so called preventive) maintenance and condition based maintenance.

Corrective Maintenance: It is also known as breakdown maintenance or unplanned maintenance, or run-to-failure maintenance. Corrective Maintenance is the earliest and simplest maintenance technique. Maintenance is performed when the failure happens. Therefore it is formed of unplanned activities and crisis management is required when the machine fails. The reason of the failure is diagnosed first and then maintenance is performed. It has high spare part and repair costs. Safety hazard is high because emergency situation is not detected and breakdown is waited to perform maintenance. On the other hand, corrective maintenance eliminates over maintenance and related costs. There is no difference between implementation of immediate and deferred maintenance except timing.

If unscheduled failure maintenance cost is not higher than preventive maintenance cost and safety and uptime are not critical issues, the usage of corrective maintenance could be the most economic way owing to usage of full life time of the component/machine. It could be useful for simple non-integrated machines if the failure is easily and cheaply repairable and it doesn't cause any other failure.

Predetermined Maintenance: It is also called as periodic, preventive or planned maintenance. The condition of a machine is not taken into account and machine age is the only criteria to execute maintenance. Maintenance is performed periodically to decrease unexpected failures; however it is not possible to eliminate all random failures. The maintenance activities could be managed and the amount of required labor and spare parts are determined earlier. Unscheduled breakdown and so down time are reduced. Although this approach reduces failure risk and down time, costs related to over-maintenance and spare parts increases.

Condition Based Maintenance: CBM is the developed preventive maintenance technique which is based on machine condition. Maintenance is performed when it is required by observing the condition of the physical asset. CBM aims to improve system reliability, availability and security and to reduce maintenance cost. This technique has significant advantages over conventional techniques. Firstly, induced failure, spare parts, downtime and production interference are reduced. System availability is increased by CBM. Secondly management and logistic activities are controlled. Labor planning, maintenance planning spare parts planning can be conducted effectively by observing machine condition. One of the greatest advantages is the extended equipment life which causes reduction in life cycle cost. Since machines condition is observed continuously or periodically, machines can be stopped in critical situations and it provides higher safety. On the other hand the implementation of this technique is complex and costly. It requires additional skills and higher investment in comparison to the other two techniques. Capital investment includes cost of experiment tests, R&D expenses, and system development cost due to new IT infrastructure, hardware, software, system integration.

Selection of the appropriate maintenance policy is based on the main concern of the user. The significance of availability, cost and safety issues may lead to implementation of different maintenance techniques. If the system is cheap, easily repairable and failure doesn't cause any serious problem, corrective maintenance could be the effective way. However if failure is avoided due to the mentioned issues, preventive maintenance or CBM could be a better alternative. Availability of condition monitoring system and skilled labor directs to the condition based maintenance option which provides higher uptime, reduced cost and higher safety. However if required infrastructure is not available, a user should trade off between investing in a CBM system, and paying for over maintenance and unscheduled breakdown.

The advantages and disadvantages of the maintenance techniques are summarized in Table 1.

Table 1: Comparison of Maintenance Techniques

	Advantages	Disadvantages
Corrective Maintenance	No over-maintenance (low cost policy) No condition related cost Requires minimal management Useful on small non-integrated plant	High production downtime Large spare inventory High cost repairs Crisis management needed Over time labor Safety hazardous
Predetermined Maintenance	Enabled management control Reduced down time Control over spare parts and costs Reduced unexpected failure Fewer catastrophic failure	Over-maintenance Unscheduled breakdown
Condition Based Maintenance	Reduced unplanned downtime, spares, induced failures Reduced production interference Enabled management and logistic control Extended equipment life Reduced life cycle cost and maintenance expenditures	Higher investment cost Additional skills are required

CBM Methodology

Three main steps in CBM should be followed in order to design a CBM decision model: (1) Data Acquisition, (2) Data Processing, (3) and Maintenance Decision Making

a. Data Acquisition

Data Acquisition is the first step which includes collecting and storing information from the capital goods. Two types of data namely condition and event data are recorded to use diagnostics and prognostics. Condition data indicates the state and health condition of capital goods whereas event data depicts the cases and taken actions.

Condition data is obtained by means of condition monitoring. Condition monitoring has been defined as “The assessment on a continuous or periodic basis of the mechanical and electrical condition of machinery, equipment and systems from the observation and/or recordings of selected measurement parameters” (Collacott 1997).

b. Data Processing

Data Processing includes the data cleaning and data analysis steps.

- **Data Cleaning**

Obtaining high quality data is the first crucial step to generate a strong CBM decision model. Data cleaning which includes detecting and correcting inaccurate data is required to enhance the data quality. Statistical tools such as Descriptive Statistics, Histograms, Scatter plot could be helpful to detect errors.

- **Data Analysis**

Which data analysis is performed depends on the data type. According to Jardine et al. (2006) data are collected in three different categories.

  - Waveform type: Data collected in the form of time series at a specific time period
  - Value type: Single value data collected at a specific time period
  - Multidimensional type: Multidimensional data collected at a specific time period

Data analysis can be performed either for only event data or for combination of event and condition data. The first type of analysis, known as reliability analysis, is to select best fitting survival distribution based on event data. The fitted distribution is used for further analysis. Secondly, in order to better understand and interpret data, combination of event and condition data is analyzed by building mathematical model. This mathematical model is the basis for maintenance decision support model (Jardine et al, 2006).

c. Maintenance Decision Making

Diagnostics and prognostics are two significant aspects of CBM decision making step. Although the aim of CBM model to do prognostics, diagnostic is required when prognostics fails to predict and fault occurs (Vismara, 2010). Peng et al. (2010) define diagnostics as dealing with fault detection, isolation, and identification when abnormity occurs and define prognostics as dealing with fault and degradation prediction before they occur.

Diagnostics analyze the system performance, degradation level and health states. Firstly the abnormal operating condition is discovered (Fault detection). Then the faulty component or subsystem is detected (Fault isolation). Finally the nature and extend of fault/failure is evaluated (Fault identification).

Prognostics refer to the capability to provide early detection of the fault condition of a component, and to predict the progression of this fault condition to component failure (Gilmartin et al., 2000). In other words failure occurrence time is estimated. Precise and reliable prognostic is critical for CBM in order to improve safety, schedule maintenance, reduce maintenance cost and increase availability.

According to Jardine et al. (2006), there are two main prediction types in machine prognostics. One of them, common type, is the prediction of machine remaining useful life (RUL). RUL, also called remaining service life, residual life or remnant life, indicate the time left before the failure occurs. The second one is to predict the chance that a machine operates without a fault or a failure up to some future time. This prediction could help to determine an inspection interval by estimating failure probability in this time period.

Although a variety of algorithms and techniques have been developed for diagnostic, prognostic algorithms for CBM have only recently been introduced in literature (Peng et al., 2010). In literature, similar approaches are used for diagnosis and prognosis which are classified in three main categories: Physical Model, Knowledge Based Model and Data Driven Model.

As mentioned above, physical models are utilized both for diagnostics and prognostics in literature. This approach uses a mathematical model related to physical processes that have direct or indirect effect on health of physical asset (Peng et al., 2010). Knowledge based model is based on a priori knowledge of state of system and its components. Expert System and Fuzzy Logic are two approaches used for knowledge based model are. Data-driven models are the models in which both previous inputs and outputs are known and measured. The main aim of data driven model is to figure out a relationship between measured input and output by using statistical and learning techniques. Peng et al. (2010) classify data-driven methods into two categories: statistical approaches and AI approaches.

CBM Applications and Results

CBM has proved to minimize the cost of maintenance, to improve operational safety and to reduce the quantity and severity of system failure. Rao (1996) explains that in 1988 a survey was conducted among 500 plants to evaluate the impact of CBM. Participants had been operating CBM for three or more years. The results of the survey show 50%-80% reduction in maintenance and repair costs and more than 30% reduction in spare part inventory emerged (Rao, 1996). Furthermore, saving of some companies due o predictive maintenance are also stated in his book.

Lee et al. (2006) introduce several case studies to compare several maintenance strategies in their study. Four maintenance strategies are defined as corrective maintenance strategy, scheduled maintenance strategy, condition based maintenance and predictive maintenance strategy based on maintenance scheduling. In this study maintenance labor availability is considered and it was assumed that any unscheduled equipment failure will be addressed when a maintenance team is available. Spare part inventory is not taken into account. Cost effects of maintenance are evaluated based on system state, total scheduled maintenance, total unscheduled maintenance maintaining time, unit cost for scheduled maintenance and unit cost for unscheduled maintenance. The result of the case studies verifies that as long as unscheduled failure maintenance is more expensive than scheduled one, cost benefit of last two strategies was higher than the corrective and scheduled maintenance strategies.

Beside the fact that superiority of CBM is proved theoretically like in Lee et al., its feasibility and practicability is also proved in many studies.

Li and Nilkitsaranont (2009) describe a prognostic approach to estimate the remaining useful life of gas turbine engines. Their approach provides valuable estimation of the engine remaining useful life and assists gas turbine users in their condition-based maintenance activities.

Blechertas et al. (2009) explain a systematic approach to US Army rotorcraft CBM and the resulting tangible benefits in their study. In this article, AH-64 Tail Rotor Gearbox case is studied, and results of cost benefit analysis of the rotorcraft Condition-Based Maintenance program which is implemented at the South Carolina Army National Guard is stated. Cost benefit analysis is done by figuring out investment cost and returns. Whether the benefits and returns exceed the investment shows the success of CBM program. As a result, $33.4 million savings in parts costs, $38.3 million savings in parts cost and operation support are observed. Furthermore productivity is increased through reduction in maintenance test flights and unscheduled maintenance and increase in mission flight time. Improvement in safety, sense of safety, morale, and performance are also verified outcomes of CBM implementation in this study. Shortly this case confirmed the CBM effect on increase in cost effectiveness, availability and safety practically.

Hoyle et al. (2007) analyze cost benefit of Integrated Systems Health Management (ISHM) in Aerospace Systems. As Condition Based Maintenance Policy, ISHM detects, assesses and isolate faults and so improves safety and reliability. It is used to determine optimum threshold level and inspection interval. Proposed ISHM framework is applied to aerospace system in their study. While calculating system cost and profit; System Availability, Cost of Detection and Cost of Risk are considered. Significant increase in profit, decrease in cost and increase in inspection interval is observed.

Kent and Murphy (2000) present cost benefit analysis of implementation of sensor based technologies for use in aerospace structure health monitoring systems (ASHMS). They focus on the cost and benefit of usage of health monitoring for maintenance. Such CBM policy requires high investment and they figure out whether the expected benefits are worth the high investment. This study leads to 30-40% improvement in maintenance. Reduced scheduled maintenance requirements, operational performance improvement, increased environmental safety are some of non-economic benefit of ASHMS.

1.3. Research Assignment

1.3.1. Problem Statement

To ensure competitiveness and getting a larger market shares, companies are forced to continuously decrease cost and increase productivity. Manufacturing companies use physical assets/capital goods to produce their end-products. The availability of these capital goods is the main concern of manufacturing companies to eliminate costly unexpected downtime and to increase productivity. Therefore maintenance becomes a significant issue for manufacturing companies.

Customers of ASML are unsatisfied with conventional maintenance techniques which are corrective and periodic maintenance. Periodic maintenance is based on the worst case scenario and customer usage. Therefore it causes over-maintenance and so extra downtime.

Furthermore it may not eliminate all unscheduled downs (USDs). On the other hand, reactive maintenance provides the use of whole life time of the machine, but an USD may lead to long down time and higher repair cost. Considering customer demand on increasing availability, ASML focuses on predictive tools to decrease down time.

Condition Based Maintenance (CBM) is a proactive maintenance strategy which increases availability of capital goods while eliminating over maintenance cost. By monitoring the condition of the system, the optimal maintenance strategy can be determined in terms of cost effectiveness, availability and safety. CBM policy helps ASML provide better maintenance solutions to customers (increased system availability and decreased associated costs).

1.3.2. Objective

The objective of this thesis is **to develop a data driven CBM decision support model which alerts the user before failures occur, and indicates the remaining useful life (RUL) of capital goods by using condition-based data**.

1.3.3. Research Scope

Implementation of condition based maintenance policy in ASML is a broad topic. The main output of the project is a data-driven decision support model which figures out the relationship between measured input (machine condition parameter) and output (machine health state) by using statistical and learning techniques. In other words, the failure prediction model will be built by using machine historical data without any system knowledge.

In general, to implement a CBM policy in ASML, the failures of all machine types have to be figured out. However there are millions of parameters to analyze and each machine consists of variety components which should be examined separately. Therefore the failure of the machine component could be seen as the root of machine failure. Rather than focusing on a machine failure, the critical component failures are taken as a starting point.

The project focuses on the development of a condition based maintenance decision support model for a single module. This module (which is referred to Module X in the rest of the report) is used on an installed base of more than 1000 systems. A high number of early lifetime failures (10 %) of the module have been observed. Furthermore maintenance of this part takes a long time and thus causes significant downtimes. Delays including diagnostics, parts delay and customer delay are the reason of 50 % of downtime caused by Module X (Figure 3). Therefore Module X is a significant component in order to keep machine operating. Explanation of this part failure contributes significantly to the explanation of machine failure. Through proactive maintenance, a significant amount of machine hours spend on unscheduled downtime (USD) could be saved.

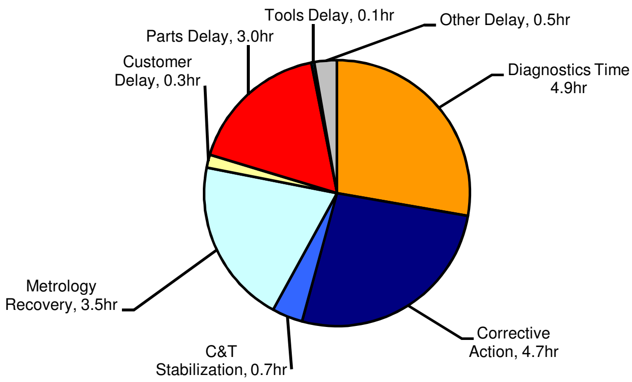


Figure 3: Breakdown of Module X Long downs

1.3.4. Research Methodology

This master thesis is a Business Problem Solving (BPS) project which focuses on the design of a solution for a business problem. Van Aken et al. (2007) state that "Problem solving projects aim at the design of a sound solution and at the realization of performance improvement through planned change." Furthermore they claim that a sound business solving project has to satisfy the following criteria, which we have adopted for this master thesis:

Performance focused: The main objective of the project should be to improve actual performance. This project points out the company problem and aims to develop a model which results in performance increase. ASML has to continuously improve Operational Expenditures and their main focus is to increase system availability. In line with ASML's objective, a model is built to increase uptimes.

Design Oriented: The projects steps are controlled by a project plan. This plan gives an insight about the project progress. Therefore while generating model, sound decisions could be taken.

Theory-based: Existing literature has been reviewed and evaluated. By contextualizing the theories for company problem, analysis and design activities are realized in this project. Therefore valid and state of art knowledge is used to solve the problem.

Client Centered: Since the proposed solution is an operational service for ASML, ASML requirements are identified and taken into account.

Justified: The solution is provided with reasoning behind it. Performance analysis is executed to justify the proposed solution.

The approach that we follow is **CRISP-DM** (CRoss-Industry Standard Process for Data Mining), which is the industry standard methodology for data mining and predictive analytics. It is a useful methodology to make large data mining projects faster, cheaper, more reliable and more manageable (Shearer, 2000). As shown in Figure 4, CRISP-DM organizes the data mining process into six phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. These phases help to understand the data mining process and guide a data mining project.

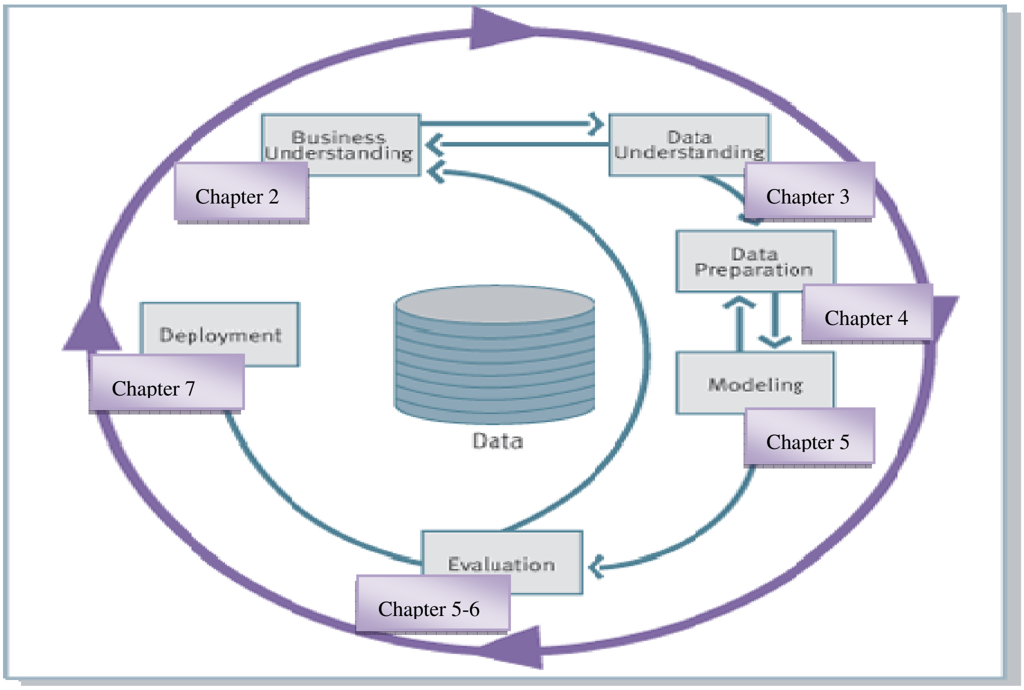


Figure 4: Phases of the Crisp- DM Process Model

1.3.5. Research Questions

In order to accomplish the objective, the following research questions have been formulated:

1. How can a condition based maintenance decision support model be designed technically?

There are three main steps in order to design a CBM decision support model which are (1) Data Acquisition, (2) Data Processing, and (3) Maintenance Decision Making.

1.1 How to perform the Data Acquisition step? *(Data Understanding-Chapter 3)*

Data Acquisition is the first step which includes collecting the condition and event data. Since ASML has recorded millions of data up to now, by assuming that the data are reliable, any more additional activity will not be performed for this step. Therefore obtained data will be used in the following steps.

1.2 How to perform the Data Processing step? *(Data Preparation-Chapter4)*

Data Processing consists of Data Cleaning and Data Analysis. Data Cleaning is required to eliminate data errors. Moreover, Data Analysis helps to understand and interpret data.

a. How to perform Data Cleaning?
b. How to perform Data Analysis?

i. What is the relationship between condition parameters and failure cases?
ii. What is the relationship among condition parameters?

1.3 How to perform Maintenance Decision Making step? *(Modeling-Chapter5)*

After data is acquired and interpreted, the decision support model is built. This model helps the user to take decisions by warning about the upcoming failure. There are various methods to predict the RUL of the module. After selecting most appropriate method, a complete CBM decision support model which alerts the user and shows RUL of capital goods, is designed.

a. What methods can be used to predict RUL?
b. What is the best method to be used for the CBM decision support model?

2. What is the difference between proposed model and physical model that has already been developed? *(Evaluation-Chapter 6)*

A physical model has already been developed by ASML by considering the physical behavior of Module X. In the final part of research assignment, the proposed data driven model will be compared with the physical model and its feasibility and success will be evaluated.

a. Does the proposed model perform better than the physical model?
b. What is the improvement amount in terms of previously identified performance measures? What is the attainment of data driven model compared to physical model?

1.4. Report Outline

This chapter provided background information about ASML and Customer Support Operational Services Department where the practical part of this master thesis was conducted. Then, brief information about maintenance, maintenance policies were given. Furthermore the condition based maintenance methodology and applications were explained in detail. Finally, the research assignment was clearly defined in this chapter. Based on the research methodology, the rest of the report is organized as follows. **Chapter 2** focuses on the understanding of ASML's business objectives and expectations. Moreover the data mining problem is designed in line with these objectives. **Chapter 3** points out understanding and exploration of the initial data. As a next step, **Chapter 4** explains all activities performed to obtain final data set from initial raw data. In **Chapter 5** the implementation of the modelling techniques, the creation of models, and the assessment of models are presented. After developing the prediction model, in **Chapter 6**, it is compared with the physical model. **Chapter 7** explains the deployment phase of the project. It gives information about the decision support model and user interface by which ASML can use the knowledge gained from the model. Finally, the conclusion and discussion are presented in **Chapter 8**.

2. Business Understanding

This chapter focuses on determining the business objectives, assessing the situation and determining the project goals.

2.1. Determination of Business Objectives:

Integrated Customer Solution is a key for semiconductor manufacturers to remain competitive. To preserve its market share and to satisfy customers, ASML provides high quality customized support services with technology.

Maintenance Support Service is an essential service provided with the technology. ASML implements periodic and corrective maintenance according to customer demand. If the customer wants to use the whole life time of the machine, reactive maintenance is performed (Figure 5). After a failure occurrence, it takes time to respond to the failure, then the customer and the service engineer discuss about the case. Service engineers starts to diagnose the reason(s) of the failure. Then required parts and tools are ordered. As soon as receiving ordered parts, maintenance is planned, executed and the machine starts working again. If the customer prefers preventive maintenance, service engineer performs maintenance periodically to eliminate failure. Periodic maintenance (scheduled maintenance) decreases downtime based on periods of worst case scenarios and customer usage assumptions. Although unscheduled down time is decreased, maintenance is performed too early and over maintenance is performed.

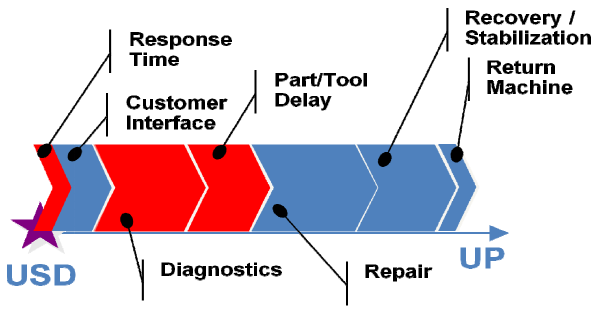


Figure 5: Steps of Corrective Maintenance

To increase availability and reduce scheduled and unscheduled downtime, predictive tools have been developed and started to use. Several years ago a pro-active initiative was started but due to many reasons the initiative was not funded. Recently local offices, Customer Support–Veldhoven and Industrial Engineering have started their own Pro-active initiatives to support demanding customers within their own competence. Local offices initiate warning monitoring, immersion parameter monitoring and scripts engineering. More than 3000 parameters per machine are recorded in a database (file archive) every day. These data are retained in the storage about 0.5 years. Such local initiatives have showed distinct improvements on unscheduled down (USD) and extreme long down (XLD) performance by monitoring systems. Therefore the Be-Warned project been started to translate local developed monitoring tooling knowledge into a sustainable solution.

CS-OS conducts the Be-warned Project to design and deliver predictive maintenance tools, methods, mindset and organization. Proactive Maintenance Models will be the basis of Be-Warned Project. As explained in the literature study conducted by Cakir (2011) (see Chapter 1.2), CBM decision models can be developed by using different modeling approaches such as Physical Model and Data Driven Model. In the scope of the Be-Warned project, ASML has already developed a physical model by using specific knowledge and theories relevant to the systems. As opposed to the physical model, data driven model without any system knowledge was developed in this project. This approach was taken to validate the expectation that, analysis of historical machine data together with the failure data, leads to correlation between particular data and the failure. This in turn is the starting point to design a model to predict the failure of the module without detailed system knowledge. The details about the data driven model are explained in the following sections.

Pilot Model: Physical Model

The physical model was developed to predict failure through understanding of the physical degradation behavior of Module X. It results in savings in labor hour per machine (LHM), increased availability of machine and decreased extreme long downtime. Furthermore, the model enables part failure prediction up to 10 weeks in advance. Performance of the model can be indicated as below:

Table 2: Confusion Matrix

		Predicted Class	
		Failure	**Non-Failure**
Actual Class	**Failure**	True Positives (TP)	False Negatives (FN)
	NonFailure	False Positives (FP)	True Negatives (TN)

$$\text{Sensitivity} = \frac{\text{Number of True Positives}}{\text{Number of True Positives} + \text{Number of False Negatives}}$$

$$\text{Specificity} = \frac{\text{Number of True Negatives}}{\text{Number of True Negatives} + \text{Number of False Positives}}$$

$$\text{Precision} = \frac{\text{Number of True Positives}}{\text{Number of True Positives} + \text{Number of False Positives}}$$

- Sensitivity: 91.0%
- Specificity: 96.0%
- Precision: 45.8%

Sensitivity and specificity shows that 91% of Failure cases and 96% of Non-Failure cases are recognized correctly respectively whereas precision indicates that only 46% of failure signal is correct. As a result, although 91% of part failures are predicted by this model, the model generates twice as much alert. In other words, this is a good model to prevent failures however too much preventive maintenance is implemented.

Business Objectives

ASML aims to develop sustainable predictive maintenance tools by using locally developed monitoring tooling knowledge. Within this context, the objective of this project is to develop a data driven, failure prediction model for Module X by using data mining methods. Besides, it is aimed to compare performance of data mining approach with the physical model.

Business Success Criteria

The success of this project can be measured by the following criteria:

- Utility of local monitoring data
- Discovery of system knowledge through data mining methods
- Validated failure prediction model which increases machine availability

2.2. Assess Situation

In order to develop a CBM decision model for Module X, large number of qualified data is required. Data is collected from customer fields and sent to the global data base. Data quality and data amount which cannot be controlled easily are significant constraints for this project.

Since it is aimed to discover knowledge through data mining methods, the knowledge about Module X working principle, the components of Module X explanation of the condition parameters, etc., which may give an idea about the part failure were not used until after the development of the model. Without any system knowledge, only data usage could be risky and may lead to misinterpretation of data and so do unreasonable models.

2.3. Determine Data Mining Goals

Data Mining Goals

Data mining which is also known as data or knowledge discovery is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining is the process of finding correlations or patterns in large relational databases (Data Mining, University of California).

Main objective is to predict the failure time and to warn the user about the upcoming failure by indicating remaining useful life (RUL) of the part.

Data Mining Success Criteria

Success of data mining can be assessed with the following criteria:

- Accuracy: the proportion of true results in the population.

Table 3 shows the example confusion matrix. Each column of the matrix represents the instances in a predicted class whereas each row represents the instances in an actual class.

Table 3: Confusion Matrix for Three Class Classifier

		Predicted Class		
		A	B	C
Actual Class	A	k	l	m
	B	n	o	p
	C	q	r	s

$$Accuracy = \frac{k+o+s}{k+l+m+n+o+p+q+r+s}$$

- True Positive Rate, False Positive Rate and Precision

These criteria explain the prediction accuracy in detail. The true positive rate (TP) is the proportion of positive cases that were correctly identified whereas the false positive rate (FP) is the proportion of negatives cases that were incorrectly classified as positive. Precision is the proportion of the true positives to all the positive results. Calculation of the true positive rate, false positive rate and precision for each class are shown in Table 4. Besides, averages of them are shown in the last row which can be taken into account if all classes are equally important. Higher true positive rate and precision and lower false positive rate indicate a better prediction model.

Table 4: Calculation of the True Positive Rate, the False Positive Rate and the Precision

	TP rate	FP rate	Precision
A	TP_A=k/(k+l+m)	FP_A=(l+m)/(k+l+m)	P_A=k/(k+n+q)
B	TP_B=o/(n+o+p)	FP_B=(n+p)/(n+o+p)	P_B=o/(l+o+r)
C	TP_C=s/(q+r+s)	FP_C=(q+r)/(q+r+q)	P_C=s/(m+p+s)
Weighted Average	(TP_A+TP_B+TP_C)/3	(FP_A+FP_B+FP_C)/3	(P_A+P_B+P_C)/3

2.4. Conclusion

This chapter has presented the evaluation of the project in terms of business perspectives. Background information about the business case has been provided and business expectations have been defined. Accordingly, data mining problem has been designed. Moreover the performance criteria of the overall system have been defined.

3. Data Understanding

This chapter aims to increase familiarity with the data which has been collected by ASML. It includes the description of data and the exploration of data.

Two types of data namely condition and event data were provided. Condition data indicates the state and health condition of the part whereas event data depicts cases and taken actions.

3.1. Condition Data

While machines operate, condition parameters, which are directly or indirectly related with Module X are recorded. The data which are retained in a database was extracted for the use of this project. Two data sets were provided, for the years 2009 and 2010, respectively.

The 2009 data set consists of 110 condition parameters which were taken from 884 machines in 1 year time period. 3,047,312 parameter values were recorded.

The 2010 data set is composed of 108 condition parameters which were taken from 106 machines in 1 year time period. 1,986,898 parameter values were recorded. The group of machines in this data set is a subset of the group of machines in the 2009 data set.

Besides, some information about Machine Type, Site Id, Customer Continent, Customer Country and Customer Number has been provided within data sets.

3.2. Event Data

Event data shows taken actions related to failure of Module X. ASML doesn't have direct information about the part failure. However, the part order time and machine failure time is known. Part ordering may not only indicate part failure but also stock demand or preventive maintenance. To make a clear link, part ordering time and machine failure time are cross checked. Then part orders because of the machine failure are specified. Although this is a reasonable approach to get failure time, its accuracy can be disputable. E1 and E2 error cases are still issues in the given event data.

E1: Although failure didn't occur, part was ordered.

E2: Although failure occurred, no part order were placed

For the years 2009-2010, 179 orders which include single or multiple parts are specified for this research. Whereas a single part order includes only Module X, multiple part order indicates both the order of Module X and some other machine components.

.

Table 5: Description of the Data Set

Attribute	Type	Description
Machine Nr	Categorical	Machine identifier
Time Stamp	Date	Identifies when the parameters were recorded
P955,...,P3780	Numeric	Condition Parameters (which take value between +4, -10).
Machine Type	Categorical	Machine Type
Site Id	Categorical	Site Identifier
Customer Id	Categorical	Customer Identifier
Customer Continent	Categorical	Customer Continent
Customer Country	Categorical	Customer Country
Part Order Time	Date	Probable failure time

3.3. Conclusion

In this chapter, general information about the data sets has been presented. 4 million numeric and categorical condition data was provided by the company. In addition to condition data, event data which reflects possible failure cases was given. Despite of the given huge data set the applicability of the data is not queried in this chapter. The following chapter includes data preparation and analysis steps that bring out applicable data set.

4. Data Preparation- Data Analysis

Obtaining high quality data is the first crucial step to generate a strong CBM decision support model. No matter how precise data is acquired, errors will still occur. This chapter presents steps to obtain qualified data and to eliminate E1and E2 error cases.

The 2009 data set and the 2010 data set were determined to use as training and test data respectively. Thus, analysis results were provided separately for each data set.

4.1. Selection of Failure Cases

179 part orders which includes single and multiple part orders were specified as potential failure cases. Multiple part orders may not be related to Module X failure. Another part failure, machine performance problems or inventory demand could be the reasons of multiple part orders. Therefore the orders including multiple parts were eliminated.

2009 Data Set: 58 failure cases with a single part order were used for further analysis. Since the part was ordered twice for 3 machines in 2009, 55 different machines were taken into consideration.

2010 Data Set: 71 failure cases with a single part order were used for further analysis. Since for 6 machines the part was ordered twice in 2010, 65 different machines were taken into consideration.

4.2. Selection of Parameters

2009 data set: 187,622 data values and 108 variable condition parameters associated with the specified 55 machine were given. However, 72 parameters were recorded just a few times (only for November-December 2009). Due to the lack of data, their effect on failure couldn't be analyzed and the only remaining 36 parameters were used to develop the model.

2010 data set: Use of 36 parameters in 2009 data set led to eliminate remaining parameters from 2010 data set. 44,802 data values for 65 machine and 36 parameters were selected to use in following sections.

4.3. Missing Data

Missing data means that valid values on one or more variables are not available for analysis. Missing data under 10% for an individual case or observation can generally be ignored except when the missing data occurs in a specific non-random fashion (Hair et al, 2009). Unless the missing data is less than 10 % or the cases with no missing data on any of variables provide the sufficient sample size for analysis, remedies should be applied. In this case, missing values are estimated by the imputation methods which substitute some value for a missing data.

Two types of missing data were recognized in the data sets. Zero (0) and minus ten (-10) parameter values indicates the unavailable and invalid data respectively.

2009 data set:

Parameter value=0 (1950 data)

Parameter value=-10 (2943 data)

(1950+2943)/187,622=0.026→3%

2010 Data Set:

Parameter value=0 (108 data)

Parameter value=-10 (477 data)

(108+477)/44,802=0.013→1.5%

For this case, missing data in both data sets can be deleted since they are less than 10%.

4.4. Data Alignment

Data set was given as a list of all independent records. Sample of given data format is shown in Appendix I. To observe the changes in the condition parameters in time, it is required to align 2009 (2010) data sets.

Step by step data alignment

1. Split the data into 55 (65) groups according to machine numbers.
2. Split the groups into subgroups according to parameter ids. 36 sub groups were obtained for each of the 55 (65) groups.
3. Sort data of subgroups chronologically (from oldest to newest).
4. Unify 36 subgroups by using the time stamp as an identifier. Per each specified time stamp, 36 valid parameters are pointed out. The other time stamps which include less than 36 parameters are eliminated.

As a result of data alignment, for each machine, during 1 year period simultaneous changes of 36 condition parameters can be observed. Sample of the aligned data format is given in Appendix I.

2009 Data Set: During this step, it was noticed that 4 machines suffers from lack of the condition data. Therefore 51 machines (54 failure cases) were used for further analysis.

2010 Data Set: In this data set, 54 machines suffer from lack of the condition data. Therefore 11 machines (11 failure cases) were used for testing. This gradual reduction in machine numbers can be explained with the changes in the parameter denotation. In other words for most of the machines after a certain period, different parameters were used to indicate the same conditions. If the parameters are translated and consistency is provided with the 2009 data set, more data can be used. However in this project, 11 failure cases were found sufficient for testing.

4.5. Detection of the Outliers

An outlier is detected by examining all metric variables to identify unique or extreme observations. Generally, outliers are defined according to standard scores or standard deviations. In small samples (80 or fewer observations), an observation is detected as an outlier if its standard score is ±2.5 or beyond. For large samples (more than 80 observations), an observation is classified as an outlier if its standard score is ±3.0 or beyond.

In this case, both misrecorded data and the part failure might be classified as outliers. To differentiate wrong data and failure cases, it is required to observe data changes in time. Whereas one time gradual change indicates the data error, continuous deviation in data illustrates the part failure.

The outlier detection methods cannot handle the classification of outliers. To detect and eliminate misrecorded data, scatter plot was used. Data points were plotted onto a graph to display the spread of condition parameters versus time. For 62 (51 +11) machines, condition parameters versus time graphs were drawn and spikes were detected and eliminated.

4.6. Analysis of the Condition Parameters

36 condition parameters were assessed as functional. They are metric data and all are measured in interval scale. Apart from those, machine type, customer id and site id may also explain the variation of failure cases. They are non-metric (categorical) data and are measured in nominal scale. By assuming that Site id includes the information about Customer Country and Continent, they weren't used in the model.

Statistic Analysis has been performed on 36 parameters. Their main features are shown in Descriptive Statistics Table in Appendix I. The parameters take values between -9.60 and 3.60 with a mean value around -2.

To understand the relationship among parameters Factor Analysis was performed. Factor Analysis is an approach for determining dimensionality of a multidimensional set of items. It examines interrelationships among a larger set of variables and then attempts to explain them in a terms of their common underlying dimensions. These common underlying dimensions are factors which attempt to explain maximum variance in variables with minimum loss of information. Principal Component Analysis (PCA) method which is a type of Factor analysis is used to handle data with complicated correlation structure (Jardine, 2006).

PCA method was implemented in order to detect underlying dimensions of the parameters. Detailed output is given in the Appendix I. Hair et al. (2009) discuss several criteria to decide on the number of factors. Firstly the pattern matrix (Table 35) and the correlation matrix (Table 36) show that the parameters are highly correlated (above 98%) in groups of six. As a second criterion which is latent root criterion, only the factors having eigen values greater than 1 are considered as significant when the number of variables differ between 20 and 50. In this case (36 variables), six factors have eigen values greater than 1 (Table 34). Lastly, the percentage of variance criterion helps to decide the number of factors by looking at the cumulative percentage of total variance. The threshold value is taken as 60% since the information is less precise. As a result, at least 2 factors should be extracted (Table 34).

Considering all these criteria the number of factors was decided to be 6. Therefore, 36 parameters are clustered in six groups consisting of six parameters. Accordingly, instead of 36 parameters, 6 parameters, which are average values of each group, were used in the model.

Table 6: Correlated Parameters

Groups	P1	P2	P3	P4	P5	P6
Parameter IDs	955	961	967	979	985	991
	956	962	968	980	986	992
	957	963	969	981	987	993
	958	964	970	982	988	994
	959	965	971	983	989	995
	960	966	972	984	990	996

4.7. Analysis of the Event Data

2009 Data Set: To see the changes in parameters, parameter values were displayed over time by scatter plot. For about 80% of the machines (39/51), parameters change significantly before and after the failure. For the remaining 20% of machines, none of the parameters change due to the failure and maintenance records, they follow a stable trend, as shown in Figure 6. These cases could be an example of E1 which indicates that although failure did not occur, the part was ordered because of some other reasons.

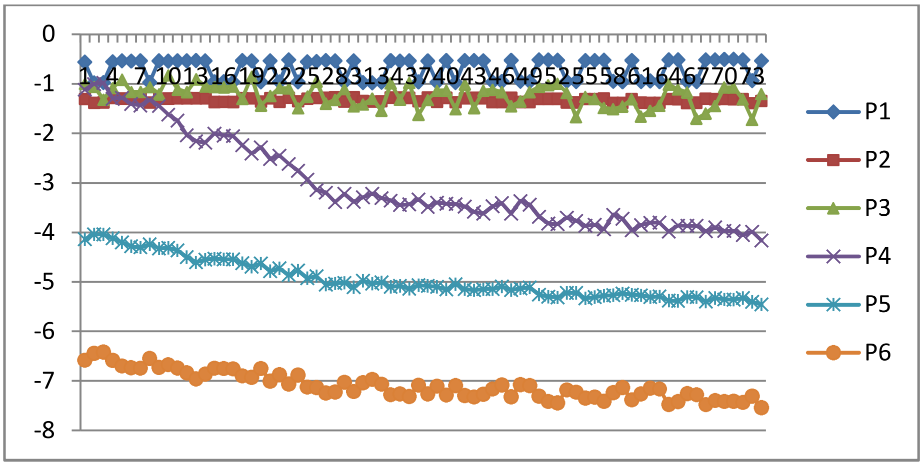


Figure 6: Parameter Values vs Time for the Machine 'M2693'

It was discovered that the part failure is related to reduction in data value. After maintenance significant and sudden increase of parameter value is observed. One or more parameters decrease until the failure. After performing maintenance, they go up to higher values.

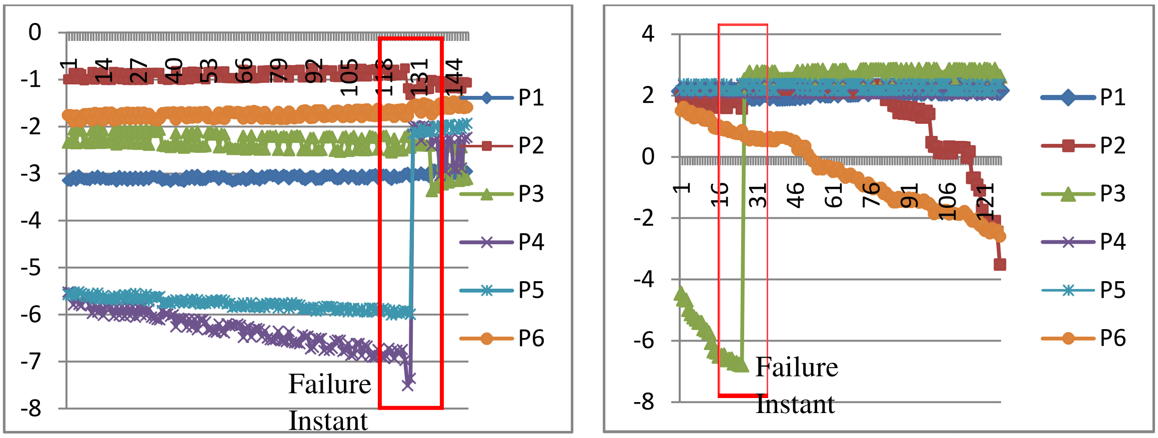


Figure 7: Failure Cases

9 of 39 failure cases show that parameters are not only affected by the particular part failure, but some other factors also affect the parameters. These factors might be unrecorded part failure, other parts' failure, machine intermittent etc. Similar situation is shown in figure below.

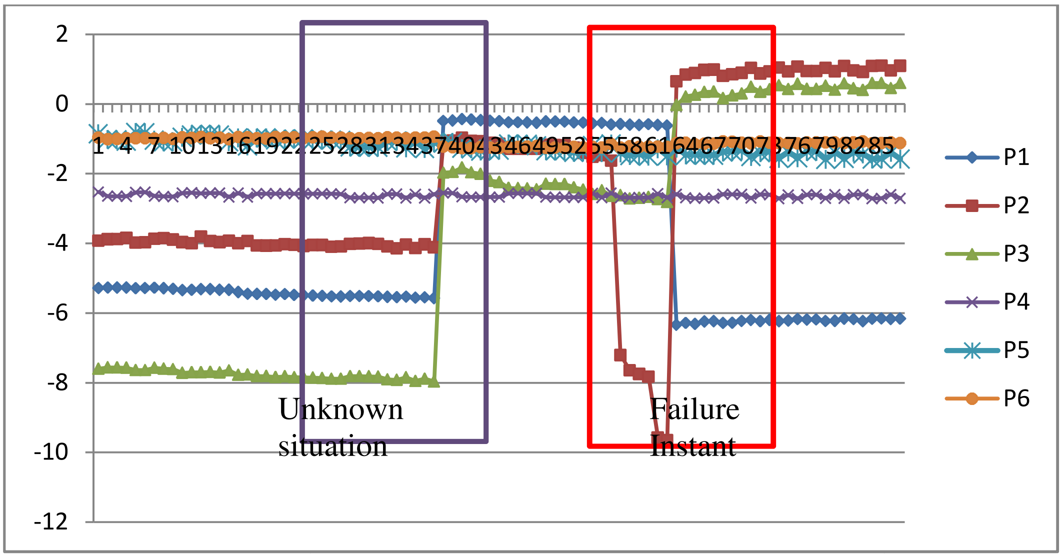


Figure 8: Effect of Unknown Factors on the Parameters

Modelling aims to detect failure cases through historical parameter data. The changes in parameter will contribute to develop the model. However such unexplained cases as no change in the parameter values (Figure 6) and uncontrolled changes in the parameter values (Figure 8) could make noise in the model. Therefore they were disposed.

2010 Data Set: For this data set, 11 failure events follow the similar trend as in Figure 7. One or more parameters decrease until the failure and they are recovered to normal values after maintenance. Since no unexplained trend is detected, all 11 cases were decided to use for model testing.

The summary of selection of the failure cases are shown in Figure 9. 41 failure cases out of 179 failure cases can be used for modelling. Although the number of cases is sufficient to develop a model, more information leads to build a better model by considering more cases.

Figure 9: Summary of the Event Selection

4.8. Gaps between the Time Stamps

Condition data were collected from functioning systems in aperiodic intervals. Sample size varies between 3-4 samples per a day and 1 sample per 54 days. Therefore, big gaps are observed for some periods. The average interval between the timestamps is about 11 days. To fill the gaps in time stamp, artificial time stamps were assigned. If the time between two time stamps is more than 20 days (more than 2 times of the average interval), artificial time stamp was created for the midpoint and the parameter takes the average value of two consecutive values. Related formula is given below.

$P_{i,t}$: Parameter i registered at time t.

t_a: artificial time stamp $\quad t_i < t_a < t_{i+1}$

$$P_{i,ta} = \frac{P_{i,t+1} + P_{i,t}}{2} \qquad i = 1..6$$

4.9. Conclusion

In this section, 2009 and 2010 data sets were analyzed elaborately. Applicable failure cases and parameters were selected. Data were organized to be used for modelling and testing. The relation between condition parameters and event cases were discovered and the unaccountable cases were eliminated. Furthermore, model inputs and so model complexity were reduced by grouping correlated parameters. As a result, the final data set which is applicable for failure prediction modelling was selected.

For supervised learning, the quality of the given data is crucial. Missing or wrong information leads to error in the model and so to inaccurate outcomes. For this case, many data were eliminated in order to prevent noise in the model. Rather, accurate and complete information should be used as an input to get more generalizable and robust model. Exact part failure time, machine states, the other parts' failure, machine performance problems etc. should be known to understand and interpret whole changes in parameters.

5. Development of a Prediction Model

In this chapter, the development of a data-driven failure prediction model is presented. As a result of the data preparation step, model inputs were defined as 6 metric condition parameters which are P1...P6 and 3 categorical condition parameters which are Machine type, Site id and Customer id. Besides that, 30 event cases from the 2009 data set and 11 event cases from the 2010 data set were selected to use for modelling and testing respectively.

Firstly, the failure threshold level and the effects of the environmental factors are analyzed. Then prediction models are developed by using three distinct approaches. Finally, models are compared in terms of their prediction accuracy.

5.1. Failure Percentage with respect to the Threshold Level

Repairs or replacements of Module X are performed once the degradation level reaches a threshold level.

Since reduction in the parameter values triggers the failure, the threshold level for each case has been determined according to the parameter value which reaches the minimum level at the failure instant. The threshold level is situation dependent and deterministic. It varies for the 30 failure cases (in the training data set) between -5 and -9.6. There is not a fixed threshold level because the maintenance demand depends on the customer expectation about the machine performance (Appendix II). Different threshold levels indicate that some customers wait for the hard failure and perform maintenance (corrective maintenance) as late as possible, whereas others suffer from the performance reduction (preventive maintenance) and perform maintenance earlier. In addition to the customer dependency, the threshold level is also dependent to the situation. A customer may define different threshold levels for different circumstances. For example, according to the demand, customers might prefer to postpone the maintenance or bring it forward.

The threshold level should be defined by the customers in order to predict the failure time. As mentioned above, the threshold level is dependent to the customer expectation on the module performance. If the unacceptable performance level, which is considered as the failure, is specified, corresponding threshold level can be discovered. While the module operates, the performance of the module decreases. AS monitoring the performance regularly, the service engineer decides the performance level at which the customer is unsatisfied with, and the customer prefers to execute maintenance. Then, the threshold level which corresponds to this performance level is specified.

Table 7 depicts the number of failure cases observed between the upper (UL) and lower (LL) threshold limits.

Table 7: Number of Failure Cases for Varying Threshold Level

Upper Threshold Limit	Lower Threshold Limit	Observed Instances	Probability	Cumulative Probability
-9	-9.6	8	0.267	1
-7.8	-9	5	0.167	0.73
-7	-7.8	5	0.167	0.56
-6.6	-7	5	0.167	0.4
-6	-6.6	4	0.133	0.23
-5	-6	3	0.1	0.1
4	-5	0	0	0

The failure percentage (cumulative probability) indicates how many cases failed before the lower limit. As the threshold level decreases, the failure percentage increases. As seen from the table, whereas for 23% of the cases threshold level is greater than -6.6; for 100% of the failure cases, the threshold level is greater than -9.6. It could be deducted that machine performance decreases as the parameter values decrease. To explain the relation between the failure percentage and the threshold level (Figure 10), piecewise linear regression, 2nd order polynomial regression and logarithmic regression methods were used. Piecewise linear regression provides the best model which explains 97% deviation in the failure percentage with the threshold level. (Appendix II)

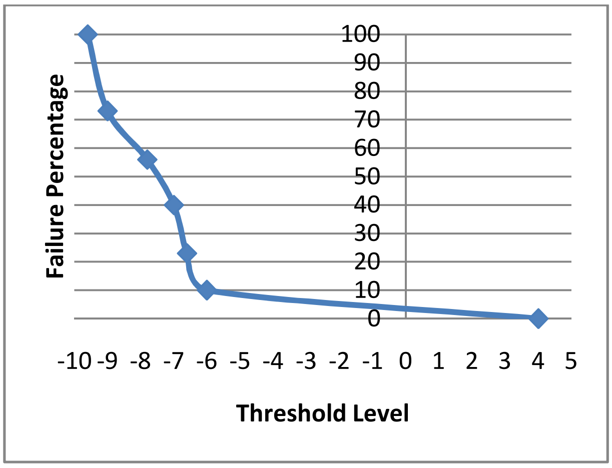


Figure 10: Failure Percentage vs Threshold Level

As a result of linear regression model, the threshold levels with the corresponding failure percentage and failure classes were defined as shown in Table 8.

Table 8: Failure Percentage with the corresponding Failure Classes

Failure Percentage		Threshold		Failure Class
LL	UL	UL	LL	
80	100	-8.94	-9.8	A
60	80	-8.1	-8.94	B
40	60	-7.22	-8.1	C
20	40	-6.37	-7.22	D
10	20	-6	-6.37	E
7	10	-3	-6	F
0	7	4	-3	G

5.2. Effects of the Environmental Factors

As explained above, the threshold level is dependent to the several factors. Machine type, customer id and site id are the environmental factors which may explain the variation in the threshold level. 30 failure cases were used to analyze and understand the effect of these environmental factors. However sample size of the factors should be sufficient to represent the groups. In this study, groups with more than 3 samples were taken into account. Appendix III shows the sample size for each factor. Therefore generalization can be done for T0007 and T0010 machine types, S1243, S366 site ids, and C1 and C1665 customer ids which have more than 3 samples.

Figure 11, Figure 12 and Figure 13 were plotted based on all data taken for the 30 failure cases. The dots in the figures indicate that there is at least one instance belongs to indicated failure class (vertical axis) in that group (horizontal axis). Among two machine types, the effect of the machine type T0010 on the failure classes is obvious (Figure 11). For this type of machine, users do not allow hard failures and maintenance is performed at the higher threshold level (greater than -7.22), which corresponds to D-E-F-G failure classes. As seen from Figure 12 and Figure 13, Sites S1243, S366 and customers C1, C1665, do not explain the variation of failure classes. Therefore environmental factors were not used for the modelling due to lack of their representativeness.

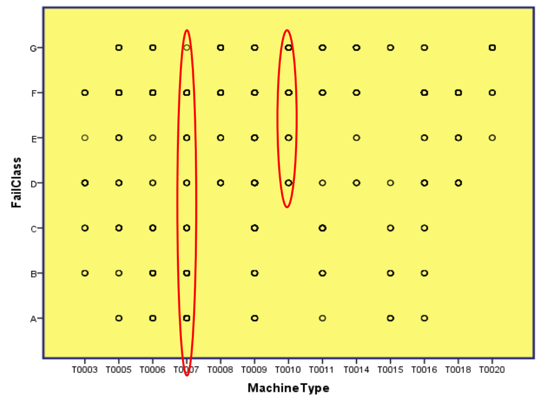


Figure 11: Machine Type Effect on the Failure Classes

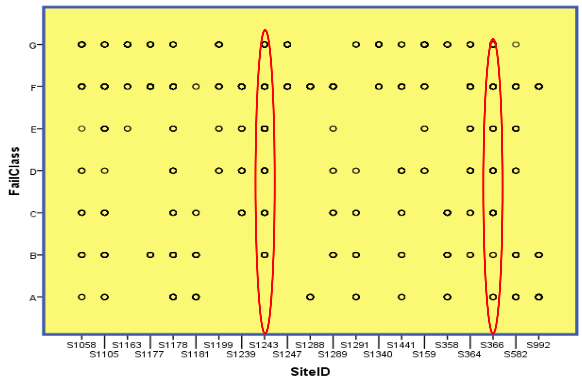


Figure 12: Site Id Effect on the Failure Classes

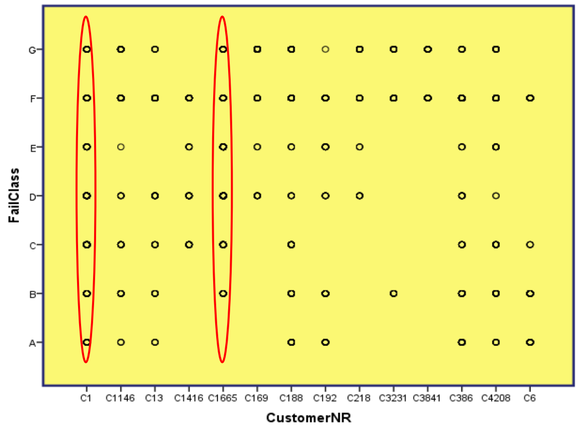


Figure 13: Customer Effect on the Failure Classes

5.3. First Modeling Approach

The classification tools Weka and KNIME were used to develop models which predict the remaining time of Module X. Only the samples which were taken before the failure time have been selected to use for the modelling. The actual remaining time to the failure is calculated for each sample by calculating the difference between the time stamp of the sample and the failure time. According to the calculated remaining time, data is clustered into three groups.

If the remaining time is greater than XX days→Cluster A

Else if the remaining time is greater than YY days→Cluster B

Else →Cluster C

It is aimed to produce explanatory signals for the user with these groups.

(A) Failure will not occur before XX days, no action is needed.
(B) Failure will occur in XX days, start to plan maintenance.
(C) Failure will occur in YY days, perform maintenance as soon as possible

To find the optimum XX and YY values, several iterations were performed on the training data by using the decision tree method. To evaluate the models, average of True Positives (TP) rate, False Positives (FP) rate and precision of the three clusters were compared. Besides, classification accuracy of the models on the cluster C has been assessed since the cluster C is the most significant alert for the user (see chapter 2.3, for calculation details).

Table 9: Assessment of Different Interval Values for Clustering

	Weighted Average			C Cluster		
XX-YY	TP	FP	Precision	TP	FP	Precision
60-20	0.560941	0.439059	0.724471	0.406542	0.593458	0.604167
50-20	0.547726	0.452274	0.733161	0.350467	0.649533	0.681818
50-15	0.546571	0.453429	0.728546	0.357895	0.642105	0.62963
40-10	0.506837	0.493163	0.762629	0.152778	0.847222	0.733333

		Predicted Class		
		A	B	C
Actual Class	A	1362	7	20
	B	144	76	37
	C	105	22	87

Therefore 60 and 20 days which results in higher TP rate and precision; and lower FP rate are selected for the XX and YY value, respectively. Whereas cluster A refers that the failure will not occur in the following 60 days, cluster C shows that the failure will occur in 20 days.

Various data mining techniques were implemented to classify the part life as Cluster A, Cluster B or Cluster C. These techniques were explained briefly in the following sections. For more information see Han and Kamber (2006). Six condition parameters (P1...P6), were taken as the inputs. 2009 data set was used to develop models and 2010 data set was kept apart for testing.

1. Artificial Neural Network

Artificial Neural Network is one of the common ways to perform machine failure prognosis and diagnosis. It is useful method when large amount of noisy and temporal data is available and physical, statistical or deterministic model is not known or impractical to apply. In other words, through ANN, complex, multidimensional, non linear systems can be modelled without a physical understanding of the system behaviour. On the other hand, such models are lacking of ability to explain themselves.

a) Multilayer Perceptron (WEKA)

A multilayer perceptron is a feed forward artificial neural network that uses back propagation to classify instances. A signal inside the neural network flows from the input layer passing hidden layers to the output layer. The goal of the training process is to find the set of weight values so that neural network output matches with the actual target values as closely as possible. While training the error correction of neural weights are done in the opposite direction. This is done by the back propagation algorithm.

To build a multilayer perceptron neural network, the network parameters which are learning rate (LR), momentum (M) and number of neuron per hidden layer (NHL) should be determined. Learning rate is the amount the weights that are updated. While lower learning rate may lead to the risk of the network to be stuck in local minimum, higher learning rate may result in undesirable oscillations. Momentum is applied to adjust the weights. Besides, number of neurons per hidden layer affects the structure of the network. Weka designs neural network with a single hidden layer. The proper choosing of learning rate and momentum is done by experience. Both values have a range between 0 and 1. Weka Classifier tool uses a default value of 0.3 for learning rate, 0.2 for momentum and ‘a’ (a= (attributes + classes) / 2) for neuron per hidden layer.

Firstly, training and the test data were defined. Several models were developed by adjusting the network parameters. Optimum network which results in higher TP rate and Precision and lower FP rate for the test data, has been searched. The optimal combination for TP, FP and Precision has been obtained when M=0.7, LR=0.2, NHL=a.

Table 10: Results of MLP-NN models for Variable Parameters

Inputs	M	LR	NHL	TP Rate (%)	FP Rate (%)	Precision (%)
P1...P6	0.1	0.2	a=4	0.407	0.377	0.355
P1...P6	0.3	0.2	a	0.413	0.385	0.309
P1...P6	0.5	0.2	a	0.413	0.386	0.311
P1...P6	0.7	0.2	a	0.515	0.347	0.459
P1...P6	0.8	0.2	a	0.475	0.363	0.340
P1...P6	0.9	0.2	a	0.479	0.360	0.343
P1...P6	0.7	0.1	a	0.413	0.386	0.313
P1...P6	0.7	0.3	a	0.475	0.363	0.340
P1...P6	0.7	0.5	a	0.498	0.353	0.524
P1...P6	0.7	0.7	a	0.413	0.387	0.314
P1...P6	0.7	0.2	1	0.423	0.434	0.271
P1...P6	0.7	0.2	2	0.416	0.449	0.224
P1...P6	0.7	0.2	3	0.459	0.368	0.328
P1...P6	0.7	0.2	5	0.275	0.394	0.397
P1...P6	0.7	0.2	6	0.266	0.380	0.223

The MLP method just considers the static nature of the neural network. Most of the software does not provide any method to implement recurrent networks. In order to capture the dynamic nature of the underlying process, different time windows were used. This approach takes the previously inscribed data into account as a new input which uses similar principal with the recurrent network. While momentum, learning rate and hidden number of layers were kept constant, variable inputs were tested through time windows method. For example, Time Window 2 indicates that current time stamp data and previous time stamp data are used as the inputs for the model, Time Window 5 indicates that current time stamp and last 4 time stamps are taken as the inputs. This approach causes to increase the number of inputs (6 additional parameters for each time window) and so complexity.

Here are the additional inputs which correspond to prior values of condition parameters (P1...P6).

$P7_t ... P12_t = P1_{t-1} ... P6_{t-1}$
$P13_t ... P18_t = P1_{t-2} ... P6_{t-2}$
$P19_t ... P24_t = P1_{t-3} ... P6_{t-3}$
$P25_t ... P30_t = P1_{t-4} ... P6_{t-4}$
$P31_t ... P36_t = P1_{t-5} ... P6_{t-5}$
$P37_t ... P42_t = P1_{t-6} ... P6_{t-6}$
$P43_t ... P48_t = P1_{t-7} ... P6_{t-7}$
t indicates current timestamp

Table 11 shows the result of MLP models for variable time windows. Time-Window-2 provides the higher TP rate and precision and lower FP rate.

Table 11: Results of MLP-NN Models for Variable Inputs

Inputs	M	LR	NHL	TP Rate (%)	FP Rate (%)	Precision (%)
P1...P6	0.3	0.2	a	0.413	0.385	0.309
P1...P12	0.3	0.2	a	0.492	0.294	0.447
P1...P18	0.3	0.2	a	0.42	0.328	0.39
P1...P24	0.3	0.2	a	0.338	0.36	0.274
P1...P36	0.3	0.2	a	0.305	0.34	0.264
P1...P48	0.3	0.2	a	0.479	0.293	0.365
P1...P60	0.3	0.2	a	0.508	0.361	0.487

Several iterations were performed to tune the network parameter values. At each iteration, one of the parameters was changed and the others were kept constant. Optimum network has been obtained when M=0.3, LR=0.2, NHL=a and inputs are P1...P12 (Table 12).

Table 12: Results of MLP-Neural Networks for Variable Parameters

Inputs	M	LR	NHL	TP Rate (%)	FP Rate (%)	Precision (%)
P1...P12	0.3	0.1	a	0.328	0.396	0.249
P1...P12	0.3	0.2	a	0.492	0.294	0.447
P1...P12	0.3	0.3	a	0.472	0.316	0.359
P1...P12	0.3	0.5	a	0.279	0.366	0.235
P1...P12	0.3	0.7	a	0.43	0.361	0.352
P1...P12	0.1	0.2	a	0.259	0.459	0.316
P1...P12	0.2	0.2	a	0.456	0.331	0.339
P1...P12	0.3	0.2	a	0.492	0.294	0.447
P1...P12	0.5	0.2	a	0.282	0.362	0.239
P1...P12	0.7	0.2	a	0.38	0.37	0.288
P1...P12	0.3	0.2	a=7	0.492	0.294	0.447
P1...P12	0.3	0.2	6	0.439	0.406	0.369
P1...P12	0.3	0.2	8	0.495	0.305	0.521
P1...P12	0.3	0.2	12	0.318	0.394	0.355

To sum up, MLP method when inputs=P1...P6 (single time window), M=0.7, LR=0.2, NHL=4 results in comparatively better prediction model. Table 13 shows the corresponding confusion matrix.

Table 13: MLP model (M=0.7, LR=0.2, NHL=4) Confusion Matrix

		Predicted Class		
		A	B	C
Actual Class	A	120	17	0
	B	64	3	16
	C	50	1	34

b) Radial Basis Function (WEKA):

A Radial Basis Function (RBF) is two layered feed forward neural network. The neurons in the hidden layer contain Gaussian transfer functions whose outputs are inversely proportional to the distance from the center of the neuron.

WEKA RBF classifier uses the k-means clustering algorithm to provide the basis functions and learns either a logistic regression (discrete class problems) or a linear regression (numeric class problems). In this approach, RBF was used to specify classes. It classifies all the samples as "cluster A" so it does not give signal about the upcoming failure.

Table 14: Results of RBF-NN on Test Data

	TP Rate (%)	FP Rate (%)	Precision (%)
A	1	1	0.449
B	0	0	0
C	0	0	0
Weighted Avg.	0.449	0.449	0.202

Neural Network models (MLP and RBF) do not help to understand the failure propagation mechanism and to diagnose faulty component. Therefore their applicability is limited.

2. Decision Tree (KNIME)

KNIME decision tree has been used to map the input variables to the Cluster A, B or C. It is developed based on C4.5 algorithm (for more information about C4.5 algorithm, see Quinlan (1993)). By changing the value of minimum number of records, optimum decision tree which results in higher prediction accuracy on test data is searched. Highest accuracy on the test data is obtained when the minimum number of records is equal to 30.

Table 15: Results of Decision Tree Models for Variable Parameters

Min number of records	Training data Accuracy (%)	Test data Accuracy (%)
50	78.01	44.91
30	81.99	58.03
20	87.15	45.90
10	91.99	44.59
5	94.14	51.47

Test Data		Predicted Class		
		A	B	C
Actual Class	A	135	0	2
	B	52	0	31
	C	43	0	42

Missing Failures

Table 16: Results of Decision Tree Model (min number= 30) on Test Data

	TP	FP	Precision
A	0.985	0.0146	0.587
B	0	1	0
C	0.494	0.505	0.560
Weighted Avg.	0.493	0.506	0.382

This model can not detect Class B. Although there are less than 20 days remained to the failure, 43 samples are classified as "A" which indicates the higher remaining time.

In order to analyze the effect of the different failure threshold levels, data is divided into 3 groups as hard failure, medium failure and soft failure (Table 17). Distinct decision trees are developed for each group.

Table 17: Failure Groups

Hard Failure		Medium Failure		Soft Failure	
M0005	-9	M3398	-8.7	M0051	-6.9
M1004	-9.6	M3407	-8.6	M1959	-6.7
M1887	-9.6	M0018	-8.4	M2789	-6.5
M3411	-9.6	M1321	-8.4	M3083	-6.5
M0017	-9.5	M0067	-8	M2683	-6.3
M1937	-9.5	M2601	-7.8	M2252	-6.2
M0041	-9.3	M0006	-7.6	M0029	-5.8
M0741	-9.3	M1186	-7.4	M1828	-5.4
M2232	-9.2	M1771	-7.3	M2417	-5.2
		M1358	-7.1		
		M0021	-7		
		M0034	-7		

Table 18 shows the decision tree accuracy for training and test data for variable parameter value (minimum number of records). For the soft failure group, all samples in the test data are classified as "A" which causes to miss the failure cases. Different decision trees do not improve the classification in the soft and medium failure groups. However significant improvement (18%) is observed in the hard failure group. Therefore medium and soft failure groups are combined and another decision tree is developed for this group which results in 47% prediction accuracy on the test data.

Table 18: Results of the Decision Trees for Different Failure Levels

	Min number of records	Train data Accuracy (%)	Test data Accuracy (%)
Hard Failure	50	76.34	64.67
	30	81.99	65.76
	20	87.09	72.28
	10	89.25	62.5
	5	90.6	62.5
Medium Failure	50	80.37	26.83
	30	82.84	26.83
	20	89.63	42.68
	10	91.6	29.26
	5	94.2	29.26
Soft Failure	50	80.56	51.35
	30	88.95	51.35
	20	90.57	51.35
	10	93.81	51.35

4th Decision Tree for Soft and Medium Failures

	Min number of records	Train data Accuracy (%)	Test data Accuracy (%)
Medium-Soft	50	82.54	36.13
	30	82.40	31.09
	20	90.53	47.90
	10	93.55	31.09
	5	94.69	36.97

As a result, using two decision trees (one for the hard failure group of which threshold level is lower than -9 and one for both soft and medium failure groups) increases the prediction accuracy. The results of the combined decision tree model are shown in Table 19.

Table 19: Results of the Combined Decision Tree Model

Test Data/ Medium+Soft		Predicted Class A	Predicted Class B	Predicted Class C
Actual Class	A	29	0	8
	B	16	0	11
	C	20	7	28

Test Data/ Hard Failure		Predicted Class A	Predicted Class B	Predicted Class C
Actual Class	A	98	1	0
	B	21	19	16
	C	13	0	16

	TP Rate (%)	FP Rate (%)	Precision (%)
A	0.934	0.066	0.645
B	0.229	0.771	0.704
C	0.524	0.477	0.557
Weighted Avg.	0.562	0.438	0.635

Combined decision tree model provides 56%, 63% TP rate and precision respectively. Besides, such trees are easy to understand and enable the user detect faulty component (parameter). Decision tree models are shown in Appendix VI.

3. Simple Cart (WEKA)

Simple Cart is another method for tree based representation of decisions. Minimal cost complexity pruning technique is used for the classification by adjusting the minimum number of objects. Cost complexity pruning is implemented on fully induced tree which is fitting the training data. It prunes the tree by aiming to increase the accuracy and decrease the complexity (for more information about Simple Cart method, see Breiman et. al. (1984)). Simple Cart model is shown in Appendix VII. When the minimum number of object is equal to 10, the optimum model which provides 0.498 TP rate and 0.532 Precision on test data is obtained (Table 20).

Table 20: Results of Simple Cart for Variable Parameters

Min # of objects	TP Rate (%)	FP Rate (%)	Precision (%)
50	0.377	0.449	0.239
20	0.456	0.314	0.435
10	0.498	0.257	0.532
5	0.479	0.26	0.437

		Predicted Class		
		A	B	C
Actual Class	A	101	35	1
	B	33	50	0
	C	15	69	1

4. K Nearest Neighbor (KNIME)

K-nearest neighbor algorithm (KNN) is a method for classifying objects based on closest training examples. K is the user defined parameter. A sample is classified by assigning the class which is most frequent among the k training samples nearest to that query point.

KNIME were run for different k values and maximum accuracy is obtained when k=4. However TP rate and precision of cluster C are relatively low.

Table 21: Results of KNN Model for Variable Parameters

k	Accuracy (%)
1	43.279
2	41.639
3	43.279
4	45.574
5	43.279
6	43.279

	TP Rate (%)	FP Rate (%)	Precision (%)
A	0.802	0.197	0.493
B	0	1	0
C	0.341	0.659	0.354
Weighted Avg.	0.381	0.619	0.282

Conclusion

Neural network, decision tree, simple cart, KNN techniques have been used to develop models to classify remaining life of Module X as Cluster A, Cluster B or Cluster C. These clusters indicate the range of remaining time. Cluster A refers to the remaining time more than 60 days whereas cluster C refers to remaining time less than 20 days. For each technique, best model is searched by adjusting model parameters. As a result, it was found that combined decision tree model which is combination of two decision tree (one for hard failures and one for soft and medium failures) provides more accurate prediction (TP=0.56, FP=0.44, Precision=0.63).

5.4. Second Modeling Approach

As a second approach, it was aimed to estimate the remaining time in days instead of clusters. Neural network and linear regression techniques of Weka have been used to estimate the remaining time of Module X. These techniques were evaluated in terms of root mean squared error and mean absolute error. The smaller error value indicates the better model.

Several iterations were performed to tune the neural network parameters.Table 22 and Table 24show the result of MLP and RBF techniques for variable parameters.

Table 22: Result of the MLP Neural Network for Variable Parameter

	Inputs	M	LR	NHL	Root Mean Squared Error (day)	Mean Absolute Error (day)
NN-Multi Layer Perceptron	P1...P6	0.1	0.2	a=3	113.45	99.1
	P1...P6	0.3	0.2	a	83.82	71.71
	P1...P6	0.4	0.2	a	80.79	68.89
	P1...P6	0.5	0.2	a	82.81	70.47
	P1...P6	0.7	0.2	a	82.11	75.08
	P1...P6	0.4	0.1	a	80.35	68.32
	P1...P6	0.4	0.5	a	98.94	80.31
	P1...P6	0.4	0.7	a	97.72	82.09
	P1...P6	0.4	0.1	1	95.21	82.13
	P1...P6	0.4	0.1	2	94.15	78.5
	P1...P6	0.4	0.1	4	115.44	92.73
	P1...P6	0.4	0.1	5	116.83	102.48
	P1...P6	0.4	0.1	6	112.26	91.09

Table 23: Results of the Linear Regression Model

	Root Mean Squared Error	Mean Absolute Error	Model
Linear Regression	101.99	88.46	RUL=3.42 * P1 + -1.634 * P2 + 12.3182 * P3 + 2.494 * P5 + -7.1724 * P6 + 154.6398

Table 24: Results of the Radial Basis Function Models

		Root Mean Squared Error	Mean Absolute Error	Model
NN-Radial Basis Function	cluster=6	86.09	73.35	RUL=-21.7573 * pCluster_0_0 + -12.6426 * pCluster_0_1 + -33.8206 * pCluster_0_2 + 1.8987 * pCluster_0_3 + 12.1479 * pCluster_0_4 + 26.92 * pCluster_0_5 + 127.9365
	cluster=5	86.83	74.56	RUL=-21.2185 * pCluster_0_0 + 2.8558 * pCluster_0_1 + -33.1454 * pCluster_0_2 + 4.8096 * pCluster_0_3 + 12.454 * pCluster_0_4 + 127.2579
	cluster=4	80.57	69.73	RUL=24.2103 * pCluster_0_0 + -11.0148 * pCluster_0_1 + -4.1696 * pCluster_0_2 + -9.5119 * pCluster_0_3 + 128.26
	cluster=3	82.12	70.8	RUL=-4.6842 * pCluster_0_0 + 4.6613 * pCluster_0_1 + 129.3493
	cluster=2	82.24	71.35	RUL=-4.6842 * pCluster_0_0 + 4.6613 * pCluster_0_1 +129.3493
	cluster=1	83.92	72.54	RUL=128.2661

The best model was developed by MLP method when M=0.4, LR=0.1 and NHL=a. Furthermore the model's warning capability was assessed with the model's prediction accuracy in the last 30 days before the failure. Mean absolute error is found as 71 days. As a result, this model does not produce accurate signals to warn the user on time.

Furthermore, the results of MLP model was translated into A, B, C clusters in order to compare with the first approach. Table 25 shows confusion matrix and detailed prediction accuracy of the model. Nevertheless, higher prediction accuracy is obtained by the first approach (see Table 19).

Table 25: Results of MLP Method (M=0.4, LR=0.1, NHL=a)

		Predicted Class		
		A	B	C
Actual Class	A	47	4	2
	B	102	21	20
	C	59	34	16

	TP Rate (%)	FP Rate (%)	Precision (%)
A	0.887	0.113	0.225
B	0.356	0.644	0.356
C	0.147	0.853	0.421
Weighted Avg.	0.463	0.536	0.334

5.5. Third Modeling Approach

Machine learning techniques have been used in the previous approaches which predict RUL of Module X with 50% accuracy. As opposed to these approaches, third approach is based on understanding and interpretation of the parameters' behaviour. Data visualisation technique is used to understand the relationships in multidimensional data.

It is discovered that the part failure is because of the reduction in the data value. After maintenance, a significant and sudden increase in the parameter value is observed. One or more parameters decrease until the failure. After performing maintenance, some parameters go up to higher values. As understood from parameters variation over time for a machine, maintenance can recover all parameters or it can be performed for recovery of specific parameters. Regardless of the maintenance coverage, it always includes recovery of the highest degraded parameter. Therefore, it is reasonable to assume that maintenance is performed based on a single parameter so called "dominant parameter". Dominant parameter is the parameter that takes the minimum value in comparison with the other parameters. The fact that dominant parameter goes beyond the user threshold level leads to the failure. During failure diagnosis, service engineer may check all other parameters and implement complete maintenance or he may prefer to fix the problematic parameters.

Since the decrease in the parameter values causes the failure, firstly the parameters are classified to detect the decreasing trend. If a parameter value is less than -4 (upper limit for the failure threshold level) and it decreases continuously, it indicates degradation of the parameter. Decreasing trend is detected by using cumulative moving average (CA) method which smoothes out short term data fluctuation. In cumulative moving average, the average of all of the data up until the latest data point is calculated. Since CA method may not smooth all the fluctuations, the current cumulative average CA_{i+1} is not only compared with CA_i, but also with CA_{i-1} to discover reduction in parameter values.6 condition parameters are checked, whether they are in the degradation period or not. If the dominant parameter decreases then life of Module X is specified as failure period ("Y").

Classification step shows only whether the part is in failure period or non failure period. However it doesn't indicate when the part will fail. Failure may occur in a few days as well as in a few months. Therefore, more tangible output is required to warn the users and make them take better decisions. Secondly, RUL which shows the user how long the part operates before it fails is calculated. Hence service engineer gains time to plan maintenance, labor and spare parts.

Threshold level is determined according to the value of the dominant parameter at the failure instant. For 30 failure cases, the dominant parameter takes different values between -4 and -9.6 at the failure instant. Therefore, there is not a fixed threshold level because the user perception about the failure depends on the user expectation. Different threshold levels indicate that some customers wait for the hard failure and perform maintenance as late as possible, whereas others suffer from the performance reduction and perform maintenance earlier. Considering these issues, RUL of the module is calculated for different threshold levels. The most suitable threshold level which is the closest to the actual value, leads to the smaller error between the actual and predicted RUL.

Since the parameters decrease linearly in the failure period, it is reasonable to calculate RUL by performing a linear extrapolation. By using the first and the last sample points, a linear line is created and it is extended to the predefined threshold level. Therefore the remaining life of the module is calculated. In Figure 14, three RUL predictions are shown for a machine. Predictions are made at different times and the prediction accuracy increases as the data points get closer to the failure instant. Prediction 3 which is calculated just before the failure leads to better prediction than Prediction 2 and Prediction 1, which are determined by using the previous data points.

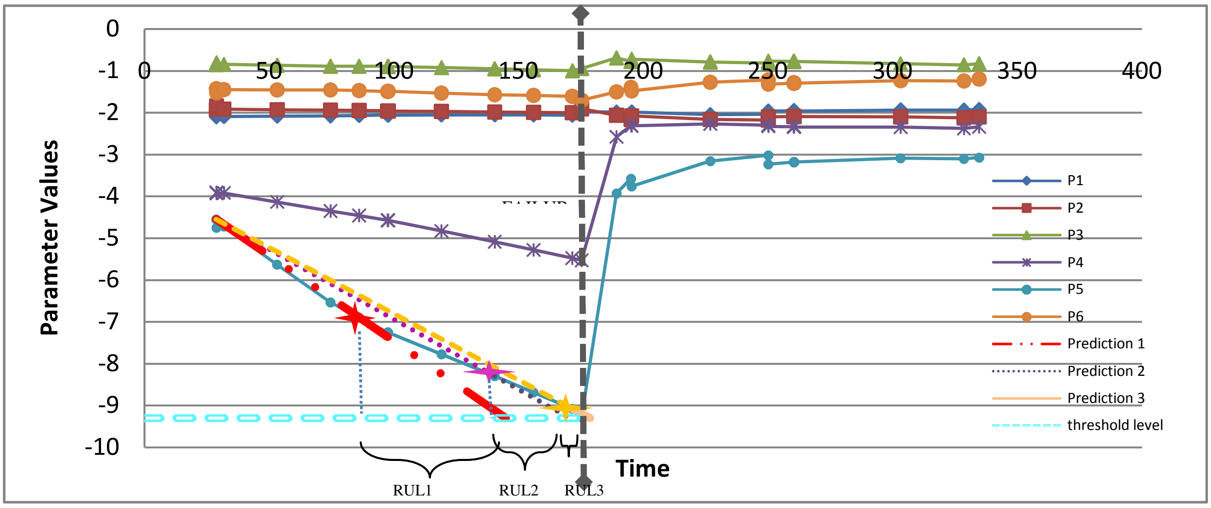


Figure 14: RUL Calculation Based on the Dominant Parameter

Formulation

Pi: Condition Parameter i where i=1...6

d: dominant parameter

t: time stamp

HS: Health Status

$CA_{i,t}$: Cumulative Moving average of parameter i at time t.

$X_{i,t}$: value of parameter i at time t

Step 1: Dominant Parameter:

Min (P1...P6)

Step 2: Maintenance Diagnosis:

Jumps in the parameter values indicate that parameter is recovered. Therefore, if a parameter is increased by 2 or more, this depicts the recovery of the parameter.

Recovery Diagnosis (on each Parameter)=$P_{i,t} - P_{i,t-1} > 2 \rightarrow P_i\ is\ recovered\ at\ time\ t.$

Step 3: Diagnosis: Degradation of Parameters

Each parameter is checked to detect whether it follows stable trend or decreasing trend

$$CA_{i,t} = \frac{x_{i,1}+\cdots+x_{i,t}}{t}$$

$$CA_{i,t+1} = \frac{x_{i,t+1} + tCA_{i,t}}{t+1}$$

IF (Pi $< -4\ and\ (CA_{i+1} < CA_i$ or $CA_{i+1} < CA_{i-1})$ "Y" (decreasing trend in critical period)

$ELSE$ $"N"$

Step 4: Classification of Health Status of Module X:

If Dominant Parameter is in degradation, Failure "Y"

Else Non-Failure "N"

Step 5: Detection of the starting point of degradation: The point where degradation starts:

HS_{t-1} = "N" and HS_t= "Y"

P_{1o}, P_{2o}, P_{3o}, P_{4o}, P_{5o}, P_{6o} ,T_o

Keep the starting point of degradation in the memory to compute RUL.

***Step 6: Prognosis: RUL Computation* (Linear Extrapolation)**

If Health Status of Module X = "Y"

$RUL_{i,t} = \frac{TL-P_{it}}{P_{it}-P_{io}} * (t - To)$ (in days)

Else RUL_i=5000 (no alert)

RUL of the component is equal to the RUL of the dominant parameter.

$RUL_{Module\ X}$= RUL_d where

This step is performed for different threshold levels.

Results

Figure 15 shows the predicted and the actual RUL for different machines (M0005, M0006, M0017, M0018). The accuracy of RUL prediction depends on the data behaviour. First few data samples might not be sufficient to indicate the trend of the dominant parameter, which causes high error of initial predictions. The predictions are improved and approximated to the actual RUL as more data is considered. As seen from Figure 15a, Figure 15c, Figure 15d, the error between the actual and predicted RUL gradually decreases. However if the parameter fluctuates at high frequency as in Figure 15b, predictions may not be consistent and reasonable.

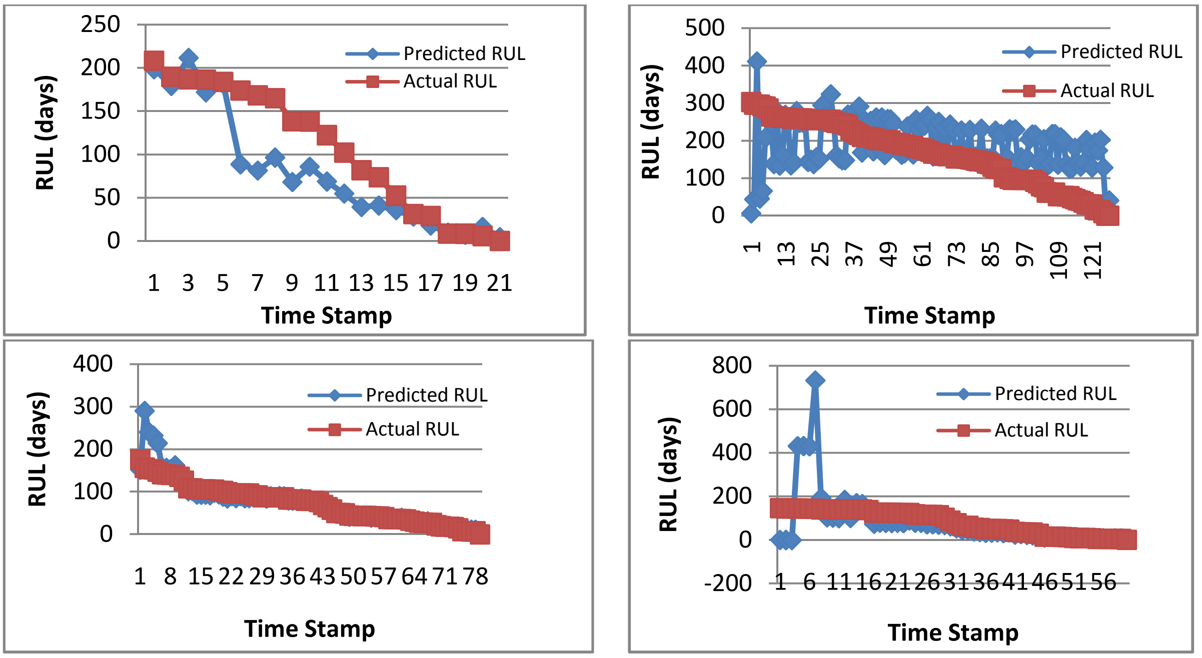


Figure 15: Changes in Actual and Predicted RUL for Machines: M0005 (a). M0006(b), M00017(c), M00018(d)

According to the given formula, RUL of the module has been predicted for 6 different threshold levels.

Table 26 and Table 27 show the mean error between actual and predicted RUL for each threshold level. If the error is smaller, the threshold level is closer to the actual level. That the error is equal to "#DIV/0!" indicates that corresponding threshold level is far beyond the actual level so RUL is not predicted. Therefore, for each failure case the smallest value, which is highlighted in the tables, is taken into account. In the last 30 days before the failure, the deviation between the predicted and the actual RUL is about 36 days and 42.5 days for the training and the test data set respectively. Although degradation level is checked if one of the parameters is less than -4 in the model, the failure of the machine M1959 occurs before this level. In other words, the failure occurs so unexpectedly early that the model cannot predict it.

Table 26: Mean Absolute Error between the Predicted RUL and the Actual RUL (Training Data Set)

	|RULact-RULpre| (last 30 days) for different Threshold Levels						
	-5	**-6**	**-7**	**-8**	**-9**	**-9.6**	**Mean Absolute Error (day)**
M0005	449.45	336.88	224.31	111.74	5.02	68.37	5.02
M0006	408.65	203.56	35.52	206.62	248.45	#DIV/0!	35.52
M0017	440.60	342.27	243.94	145.62	47.29	11.71	11.71
M0018	140.05	98.75	57.46	16.17	25.12	49.90	16.17
M0021	52.48	23.64	5.28	34.03	62.86	80.16	5.28
M0029	330.92	212.33	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	212.33
M0034	86.53	41.48	5.33	48.63	93.68	120.72	5.33
M0041	138.91	106.55	74.20	41.84	9.48	9.93	9.48
M0051	160.74	72.29	16.17	104.62	193.08	246.15	16.17
M0067	466.62	279.39	92.15	95.08	185.94	278.35	92.15
M0741	1617.41	1240.56	863.70	486.84	109.99	116.13	109.99
M1004	549.16	429.76	310.37	190.97	71.58	3.38	3.38
M1186	266.82	155.59	44.36	66.86	178.09	244.82	44.36
M1321	584.07	410.07	236.07	62.05	111.95	216.35	62.05
M1358	1645.53	854.22	62.90	#DIV/0!	#DIV/0!	#DIV/0!	62.90
M1771	186.07	103.92	21.76	60.39	142.55	191.84	21.76
M1828	75.87	99.09	189.35	#DIV/0!	#DIV/0!	#DIV/0!	75.87
M1887	1109.67	861.37	613.07	364.77	116.47	32.52	32.52
M1937	368.91	287.56	206.21	124.86	43.51	8.77	8.77
M1959	38.04	15.31	7.42	30.15	52.88	66.52	7.42
M2232	114.79	88.86	62.94	37.02	11.10	5.68	5.68
M2252	218.47	43.59	131.28	217.92	#DIV/0!	#DIV/0!	43.59
M2417	1.04	8.05	16.99	25.94	34.89	40.26	1.04
M2601	1414.47	911.82	409.17	93.49	#DIV/0!	#DIV/0!	93.49
M2683	21.65	6.04	9.56	25.17	40.78	50.15	6.04
M2789	13.15	4.09	4.96	14.02	23.07	28.50	4.09
M3083	65.79	26.01	14.21	53.56	93.35	117.22	14.21
M3398	743.70	543.78	343.87	143.95	65.19	125.80	65.19
M3407	162.65	118.92	75.19	31.47	12.26	38.50	12.26
M3411	419.36	328.17	236.98	145.79	54.60	3.48	3.48
						Average	36.24 days

Table 27: Mean Absolute Error between the Predicted RUL and the Actual RUL (Test Data Set)

	|RULact-RULpre| (last 30 days)						
	-5	-6	-7	-8	-9	-9.6	Mean Absolute Error (day)
M0019	302.99	235.47	167.95	100.43	32.91	7.85	7.85
M0067	441.41	310.57	179.73	48.89	81.95	160.46	48.89
M0885	320.25	241.24	162.24	83.24	17.38	44.63	17.38
M0939	682.92	526.37	369.83	213.28	56.73	37.20	37.20
M1053	1161.51	764.28	367.06	30.16	#DIV/0!	#DIV/0!	30.16
M1859	942.67	587.11	231.56	124.00	146.11	212.78	124.00
M1959	N/A	N/A	N/A	N/A	N/A	N/A	N/A
M2258	343.81	178.63	13.45	58.91	128.11	169.63	13.45
M2417	114.17	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	114.17
M3407	54.24	31.38	8.53	14.32	37.18	50.89	8.53
M3409	620.74	481.25	341.76	202.27	62.78	20.91	20.91
						Average	42.25 days

The average mean absolute errors in the last 30 days before the failure are found as 71 days and 43 days for the second and the third approaches, respectively. Therefore 3rd approach overflanks the second approach by predicting RUL more accurately. Moreover, in order to compare with the first approach, RUL predictions corresponding to the specified threshold level are clustered in the previously defined groups (see chapter 5.3). Combined decision tree prediction accuracy is higher than the third approach which provides 53 % and 56% TP rate and Precision respectively on the test data.

Table 28: Results of the Third Approach

Classification on the training data

Training Data		Predicted Class A	Predicted Class B	Predicted Class C
Actual Class	A	1214	142	103
	B	61	95	61
	C	23	34	127

Accuracy: 77%

Classification on the test data:

Test Data		Predicted Class A	Predicted Class B	Predicted Class C
Actual Class	A	122	14	14
	B	43	24	6
	C	21	23	38

Accuracy: 60%

	TP Rate (%)	FP Rate (%)	Precision (%)
A	0.813333	0.186667	0.655914
B	0.328767	0.671233	0.393443
C	0.463415	0.536585	0.655172
Weighted Avg.	0.535172	0.464828	0.568176

Six different values were tried to specify the unknown threshold level. However, if more than 6 values are tried for the threshold level, the results will be significantly improved. By assuming that the threshold level is known for each case (which is equal to actual value), RUL of the module have been computed, following results (Table 29) have been obtained.

Table 29: Results of the Third Approach with given Threshold Level

Classification on the training data

Training Data		Predicted Class A	Predicted Class B	Predicted Class C
Actual Class	A	1306	135	18
	B	62	118	37
	C	24	17	144

Accuracy: 84%

Classification on the test data:

Test Data		Predicted Class A	Predicted Class B	Predicted Class C
Actual Class	A	122	26	2
	B	29	42	2
	C	21	17	44

Accuracy: 68%

⇩

	TP	FP	Precision
A	0.813333	0.186667	0.709302
B	0.575342	0.424658	0.494118
C	0.536585	0.463415	0.916667
Weighted Avg.	0.641754	0.358246	0.706696

RUL was predicted for each machine and the absolute error between the actual and the predicted RUL was calculated during last 30 days before the failure. Although predictions are not accurate for 6 machines which are highlighted in the Table 30, the average error between the actual and the predicted RUL is about 15 days.

Table 30: Mean Absolute Error of RUL Predictions

Training Data

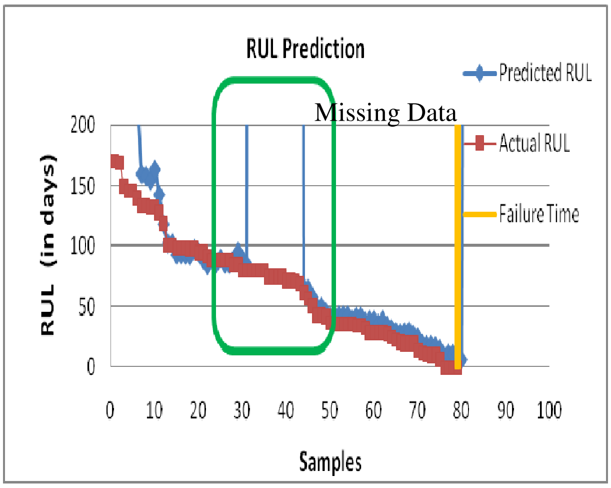
	Sample Size	Threshold	Mean Absolute Error (last 30 days)
M0005	32	-9	5.019
M0006	149	-7.6	124.586
M0017	94	-9.5	1.952
M0018	152	-8.4	2.554
M0021	127	-7	5.277
M0029	90	-5.8	48.270
M0034	68	-7	5.327
M0041	26	-9.3	1.884
M0051	155	-6.9	7.321
M0067	128	-8	95.084
M0741	79	-9.3	3.069
M1004	99	-9.6	3.377
M1186	46	-7.4	7.005
M1321	71	-8.4	10.983
M1358	72	-7.1	18.301
M1771	93	-7.3	3.379

Test Data

	Sample Size	Threshold	Mean Absolute Error (last 30 days)
M0019	97	-9.6	7.849
M0067	42	-8.5	16.532
M0885	86	-9.5	38.748
M0939	78	-9.4	6.815
M1053	26	-8	30.163
M1859	23	-7.7	17.333
M1959	31	-4	xxxxx
M2258	26	-7.1	7.552
M2417	75	-4.7	16.070
M3407	61	-7.5	2.935
M3409	74	-9.5	7.274
		Mean Absolute Error	**15.127 days**

M1828	90	-5.4	24.099
M1887	105	-9.6	32.515
M1937	51	-9.5	8.025
M1959	90	-6.7	1.377
M2232	74	-9.2	6.118
M2252	59	-6.2	10.213
M2417	151	-5.2	1.023
M2601	65	-7.8	7.740
M2683	68	-6.3	1.606
M2789	55	-6.5	0.435
M3083	226	-6.5	6.137
M3398	33	-8.7	14.138
M3407	184	-8.6	6.987
M3411	32	-9.6	3.478
	Mean Absolute Error		**15.576 days**

The output of this model is updated through online data. With each new record, failure status is reclassified and RUL is recalculated. Effects of the external factors on the parameters cannot be taken into account easily by the static models. However, as the proposed dynamic model modifies predictions with each record, it externalizes the changes in the system immediately. Besides, the effects of discontinuous changes which might be caused by missing or misrecorded data are smoothed in a short time. Figure 16 shows the effect of missing and misrecorded data. Although these situations cause to high error at those instants, the predictions are modified in the short period.



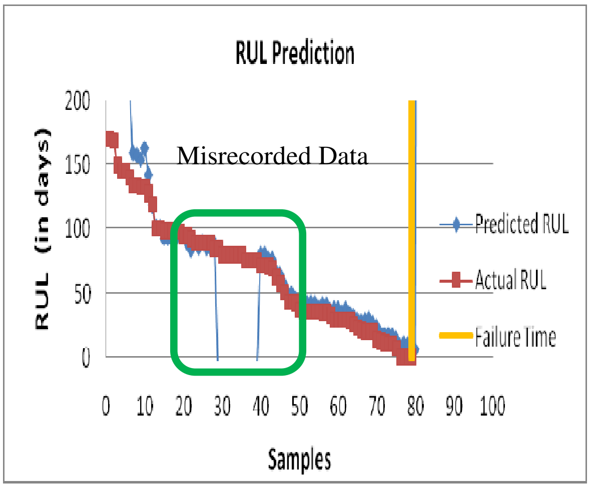


Figure 16: Effects of the Missing and Misrecorded Data on the Model

5.6. Assessment of the Models

Several models were developed on the full training data (2009 data) by using three distinct approaches. Created models were tested and ranked based on their weighted average TP rate, FP rate, Precision and Prediction Accuracy on the test data set (2010 data).

Firstly, data has been divided into three classes which indicate the distance to the failure time. Cluster A refers that the failure will not occur in the following 60 days, Cluster B refers that failure will occur in 60 days and Cluster C shows that the failure will occur in 20 days. To predict these classes, ANN-MLP, ANN-RBF, Decision Tree, Simple Cart and KNN methods have been used. The best model was obtained with the combination of two decision trees which results in 56% true classification and 63% precision. It should be pointed that classification of C cluster is important in terms of warning the user before the failure instant. The model output has been analyzed in detail to assess the warning capability. The model predicts 5 failure cases as cluster A or B at the failure instant. However for about 55% of the cases, it warns the user on time. Moreover the decision tree helps to diagnose the reason of the failure. The parameter that causes the failure could be detected by analyzing the decision tree steps.

Secondly, ANN-MLP, ANN RBF and linear regression models were developed to estimate the remaining time of Module X in days. Although MLP outperforms the other methods, it does not make satisfactory predictions. For the last 30 days before the failure, mean absolute error is found as 71 days which indicates the high deviation from actual RUL (<=30 days). Warning signal is produced for only 18 % of the upcoming failures. Hence user can not be warned for most of the failure cases. Besides, this model does not help to explain failure reasons.

Last but not least, data visualization technique was used to understand relationship between the parameters and the failure events. This understanding was translated into mathematical formulation which leads to diagnostics and prognostics of the failure. Since the actual threshold level is unknown, several values were tried and the threshold level results in minimum mean absolute error was selected for each case. This approach predicts with higher accuracy than the second approach. However, the combined decision tree model outperforms this mathematical model. By assuming that the threshold level is known for each case, the calculations were performed again. Although for 6 failure cases (4 from training data and 2 from test data), the remaining useful time is predicted with higher error, the model achieves to warn the user about 82 % (9 cases out of 11 cases) of the upcoming failures with an accurate indication of RUL in comparison with the other approaches. Significant increase in TP rate and Precision was observed.

Known threshold level increased prediction accuracy for the third approach. In order to analyze the effect of the given threshold level on second approach (MLP model), another model has been developed (see Appendix X). However given threshold level does not improve the RUL predictions of the MLP model.

Table 31 shows the results of three approaches. The first three columns indicate the warning capability of the models at the failure instants. Since the model should predict the class as C at the failure instant, the higher percentage of C predictions indicates better model. On the other hand, prediction of cluster A leads to missing failures. The latter three columns in the table depict the detailed prediction accuracy of the models.

Table 31: Comparison of Three Approaches

		Failure Instant Classification			Detailed Prediction Accuracy		
		A	**B**	**C**	**TP**	**FP**	**Precision**
1st Approach (combined decision tree)		0.364 (4/11)	0.091 (1/11)	0.545 (6/11)	0.562	0.438	0.635
2nd Approach (MLP)		0.454 (5/11)	0.362 (4/11)	0.182 (2/11)	0.463	0.537	0.334
3rd Approach	Unknown Threshold Level	0.362 (4/11)	0.182 (2/11)	0.454 (5/11)	0.535	0.464	0.568
	Known Threshold Level	0.091 (1/11)	0.091 (1/11)	0.818 (9/11)	0.642	0.358	0.707

To conclude, the third approach with the given threshold level outflanks the other approaches and predicts failure with higher accuracy. If the threshold level is not specified, it is recommended to use combined decision tree which predicts 50% of the upcoming failures.

5.7. Conclusion

In the first part of this section, failure threshold level was analyzed and according to the threshold limits, different failure classes were defined. Then the environmental factors were assessed in terms of their representativeness. Because of lack of samples, it was decided not to use environmental factors such as machine type, site id and customer id in the modelling.

Next, several models were developed to estimate failure time of Module X. Their prediction accuracy was evaluated. As a result, it was found that the third approach with the given threshold level provides highest prediction accuracy. Besides, the model helps to diagnose faulty subcomponent.

6. Comparing the Data Driven Model with the Physical Model

Domain experts in ASML have already developed a physical model to predict the upcoming failure by using mechanistic knowledge and theories related to Module X. In the previous section, the data driven model was developed without any system knowledge. In this section, the performance of the data driven model is compared with that of the physical model.

The physical model was built with Matlab code. The required inputs for the model are the 36 condition parameters and the machine type. It provides the remaining time of the sub-modules of Module X in weeks. It was processed with the raw data of specified machines which were used to build and test the data driven model. The model produced biweekly reports which show the failure status of Module X for each machine. In these reports, the failure status of 6 different sub-modules and remaining weeks to the failure for each sub-module are shown. Health status of Module X is determined based on the health status of the sub-module which has the lowest value and the least time to the failure.

The health status of Module X is denoted by F1, F2, F3 and F4 which are explained below.

F1: one or more sub-modules has a value below the certain level.

F2: one or more sub-modules will fail within 10 weeks.

F3: Both F1 and F2 occur.

F4: non -failure

Data driven model was developed by considering six parameters and one of the parameters was specified as the dominant parameter which has distinctive effect on the failure of Module X. Selected six parameters correspond to the 6 different sub-modules. Besides, dominant parameter refers to the sub-module which has the lowest value and the least time to the failure. As a result, precious and accurate knowledge in line with physical theories was discovered by data mining methods.

The performance of the physical model was evaluated after adjusting the model output format. F1, F2 and F3 were specified as failure class whereas F4 was defined as non failure class in order to provide consistency. The physical model gives fewer warning signal than the proposed data driven model (DDM). Fewer signals may not be problem if it warns the user before the failure. However, the model does not predict remaining useful life for 6 cases (Table 32). Therefore the models cannot be completely compared.

Table 32: Results of PM Model and DDM Model

	Failure Instant Classification			Detailed Prediction Accuracy		
	A	**B**	**C**	**TP**	**FP**	**Precision**
PM	0.545 (6/11)	0.182 (2/11)	0.273 (3/11)			
3rd Approach	0.091 (1/11)	0.091 (1/11)	0.818 (9/11)	0.642	0.358	0.707

Only for 5 cases out of 11 cases, RUL predictions of the PM model are reasonable and comparable with predictions of the DDM model. Figure 17 illustrates the two of these cases.

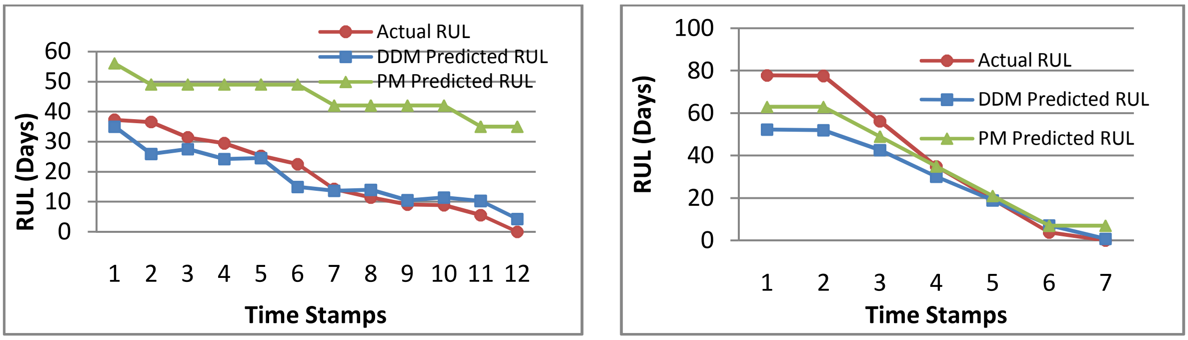


Figure 17: Comparison of the RUL Predictions

Models were also evaluated in terms of false alerts. False alerts means that the model produces warning signal although there will not be any failure in 60 days. Both DDM and PM models do not generate false alerts for 11 failure cases. Therefore they avert over-maintenance and associated costs and down time.

To sum up, in this chapter, the outputs of the physical model has been discussed and compared with the data driven model. Data driven model provides more accurate predictions than the physical model. Besides, the knowledge gained from data driven model has been validated with the principles of the physical model.

7. Deployment

In this chapter, decision support model which is the integration of the knowledge gained from prediction model with ASML decision making process is explained. Furthermore user interface of the prediction model is presented.

Decision Support Model

ASML provides condition based maintenance to all customers during warranty period (first 2 years) and to the customers who have the service contract for preventive maintenance after the warranty period. The decision support model consists of two processes.

Monitoring Process: While the prediction model is running, its outputs are monitored continuously. These outputs are listed in the real time personalized Web pages. However in the case of RUL<20 days, field service engineer is also notified by mail. This notification shows the remaining useful life of Module X and faulty sub-modules.

Maintenance Planning Process: When the field service engineer is notified about the forthcoming failure, he starts to plan maintenance. First he discusses with the customer about the timeline. Accordingly, he arranges labors and spare parts. Then maintenance is executed. At the end of this step, the details about maintenance and failure are reported.

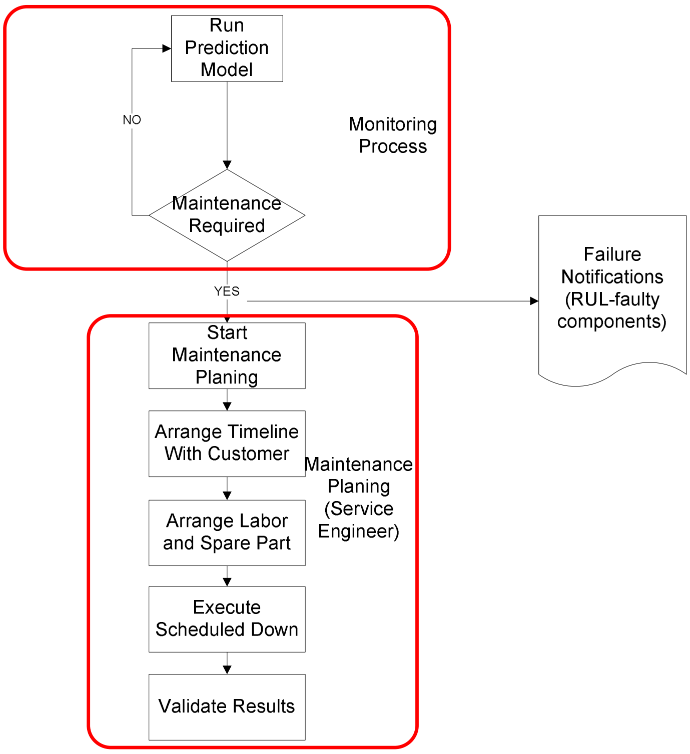


Figure 18: Decision Support Model

The performance of the prediction model is monitored regularly. If the model doesn't generate notification or if it generates false signals, model improvement would be required. Field service engineers, modelling team and development team are responsible for the model building/updating. They design, validate and release new models.

User Interface

The user interface by which ASML can use the output of the prediction model was developed with MS Excel tool. The interface aims to show the status of Module X and the remaining time before the failure. Therefore the service engineer can be informed about the upcoming failure.

Firstly, it is requested to enter the machine number and the threshold level. Secondly the parameter values (P1...P6) and the time stamp are entered. The interface shows the sample size (number of entered records for a machine), parameters in degradation, dominant parameter, health status of Module X, Failure Class, RUL of Module X (if it is in failure period), and maintenance diagnosis which indicates whether the maintenance is performed. According to the RUL of each sub-module, the user is warned about the faulty sub-modules and necessary repair process. As long as the second step is repeated, the new output and the new warning signals are provided.

If the threshold level has not been specified, the model should run to determine the level at which machine performance is not acceptable. By step by step decreasing the threshold value from -4 to -9.6, the level at which the customer is dissatisfied with the module performance is decided. For example firstly the threshold level can be determined as -5. When the dominant parameter reaches this level (when RUL<10 days), module performance is checked, if the customer is satisfied with the performance, another threshold level such as "-6" is defined. This process is repeated until the customer settles the level where the maintenance is required. After specifying the threshold level, the model can run without any interruptions and indicate the remaining time of the module.

It should be remarked that fewer data sample may lead to the large prediction errors. The predictions are improved and approximated to the actual RUL as more data is considered (see Figure 15). Therefore, it could be better to consider the outputs when the sample size is larger than 10.

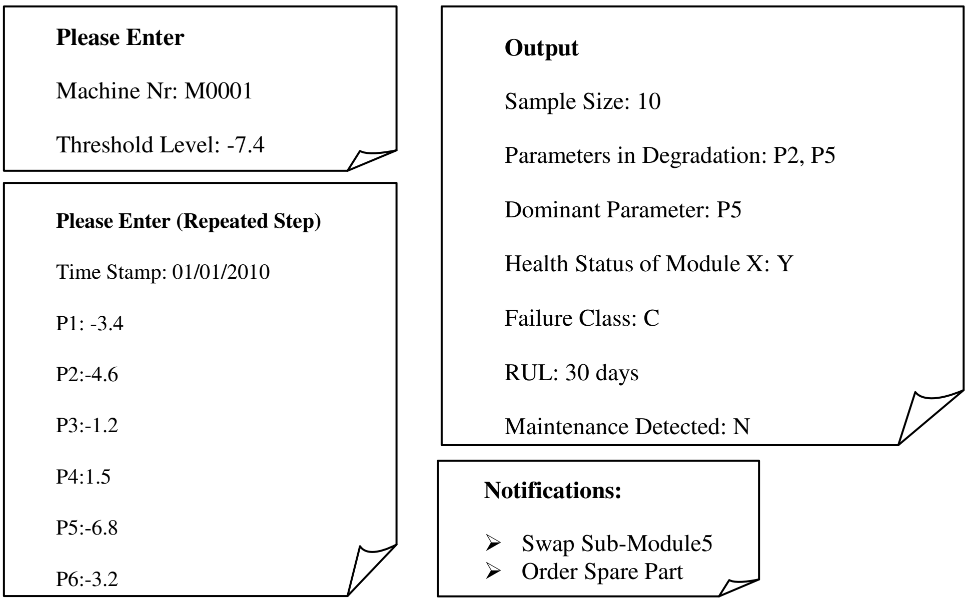


Figure 19: Sample Format of the User Interface

8. Conclusion

This chapter summarizes the main findings provided in the previous chapters. Moreover it describes limitations and provides recommendations. Last but not the least, some future research options are explained.

8.1. Main Findings, Limitations and Recommendations

In this project, a failure prediction model for Module X has been developed by using the condition monitoring data in line with ASML objectives. Business success criteria (see Chapter 2.1) of ASML have been considered to evaluate the main findings.

Utility of Local Monitoring Data

The utility of the local monitoring data has been assessed in the project. Although given data helps to predict the failure, its applicability is limited.

(1) Limitation on the Condition Data: Imperfect condition monitoring is the main limitation for the research. This results in lack of the condition data and non-periodic monitoring intervals between the samples.

(2) Limitation on the Event Data: To understand the parameter-failure relation in depth, more information is required. All data about the taken actions such as the exact failure time of Module X and other machine components, performed maintenance actions and their coverage, machine intermittent and machine performance problems should be stated clearly to understand and interpret changes in the condition parameters and accordingly to predict the failures.

Success of the local monitoring is indisputable. The failures are predicted despite of the limitations. Furthermore if more complete and accurate data is provided, the model can be improved.

Discovering System Knowledge through Data Mining

System knowledge was discovered through data mining methods and the accuracy of the gained knowledge was validated by the physical model. This situation supports sufficiency and applicability of the data mining techniques to model the systems about which we do not have extensive knowledge.

Decision Support Model

The validated failure prediction model was developed in this project. The model outperforms the physical model which can not produce accurate predictions and warn the customer on time. It is recommended to the use the data driven model which warns the user about 82% of the upcoming failures by indicating the remaining time of Module X accurately. Therefore unscheduled down time can be decreased significantly. Moreover prediction of the upcoming failure contributes to executing maintenance right on time and to eliminate over-maintenance because the maintenance decision is based on degradation of Module X.

ASML implements maintenance periodically or in case of a failure. Under both circumstances, the service engineer is unaware of the faulty sub-modules. Maintenance may be performed to recover specific components even if some other sub-modules are about to fail or complete maintenance may be performed even if only one component is faulty. Thanks to condition monitoring, a service engineer can control the degradation level of all of the sub-modules and accordingly the coverage of maintenance is specified. Therefore maintenance is performed based on necessity of the components. This leads to a decrease in maintenance expenditures.

The model indicates the remaining useful time, that makes the engineer decide about when to plan maintenance, when to arrange labor, and when to order a spare part. These steps are performed on-time and cost effectively based on the predicted failure time.

As a result, despite of the limitations, a data driven prediction model is a very favorable model for maintenance planning. By means of this model, ASML can provide better maintenance solutions to customers, leading to increased system availability and decreased associated costs.

8.2. Future Prospects

Model Extension for Multiple Component Setting

The model was developed for a single component (Module X) of the Machine. In order to predict the machine failure, the decision model should be extended for other components of the machine.

Control of Condition Monitoring Interval

Condition monitoring can be continuous or periodic. Since continuous monitoring could be expensive and gives inaccurate information, periodic monitoring is recommended as an effective approach. However, optimal condition monitoring interval should be determined based on cost and wear rate.

Determination of Optimal Replacement Time

Maintenance decision support model indicates the upcoming failure and the severity of the failure however the service engineer should decide about the optimum time for the maintenance by considering labor availability, maintenance urgency, spare part availability and machine necessity. Therefore another research direction could be to develop maintenance optimization model.

Improvement of the Model for Logistic Decisions

Prediction of the machine failure helps to take effective decisions about spare part inventory. To improve a proactive maintenance model with proactive logistics could be another research topic.

References

ASML internet site, Company about ASML, ASML Profile, http://www.asml.com/asml/show.do?ctx=272&rid=362 Last retrieved on 15.02.2011.

ASML intranet site. Last retrieved on 15.02.2011.

Blechertas, V., Bayoumi, A., Goodman, N., Shah, R., Shin, Y. (2009) CBM Fundamental Research at the University of South Carolina: A Systematic Approach to U.S. Army Rotorcraft CBM and the Resulting Tangible Benefits, Proceedings of AHS International Specialists' Meeting on Condition Based Maintenance

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). “Classification and Regression Trees”, Wadsworth International Group, Belmont, California.

British Standards Institution (1984) BS3811 Glossary of maintenance terms in Terotechnology, BSI, London.

Cakir, G. S. (2011). Implementation of Proactive Maintenance Policy in ASML: Literature study, in Operations Management and Logistics, Eindhoven University of Technology.

Collacott, R.A. (1997) Mechanical fault diagnosis and condition monitoring, Chapman and Hall Ltd., London

Cross Industry Standard Process for Data Mining (CRISP-DM) internet site, http://www.crisp-dm.org. Last retrieved on 15.05.2011.

Drazin, S., Montag, M. Decision Tree Analysis using Weka Machine Learning Project II, University of Miami

Gilmartin, B.J., Bongort K., Engel, A.H.S.J. (2000) Prognostics, The Real Issues Involved With Predicting Life Remaining, in Proceedings IEEE Aerospace Conference, vol. 6, pp. 457-469.

Hair, J. F., Black, W. C., Babin, B. J., Andersin, R.E. (2009), Multivariate Data Analysis, 7th edition, Prentice Hall

Han, J., Kamber, M.(2006), Data Mining: Concepts and Techniques, Elsevier, 2006.

Hoyle, W.C.C, Mehr, A., Tumer I. (2007) "On quantifying cost benefit of ISHM in aerospace systems," in ASME 2007 International Design Engineering Technical Conferences, pp. 1-10

Jardine, A. K.S., Lin, D., Banjevic, D. (2006). A review on machinery diagnositcs and prognostics implementing condition-based maintenance. Elsevier Ltd. , 1483-1510

Kent, R.M., Murphy, D.A. (2000) Health Monitoring System Technology Assessments Cost Benefits Analysis," Hampton, Virginia.

Lee, J., Ni, J., Djurdjanovic, D., Qiu, H., Liao, H. (2006) Intelligent prognostics tools and e-maintenance, Computers in Industry, Volume 57, Issue 6, E-maintenance Special Issue, Pages 476-489.

Li, Y.G., Nilkitsaranont, P. (2009) Gas turbine performance prognostic for condition-based maintenance, Applied Energy, Vol. 86, Iss. 10, pp 2152-2161

Peng, Y., Dong, M., Zuo, M. (2010) Current status of machine prognostics in condition-based maintenance: a review, The International Journal of Advanced Manufacturing Technology, Volume 50, Numbers 1-4, pp. 297-313(17)

Quinlan, R. (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA.

Rao, B. (1996) Handbook of condition monitoring, Amsterdam,: Elsevier

Shearer, C. (2000) The CRISP-DM Model: The New Blueprint for Data Mining, Journal of Data Warehousing

Tsang, A.H.C., Jardine, A. K.S., Kolodny, H. (1999) Measuring maintenance performance: a holistic approach, International Journal of Operations and Production Management volume 19, Iss. 1, pp 691-715.

Data Mining: What is Data Mining, University of California internet site: http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.html Last retrieved on 15.02.2011

Van Aken, J. E., Berends, H., and Van der Bij, H. (2007). Problem-solving in Organizations: A Methodological Handbook for Business Students. *University press Cambridge*.

Vismara, M.G. (2010) An Integrated Approach to a Condition Based Maintenance policy and applications, Politecnico Di Milano.

Zhao Z., Wang F., Jia M., Wang S. (2010) Predictive maintenance policy based on process data, Chemometrics and Intelligent Laboratory Systems Volume 103, Iss 2, pp 137-143.

Appendices

Appendix I. Data Preparation-Data Analysis

SAMPLE OF THE GIVEN DATA FORMAT

Data sets were provided as a list of independent records.

MACHINE NUMBER	TIME STAMP	VALUE	MACHINE TYPE	SITE ID	CUSTOMER CONTINENT	CUSTOMER COUNTRY	CUSTOMER NUMBER	PARAM ID
M1297	17-Dec-09	-8.856	T0010	1288	Asia	South Korea	188	3756
M2572	22-Oct-09	-8.9597	T0005	665	Asia	Singapore	2046	990
M2488	30-Jul-09	-3.9977	T0083	755	Other	Other	OT01	981
M0822	14-Jul-09	-4.0141	T0016	1284	Asia	South Korea	188	960
M1621	08-May-09	-3.8854	T0010	1294	Asia	South Korea	1146	957
M1647	23-Oct-09	-3.9167	T0001	277	North America	USA	196	966
M0003	21-Jul-09	-3.873	T0010	1291	Asia	South Korea	188	990
M0004	21-Feb-09	-3.8264	T0010	1291	Asia	South Korea	188	966
M2862	27-Aug-09	-3.7398	T0004	629	Asia	Taiwan	222	993
M2631	06-Jan-09	-8.551	T0004	801	Europe	France	192	972
M1022	10-Jul-09	-8.9597	T0020	1163	North America	USA	188	960
M2108	10-Jul-09	-3.9977	T0005	613	North America	USA	558	969
M1141	10-Aug-09	-6.8885	T0011	1290	Asia	South Korea	1146	966
M3241	22-Apr-09	-8.551	T0010	629	Asia	Taiwan	222	963
M0051	05-Sep-09	-8.9597	T0008	1178	Asia	Taiwan	386	996
M1171	28-Feb-09	-3.9977	T0006	629	Asia	Taiwan	222	987
M1171	12-Aug-09	-6.8885	T0006	629	Asia	Taiwan	222	990
M1614	04-Dec-09	-8.551	T0007	1284	Asia	South Korea	188	990
M1951	16-Jul-09	-8.9597	T0019	1286	Asia	South Korea	188	960
M2785	05-Jul-09	-3.9977	T0010	1291	Asia	South Korea	188	996

SAMPLE OF THE ALIGNED DATA FORMAT

Data sets were aligned in order to understand the changes in the condition parameters in time.

Categorical parameters + 36 numeric condition parameters

MACHINE NUMBER	TIMESTAMP	MACHINE TYPE	CUSTOMER ID	P955	P956	P957	P958	P959	P960	P961
M0005	07-Jan-09	T0007	C1058	-6.8961	-6.8860	-6.8910	-6.9177	-6.8542	-6.8860	-3.7979
M0005	13-Feb-09	T0007	C1058	-7.3892	-7.3831	-7.3862	-7.4487	-7.3123	-7.3805	-4.3121
M0005	16-Feb-09	T0007	C1058	-7.4847	-7.4738	-7.4792	-7.5021	-7.4451	-7.4736	-4.3567
M0005	17-Feb-09	T0007	C1058	-7.5400	-7.5320	-7.5360	-7.5974	-7.4640	-7.5307	-4.3823
M0005	19-Feb-09	T0007	C1058	-7.5305	-7.5201	-7.5253	-7.5479	-7.4915	-7.5197	-4.3793
M0005	01-Mar-09	T0007	C1058	-7.6871	-7.6767	-7.6819	-7.7032	-7.6489	-7.6761	-4.5276
M0005	07-Mar-09	T0007	C1058	-7.7783	-7.7686	-7.7734	-7.7954	-7.7414	-7.7684	-4.6072
M0005	10-Mar-09	T0007	C1058	-7.7222	-7.7109	-7.7165	-7.7396	-7.6819	-7.7107	-4.6219
M0005	06-Apr-09	T0007	C1058	-8.0890	-8.0804	-8.0847	-8.1049	-8.0527	-8.0788	-4.9698
M0006	06-Apr-09	T0007	C1058	-7.9700	-7.9602	-7.9651	-7.9858	-7.9337	-7.9597	-4.9725
M0006	22-Apr-09	T0007	C1058	-8.1674	-8.1635	-8.1654	-8.2214	-8.0994	-8.1604	-5.1675
M0006	01-Jun-09	T0007	C1058	-8.5568	-8.5504	-8.5536	-8.5730	-8.5239	-8.5484	-5.6113
M0006	09-Jun-09	T0007	C1058	-8.5599	-8.5553	-8.5576	-8.6090	-8.4949	-8.5519	-5.6860
M0006	22-Jul-09	T0007	C1058	-8.7384	-8.7350	-8.7367	-8.7869	-8.6755	-8.7312	-6.2033
M0006	24-Jul-09	T0007	C1058	-8.8330	-8.8257	-8.8293	-8.8461	-8.8004	-8.8232	-6.1697
M0006	14-Aug-09	T0007	C1058	-8.9163	-8.9142	-8.9153	-8.9639	-8.8553	-8.9096	-6.3815
M0006	14-Aug-09	T0007	C1058	-8.9392	-8.9371	-8.9381	-8.9871	-8.8776	-8.9323	-6.3531
M0006	16-Aug-09	T0007	C1058	-8.8599	-8.8574	-8.8586	-8.9078	-8.7988	-8.8533	-6.4169

Chronological Order

.

Principal Component Analysis

Table 33: Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
P955	2738	-9.5831	3.3923	-1.755095	2.8740200
P956	2738	-9.5782	3.2596	-1.782281	2.8867572
P957	2738	-9.5807	3.2651	-1.768688	2.8791535
P958	2738	-9.6075	3.2425	-1.746095	2.8264964
P959	2738	-9.5596	3.5233	-1.791696	2.9422333
P960	2738	-9.5836	3.2681	-1.768896	2.8799768
P961	2738	-8.6859	2.7200	-1.730782	2.8300853
P962	2738	-8.6826	2.7277	-1.684138	2.8326117
P963	2738	-8.6842	2.7235	-1.707460	2.8291627
P964	2738	-8.6673	2.7777	-1.728918	2.7860550
P965	2738	-8.6929	2.6773	-1.683816	2.8801639
P966	2738	-8.6801	2.7274	-1.706367	2.8279740
P967	2738	-9.5355	2.9391	-1.983609	2.6068693
P968	2738	-9.5462	2.9428	-1.969870	2.6053008
P969	2738	-9.5409	2.9410	-1.976740	2.6047331
P970	2738	-9.5651	3.1064	-1.943875	2.5675943
P971	2738	-9.5682	3.0151	-2.010192	2.6466408
P972	2738	-9.5393	2.9419	-1.977032	2.6028328
P979	2738	-7.8873	3.5294	-2.347382	2.7141283
P980	2738	-7.8906	3.5544	-2.394532	2.7073025
P981	2738	-7.8889	3.5419	-2.370957	2.7085981
P982	2738	-7.9202	3.4833	-2.419629	2.7000366
P983	2738	-7.8607	3.6026	-2.322263	2.7209915
P984	2738	-7.8905	3.5429	-2.370947	2.7073187
P985	2738	-9.5123	3.2944	-2.952895	3.1781294
P986	2738	-9.6146	3.3334	-2.847936	3.2572150
P987	2738	-9.5634	3.3139	-2.900415	3.2150113
P988	2738	-9.5547	3.2684	-2.837495	3.2553084
P989	2738	-9.5795	3.3661	-2.960867	3.1797931
P990	2738	-9.5671	3.3173	-2.899180	3.2152049
P991	2738	-9.1248	2.2849	-1.976129	2.5244264
P992	2738	-9.1248	2.3166	-2.071046	2.5117088
P993	2738	-9.1248	2.2984	-2.023587	2.5162935
P994	2738	-9.1418	2.2437	-1.997085	2.5602084
P995	2738	-9.0903	2.3737	-2.049275	2.4758564
P996	2738	-9.1161	2.2954	-2.023181	2.5146595
Valid N (listwise)	2738				

Table 34: Total Variance Explained

Component	Initial Eigenvalues			Rotation Sums of Squared Loadings[a]
	Total	% of Variance	Cumulative %	Total
1	20.210	56.139	56.139	15.773
2	6.623	18.396	74.535	9.730
3	3.262	9.062	83.597	12.457
4	2.224	6.178	89.775	12.079
5	2.117	5.881	95.656	14.517
6	1.519	4.220	99.876	14.600
7	.015	.043	99.919	
8	.011	.031	99.950	
9	.008	.021	99.971	
10	.004	.010	99.981	
11	.002	.006	99.987	
12	.002	.005	99.992	
13	.001	.003	99.996	
14	.001	.002	99.997	
15	.000	.001	99.998	
16	.000	.001	99.999	
17	.000	.001	100.000	
18	6.364E-5	.000	100.000	
19	3.352E-6	9.312E-6	100.000	
20	2.299E-6	6.387E-6	100.000	
21	2.040E-6	5.667E-6	100.000	
22	1.483E-6	4.119E-6	100.000	
23	8.456E-7	2.349E-6	100.000	
24	6.749E-7	1.875E-6	100.000	
25	1.352E-10	3.755E-10	100.000	
26	1.296E-10	3.599E-10	100.000	
27	1.219E-10	3.386E-10	100.000	
28	1.199E-10	3.331E-10	100.000	
29	1.189E-10	3.302E-10	100.000	
30	1.107E-10	3.076E-10	100.000	
31	1.035E-10	2.875E-10	100.000	
32	1.026E-10	2.851E-10	100.000	
33	9.832E-11	2.731E-10	100.000	
34	9.293E-11	2.581E-10	100.000	
35	7.894E-11	2.193E-10	100.000	
36	7.652E-11	2.126E-10	100.000	

Table 35: Pattern Matrix

	Component					
	1	2	3	4	5	6
P955	.996	-.006	-.001	-.002	.003	.008
P956	1.001	.007	.002	.004	-.002	-.007
P957	.999	.000	.000	.001	.000	.000
P958	.993	-.019	.001	.004	.016	-.003
P959	1.001	.018	.000	-.004	-.014	.003
P960	.999	.000	.001	.000	.001	.000
P961	.009	.010	.001	.004	-.009	.994
P962	-.009	-.010	.000	-.004	.009	1.003
P963	.000	.000	.000	.000	.000	1.000
P964	-.006	.002	.001	.013	-.005	.999
P965	.006	-.002	.000	-.013	.005	.996
P966	.000	.000	.000	.000	.000	.999
P967	.004	.007	1.000	.002	.000	-.008
P968	-.002	-.006	.997	-.002	.000	.009
P969	.001	.001	.999	.000	.000	.000
P970	-.025	-.020	1.012	.005	.016	-.003
P971	.026	.020	.984	-.004	-.016	.004
P972	.001	.000	.999	.000	.000	.000
P979	-.002	1.000	-.001	.004	-.002	-.008
P980	.002	.997	.002	-.003	.003	.008
P981	.000	1.000	.000	.001	.000	.000
P982	-.006	1.000	.000	-.003	.014	-.013
P983	.007	.997	.000	.003	-.013	.013
P984	.000	1.000	.000	.000	.000	.000
P985	.000	-.001	.001	.996	.006	-.002
P986	.000	.001	-.001	1.001	-.006	.002
P987	.000	.000	.000	1.000	.000	.000
P988	-.001	-.008	.002	.998	.006	.008
P989	.002	.009	-.002	1.000	-.005	-.008
P990	.000	.000	.000	1.000	.000	.000
P991	.001	-.001	-.002	.005	.995	.002
P992	-.001	.001	.002	-.004	1.001	-.001
P993	.000	.000	.000	.001	.999	.000
P994	.000	-.007	.002	.018	.990	.001
P995	.001	.008	-.002	-.018	1.006	.000
P996	.000	.000	.000	.000	.999	.001

Extraction Method: Principal Component Analysis.

Rotation Method: Oblimin with Kaiser Normalization.

a. Rotation converged in 10 iterations.

Table 36:Correlation Matrix

CORRELATION	P955	P956	P957	P958	P959	P960	P961	P962	P963	P964	P965	P966	P967	P968	P969	P970	P971	P972
P955	1	.998**	1.000**	.999**	.998**	1.000**	.737**	.729**	.734**	.730**	.734**	.734**	.625**	.626**	.626**	.611**	.638**	.626**
P956	.998**	1	1.000**	.997**	.999**	1.000**	.731**	.720**	.726**	.725**	.725**	.726**	.626**	.624**	.625**	.609**	.639**	.625**
P957	1.000**	1.000**	1	.998**	.999**	1.000**	.735**	.725**	.730**	.728**	.730**	.730**	.626**	.625**	.626**	.610**	.639**	.626**
P958	.999**	.997**	.998**	1	.994**	.998**	.734**	.727**	.731**	.729**	.731**	.731**	.626**	.626**	.626**	.614**	.637**	.626**
P959	.998**	.999**	.999**	.994**	1	.999**	.733**	.720**	.727**	.725**	.728**	.727**	.625**	.623**	.625**	.606**	.640**	.624**
P960	1.000**	1.000**	1.000**	.998**	.999**	1	.735**	.725**	.730**	.728**	.730**	.730**	.627**	.626**	.626**	.611**	.639**	.626**
P961	.737**	.731**	.735**	.734**	.733**	.735**	1	.997**	.999**	.999**	.996**	.999**	.597**	.603**	.600**	.591**	.607**	.600**
P962	.729**	.720**	.725**	.727**	.720**	.725**	.997**	1	.999**	.996**	.999**	.999**	.591**	.600**	.596**	.589**	.601**	.596**
P963	.734**	.726**	.730**	.731**	.727**	.730**	.999**	.999**	1	.998**	.998**	1.000**	.595**	.602**	.599**	.591**	.604**	.599**
P964	.730**	.725**	.728**	.729**	.725**	.728**	.999**	.996**	.998**	1	.993**	.998**	.594**	.599**	.597**	.590**	.602**	.597**
P965	.734**	.725**	.730**	.731**	.728**	.730**	.996**	.999**	.998**	.993**	1	.998**	.593**	.602**	.598**	.590**	.605**	.598**
P966	.734**	.726**	.730**	.731**	.727**	.730**	.999**	.999**	1.000**	.998**	.998**	1	.595**	.602**	.599**	.591**	.604**	.599**
P967	.625**	.626**	.626**	.626**	.625**	.627**	.597**	.591**	.595**	.594**	.593**	.595**	1	.998**	.999**	.997**	.999**	.999**
P968	.626**	.624**	.625**	.626**	.623**	.626**	.603**	.600**	.602**	.599**	.602**	.602**	.998**	1	.999**	.999**	.997**	.999**
P969	.626**	.625**	.626**	.626**	.625**	.626**	.600**	.596**	.599**	.597**	.598**	.599**	.999**	.999**	1	.998**	.998**	1.000**
P970	.611**	.609**	.610**	.614**	.606**	.611**	.591**	.589**	.591**	.590**	.590**	.591**	.997**	.999**	.998**	1	.993**	.998**
P971	.638**	.639**	.639**	.637**	.640**	.639**	.607**	.601**	.604**	.602**	.605**	.604**	.999**	.997**	.998**	.993**	1	.998**
P972	.626**	.625**	.626**	.626**	.624**	.626**	.600**	.596**	.599**	.597**	.598**	.599**	.999**	.999**	1.000**	.998**	.998**	1
P979	.318**	.331**	.324**	.311**	.335**	.324**	.196**	.175**	.186**	.191**	.180**	.186**	.210**	.196**	.203**	.185**	.220**	.203**
P980	.334**	.348**	.341**	.328**	.352**	.341**	.217**	.195**	.206**	.212**	.201**	.206**	.226**	.211**	.218**	.200**	.236**	.218**
P981	.326**	.340**	.333**	.320**	.344**	.333**	.206**	.185**	.196**	.201**	.191**	.196**	.218**	.204**	.211**	.192**	.228**	.211**
P982	.317**	.331**	.324**	.311**	.335**	.324**	.196**	.174**	.185**	.191**	.180**	.186**	.212**	.197**	.205**	.186**	.222**	.204**
P983	.335**	.348**	.342**	.329**	.352**	.341**	.217**	.196**	.207**	.212**	.201**	.207**	.224**	.210**	.217**	.199**	.234**	.217**
P984	.326**	.340**	.333**	.320**	.344**	.333**	.207**	.186**	.196**	.202**	.191**	.197**	.218**	.204**	.211**	.193**	.228**	.211**
P985	.449**	.457**	.453**	.452**	.452**	.453**	.388**	.375**	.382**	.390**	.373**	.382**	.269**	.261**	.265**	.258**	.272**	.265**
P986	.446**	.454**	.450**	.450**	.448**	.450**	.386**	.373**	.380**	.389**	.369**	.380**	.265**	.257**	.261**	.255**	.267**	.262**
P987	.448**	.456**	.452**	.451**	.451**	.452**	.387**	.374**	.381**	.390**	.371**	.381**	.267**	.259**	.263**	.256**	.270**	.264**
P988	.453**	.461**	.457**	.457**	.455**	.457**	.395**	.383**	.389**	.398**	.380**	.390**	.272**	.265**	.269**	.263**	.274**	.269**
P989	.442**	.451**	.447**	.445**	.446**	.446**	.378**	.365**	.372**	.381**	.362**	.372**	.261**	.253**	.257**	.249**	.265**	.258**
P990	.448**	.456**	.452**	.452**	.451**	.452**	.387**	.374**	.381**	.390**	.372**	.381**	.267**	.259**	.263**	.256**	.270**	.264**
P991	.628**	.626**	.628**	.634**	.620**	.628**	.602**	.603**	.603**	.602**	.602**	.603**	.466**	.464**	.465**	.464**	.465**	.465**
P992	.626**	.625**	.626**	.632**	.619**	.626**	.601**	.600**	.601**	.600**	.599**	.601**	.468**	.465**	.467**	.465**	.467**	.467**
P993	.627**	.626**	.627**	.633**	.620**	.627**	.602**	.602**	.602**	.602**	.601**	.602**	.467**	.465**	.466**	.464**	.467**	.466**
P994	.628**	.627**	.628**	.635**	.620**	.628**	.603**	.603**	.603**	.603**	.600**	.603**	.468**	.465**	.467**	.466**	.466**	.467**
P995	.626**	.624**	.625**	.630**	.618**	.625**	.598**	.600**	.600**	.598**	.599**	.600**	.465**	.463**	.464**	.461**	.465**	.464**
P996	.628**	.626**	.627**	.633**	.620**	.627**	.602**	.602**	.602**	.602**	.601**	.602**	.467**	.465**	.466**	.464**	.466**	.466**

CORRELATION	P979	P980	P981	P982	P983	P984	P985	P986	P987	P988	P989	P990	P991	P992	P993	P994	P995	P996
P955	.318**	.334**	.326**	.317**	.335**	.326**	.449**	.446**	.448**	.453**	.442**	.448**	.628**	.626**	.627**	.628**	.626**	.628**
P956	.331**	.348**	.340**	.331**	.348**	.340**	.457**	.454**	.456**	.461**	.451**	.456**	.626**	.625**	.626**	.627**	.624**	.626**
P957	.324**	.341**	.333**	.324**	.342**	.333**	.453**	.450**	.452**	.457**	.447**	.452**	.628**	.626**	.627**	.628**	.625**	.627**
P958	.311**	.328**	.320**	.311**	.329**	.320**	.452**	.450**	.451**	.457**	.445**	.452**	.634**	.632**	.633**	.635**	.630**	.633**
P959	.335**	.352**	.344**	.335**	.352**	.344**	.452**	.448**	.451**	.455**	.446**	.451**	.620**	.619**	.620**	.620**	.618**	.620**
P960	.324**	.341**	.333**	.324**	.341**	.333**	.453**	.450**	.452**	.457**	.446**	.452**	.628**	.626**	.627**	.628**	.625**	.627**
P961	.196**	.217**	.206**	.196**	.217**	.207**	.388**	.386**	.387**	.395**	.378**	.387**	.602**	.601**	.602**	.603**	.598**	.602**
P962	.175**	.195**	.185**	.174**	.196**	.186**	.375**	.373**	.374**	.383**	.365**	.374**	.603**	.600**	.602**	.603**	.600**	.602**
P963	.186**	.206**	.196**	.185**	.207**	.196**	.382**	.380**	.381**	.389**	.372**	.381**	.603**	.601**	.602**	.603**	.600**	.602**
P964	.191**	.212**	.201**	.191**	.212**	.202**	.390**	.389**	.390**	.398**	.381**	.390**	.602**	.600**	.602**	.604**	.598**	.602**
P965	.180**	.201**	.191**	.180**	.201**	.191**	.373**	.369**	.371**	.380**	.362**	.372**	.602**	.599**	.601**	.600**	.599**	.601**
P966	.186**	.206**	.196**	.186**	.207**	.197**	.382**	.380**	.381**	.390**	.372**	.381**	.603**	.601**	.602**	.603**	.600**	.602**
P967	.210**	.226**	.218**	.212**	.224**	.218**	.269**	.265**	.267**	.272**	.261**	.267**	.466**	.468**	.467**	.468**	.465**	.467**
P968	.196**	.211**	.204**	.197**	.210**	.204**	.261**	.257**	.259**	.265**	.253**	.259**	.464**	.465**	.465**	.465**	.463**	.465**
P969	.203**	.218**	.211**	.205**	.217**	.211**	.265**	.261**	.263**	.269**	.257**	.263**	.465**	.467**	.466**	.467**	.464**	.466**
P970	.185**	.200**	.192**	.186**	.199**	.193**	.258**	.255**	.256**	.263**	.249**	.256**	.464**	.465**	.464**	.466**	.461**	.464**
P971	.220**	.236**	.228**	.222**	.234**	.228**	.272**	.267**	.270**	.274**	.265**	.270**	.465**	.467**	.467**	.466**	.465**	.466**
P972	.203**	.218**	.211**	.204**	.217**	.211**	.265**	.262**	.264**	.269**	.258**	.264**	.465**	.467**	.466**	.467**	.464**	.466**
P979	1	.997**	.999**	.998**	.998**	.999**	.534**	.536**	.536**	.530**	.541**	.536**	.371**	.369**	.370**	.369**	.371**	.370**
P980	.997**	1	.999**	.998**	.998**	.999**	.538**	.537**	.538**	.532**	.544**	.538**	.384**	.384**	.384**	.383**	.384**	.384**
P981	.999**	.999**	1	.999**	.999**	1.000**	.537**	.537**	.537**	.531**	.543**	.537**	.378**	.376**	.377**	.376**	.378**	.378**
P982	.998**	.998**	.999**	1	.995**	.999**	.534**	.532**	.534**	.526**	.540**	.534**	.377**	.377**	.377**	.375**	.378**	.377**
P983	.998**	.998**	.999**	.995**	1	.999**	.538**	.541**	.540**	.535**	.544**	.540**	.378**	.375**	.377**	.377**	.377**	.377**
P984	.999**	.999**	1.000**	.999**	.999**	1	.537**	.537**	.537**	.531**	.543**	.537**	.378**	.377**	.378**	.377**	.378**	.378**
P985	.534**	.538**	.537**	.534**	.538**	.537**	1	.997**	.999**	.997**	.999**	.999**	.561**	.557**	.559**	.567**	.550**	.559**
P986	.536**	.537**	.537**	.532**	.541**	.537**	.997**	1	.999**	.999**	.997**	.999**	.555**	.547**	.552**	.562**	.540**	.552**
P987	.536**	.538**	.537**	.534**	.540**	.537**	.999**	.999**	1	.999**	.999**	1.000**	.559**	.552**	.556**	.565**	.545**	.556**
P988	.530**	.532**	.531**	.526**	.535**	.531**	.997**	.999**	.999**	1	.997**	.999**	.565**	.557**	.561**	.572**	.550**	.562**
P989	.541**	.544**	.543**	.540**	.544**	.543**	.999**	.997**	.999**	.997**	1	.999**	.552**	.547**	.550**	.558**	.540**	.550**
P990	.536**	.538**	.537**	.534**	.540**	.537**	.999**	.999**	1.000**	.999**	.999**	1	.559**	.552**	.556**	.565**	.545**	.556**
P991	.371**	.384**	.378**	.377**	.378**	.378**	.561**	.555**	.559**	.565**	.552**	.559**	1	.997**	.999**	.998**	.998**	.999**
P992	.369**	.384**	.376**	.377**	.375**	.377**	.557**	.547**	.552**	.557**	.547**	.552**	.997**	1	.999**	.998**	.998**	.999**
P993	.370**	.384**	.377**	.377**	.377**	.378**	.559**	.552**	.556**	.561**	.550**	.556**	.999**	.999**	1	.999**	.999**	1.000**
P994	.369**	.383**	.376**	.375**	.377**	.377**	.567**	.562**	.565**	.572**	.558**	.565**	.998**	.998**	.999**	1	.995**	.999**
P995	.371**	.384**	.378**	.378**	.377**	.378**	.550**	.540**	.545**	.550**	.540**	.545**	.998**	.998**	.999**	.995**	1	.999**
P996	.370**	.384**	.378**	.377**	.377**	.378**	.559**	.552**	.556**	.562**	.550**	.556**	.999**	.999**	1.000**	.999**	.999**	1

**. Correlation is significant at the 0.01 level (2-tailed).

Appendix II. Threshold level

Table 37 shows the parameter values at the failure instant. The threshold level is determined according to the minimum parameter value at that time.

Table 37: Parameter Values at the Failure Instant

MACHINE NR	Machine Type	Site ID	P1	P2	P3	P4	P5	P6	Threshold Level =Min(P1...P6)
M0005	T0007	S1058	-8.960	-6.435	0.158	1.048	-8.534	-2.277	-8.960
M0006	T0009	S1243	-3.071	-1.329	-2.444	-7.354	-5.994	-1.616	-7.354
M0017	T0007	S1178	-6.784	-3.120	-5.332	-5.012	-9.392	-5.267	-9.392
M0018	T0006	S1289	-2.705	-8.336	-3.801	-1.913	-3.090	0.297	-8.336
M0021	T0014	S159	1.974	1.572	-6.784	2.339	2.387	0.867	-6.784
M0029	T0007	S1177	-5.449	-2.057	-2.452	-2.063	-5.756	-1.987	-5.756
M0034	T0008	S1199	0.582	1.004	0.134	3.194	-6.816	2.260	-6.816
M0041	T0011	S1058	-1.975	-1.904	-0.944	-5.531	-9.275	-1.702	-9.275
M0051	T0008	S1178	1.762	1.240	0.261	-5.712	-6.884	-0.160	-6.884
M0067	T0009	S1243	-4.524	-6.386	-3.221	-5.712	-7.381	-6.389	-7.381
M0741	T0007	S992	-2.844	-2.180	-3.303	-7.705	-9.278	-6.245	-9.278
M1004	T0006	S358	0.776	0.203	1.070	-2.748	-9.565	-5.170	-9.565
M1186	T0005	S366	-3.032	-0.469	-2.281	-7.315	-2.827	-0.415	-7.315
M1321	T0003	S364	-5.558	-6.835	-5.788	-1.420	-3.832	-4.796	-6.835
M1358	T0018	S366	-6.680	-3.295	-1.642	-7.066	-1.332	-4.106	-7.066
M1771	T0016	S1239	-2.051	-1.539	-4.028	-3.528	-4.116	-7.227	-7.227
M1828	T0010	S1243	-1.027	-0.091	-0.825	-5.399	-4.024	-4.734	-5.399
M1887	T0007	S992	-5.521	-1.815	-9.540	-0.042	-2.782	-3.545	-9.540
M1937	T0005	S366	-5.779	-4.856	-9.396	-0.905	0.856	-4.178	-9.396
M1959	T0010	S1058	-0.810	0.658	-6.629	1.629	2.222	-2.485	-6.629
M2232	T0016	S1441	-9.173	-5.026	-4.509	-3.084	-1.597	-2.157	-9.173
M2252	T0018	S1105	-2.262	-1.168	-5.365	-5.128	-6.091	-6.198	-6.198
M2417	T0010	S1247	0.093	-5.080	-0.445	-2.030	-1.039	0.310	-5.080
M2601	T0011	S366	-6.815	-7.794	-2.933	-3.174	-3.658	-4.952	-7.794
M2683	T0020	S1163	0.645	1.322	2.885	1.260	-6.256	2.123	-6.256
M2789	T0010	S1243	-0.916	-2.248	-6.452	0.591	2.005	2.170	-6.452
M3083	T0010	S364	0.896	-2.017	-6.495	-4.018	-0.639	0.598	-6.495
M3398	T0011	S1177	-6.647	-8.660	-4.918	-1.075	-1.906	-3.963	-8.660
M3407	T0009	S1243	-8.521	-7.523	-6.402	-2.225	-3.176	-3.607	-8.521
M3411	T0015	S1291	-9.534	-3.900	-7.526	-7.832	-1.237	-8.893	-9.534

As seen from the Table 37, the threshold level changes between -5 and -9.6. Figure 20 illustrate the changes in the threshold level.

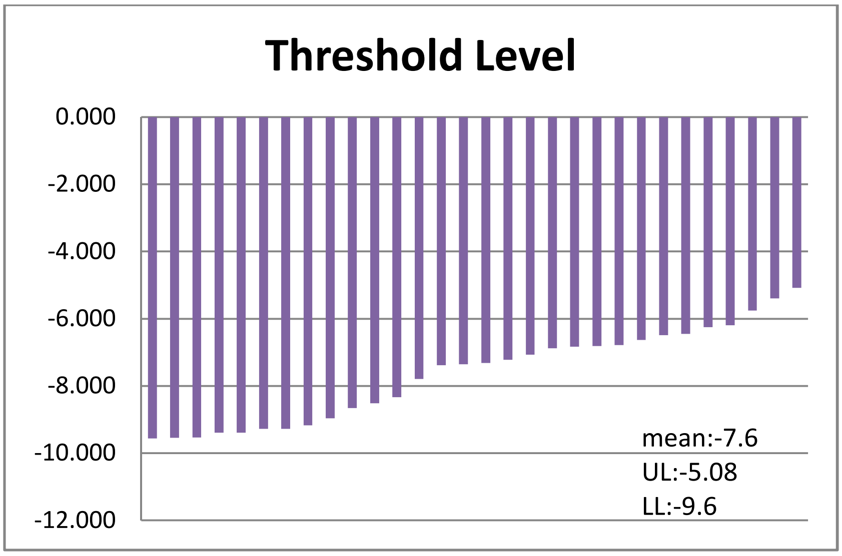
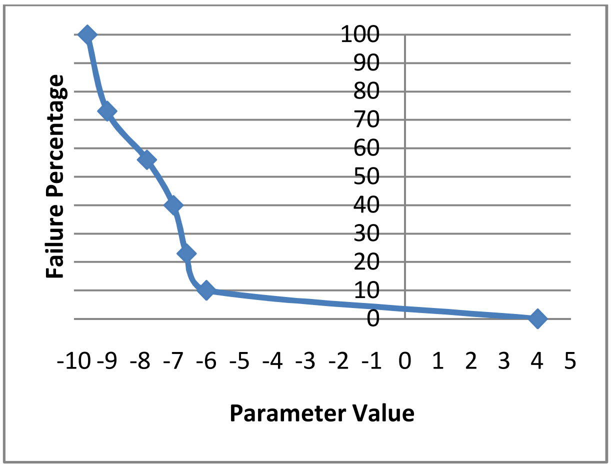


Figure 20: Variability of Threshol Level for 30 failure Cases

Upper Limit	Lower Limit	Observed Instances	Prob.	Cum. Prob.
-9	-9.6	8	0.267	1
-7.8	-9	5	0.167	0.73
-7	-7.8	5	0.167	0.56
-6.6	-7	5	0.167	0.4
-6	-6.6	4	0.133	0.23
-5	-6	3	0.1	0.1
4	-5	0	0	0



Failure percentage refers to the cumulative probability which shows the percentage of failure cases occurs before the lower threshold limit. To explain the relation between the failure percentage and the threshold level, piecewise linear regression, 2nd order polynomial regression and logarithmic regression methods were used. Adjusted R square was used to evaluate the methods.

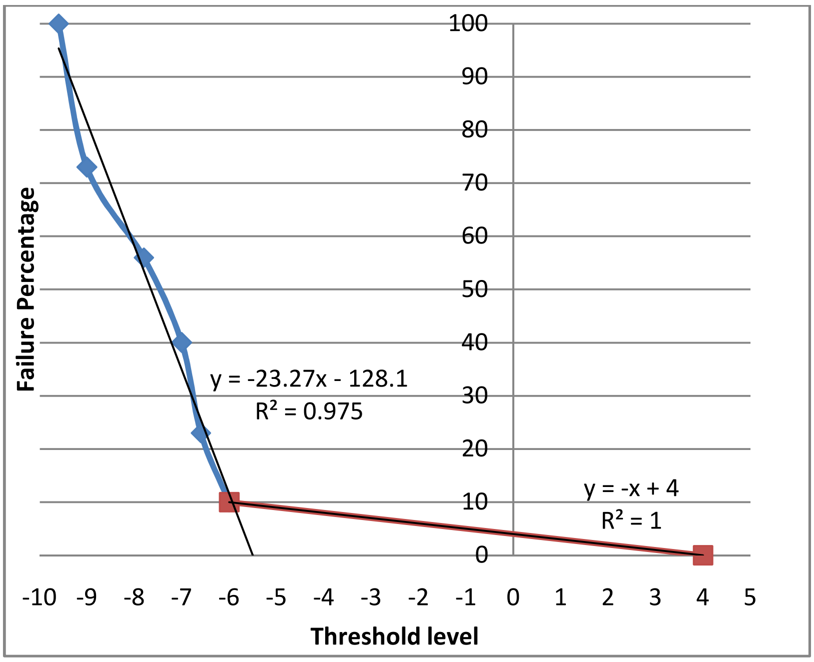
Modelling to Explain changes in Failure Percentage with the Threshold Level

Failure percentage (dependent variable) is defined as a function of threshold level (independent variable).

Threshold Level	Failure Percentage
-9.6	1
-9	0.73
-7.8	0.56
-7	0.4
-6.6	0.23
-6	0.1
4	0

1. <u>Piecewise Linear Regression</u>

Piecewise linear regression is used to explain abrupt changes of the response function. Independent variable is partitioned into intervals in which it exhibits different relations between the dependent variable.



"-6" is the breakpoint for the threshold level. Two linear regression models have been developed to explain the relation between failure percentage and threshold level. First one is valid when the threshold value is less than -6 and the other is valid when threshold value is equal or greater than -6.

If TL<-6

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.988[a]	.976	.970	5.78559	2.661

a. Predictors: (Constant), TL

b. Dependent Variable: percent

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-128.108	14.299		-8.960	.001	-167.807	-88.409
	TL	-23.275	1.839	-.988	-12.654	.000	-28.382	-18.168

a. Dependent Variable: percent

If TL>=-6

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	1.000[a]	1.000	.	.	.[b]

a. Predictors: (Constant), TL

b. Not computed because there is no residual variance.

c. Dependent Variable: Percent2

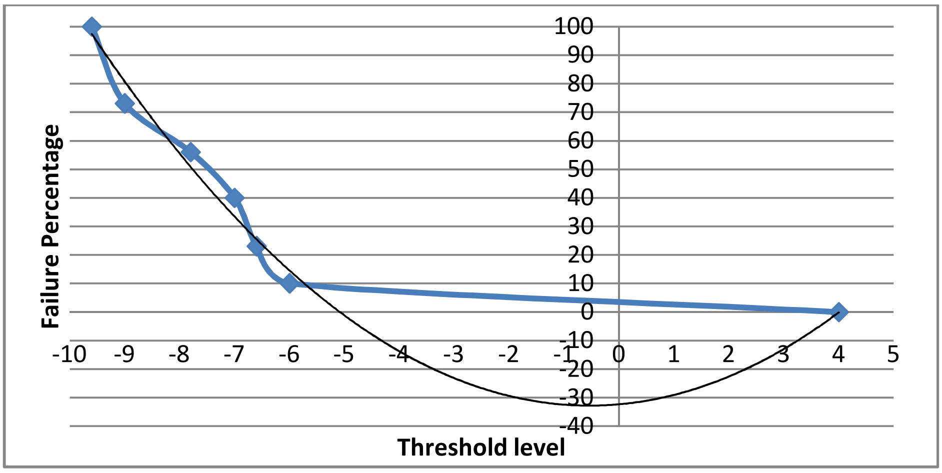
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	4.000	.000		.	.	4.000	4.000
	TL2	-1.000	.000	-1.000	.	.	-1.000	-1.000

a. Dependent Variable: Percent2

If TL<-6	*Percentage=-128.1-23.2 TL*	$R^2_a=97\%$
If TL>=-6	*Percentage=4-TL*	$R^2_a=100\%$

2. 2^{nd} order Polynomial Regression

TL	TL^2	Cum. Prob.
-9.6	92.16	100
-9	81.00	73
-7.8	60.84	56
-7	49.00	40
-6.6	43.56	23
-6	36.00	10
4	16.00	0



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.989[a]	.979	.968	6.38793	2.345

a. Predictors: (Constant), TL^2, TL

b. Dependent Variable: percent

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-32.288	6.180		-5.225	.006	-49.445	-15.130
	TL	1.706	1.015	.219	1.682	.168	-1.111	4.523
	TL^2	1.584	.178	1.163	8.924	.001	1.091	2.077

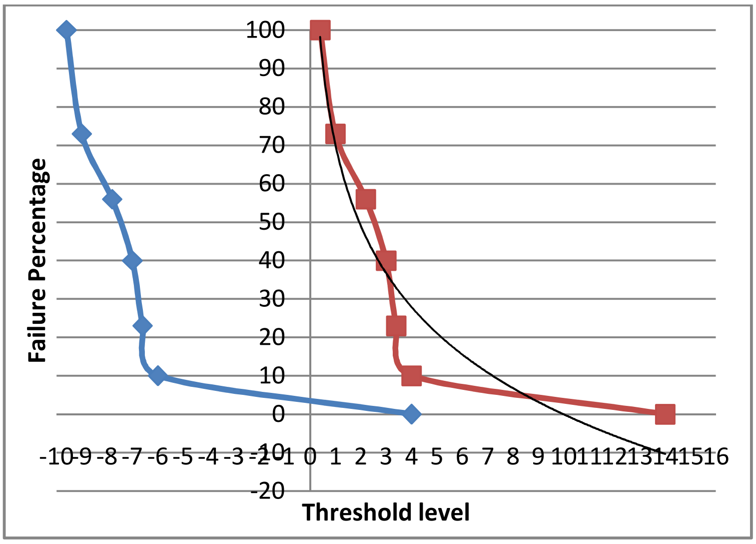
a. Dependent Variable: percent

$$Percentage=-32.28+1.706\ TL+1.584\ TL^2 \qquad R^2_a=96.8\%$$

3. Logarithmic regression

In order to apply logarithmic regression, it is required to transform negative values to positive values. Therefore TL+10 is used instead of TL

TL	TL+10	ln(TL+10)	Cum. Prob.
-9.6	0.4	-0.92	100
-9	1.00	0	73
-7.8	2.20	0.79	56
-7	3.00	1.09	40
-6.6	3.40	1.22	23
-6	4.00	1.38	10
4	14.00	2.64	0



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.957[a]	.916	.899	11.34589	1.757

a. Predictors: (Constant), ln(TL+10)

b. Dependent Variable: percent

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	70.284	5.648		12.445	.000	55.766	84.801
	ln(TL+10)	-30.545	4.136	-.957	-7.385	.001	-41.177	-19.913

a. Dependent Variable: percent

*Percentage=70.28-30.5*ln(TL+10)*	R^2_a =89.9%

Appendix III. Effect of the Environmental Factors

Sample Size for Each Group:

Machine Type	Machine Number
T0007	5
T0010	5
T0009	3
T0011	3
T0005	2
T0006	2
T0008	2
T0016	2
T0018	2
T0003	1
T0014	1
T0015	1
T0020	1

Site ID	Machine Number
S1243	5
S366	4
S1058	3
S1177	2
S1178	2
S364	2
S992	2
S1105	1
S1163	1
S1199	1
S1247	1
S1239	1
S1289	1
S1291	1
S1441	1
S159	1
S358	1

Customer ID	Machine Number
C1	7
C1665	5
C1146	3
C188	3
C6	2
C3231	2
C386	2
C13	1
C1416	1
C169	1
C218	1
C3841	1
C4208	1

Appendix IV. Neural Network Model-MLP

Neural network was built by using multilayer perceptron function. The model parameters were tuned to obtain the better model.

Inputs: P1...P6

Network Parameters: LR=0.7, M=0.2, NHL=a

Output: A/B/C

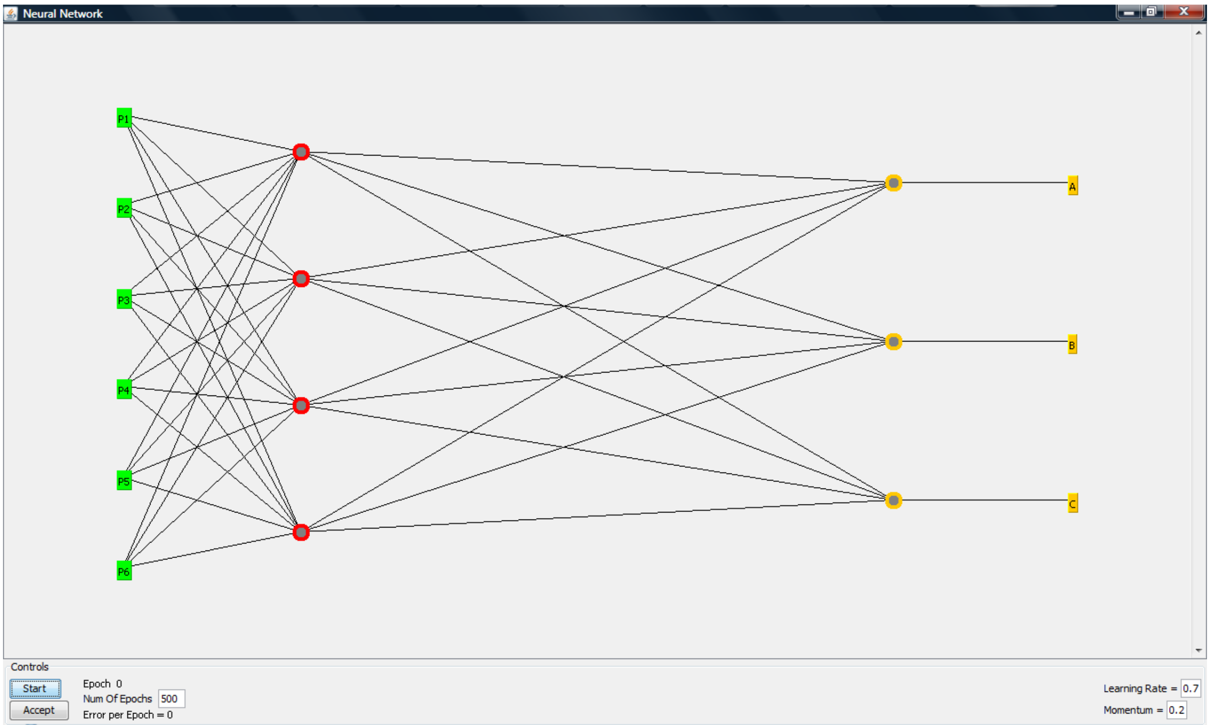


Figure 21: Screenshot WEKA NN-MLP model

=== Classifier model (full training set) ===

Sigmoid Node 0

Inputs Weights

Threshold 0.04368478839871569

Node 3 -5.241075085065454

Node 4 -9.626191787137266

Node 5 4.4206423031749305

Node 6 2.278513263179502

Sigmoid Node 1

Inputs Weights

Threshold -12.09741926200117

Node 3 3.5755479916929227

Node 4 6.108590741274521

Node 5 -4.3281772778173595

Node 6 10.195975922401853

Sigmoid Node 2

Inputs Weights

Threshold -0.8124151450850194

Node 3 14.193319884462513

Node 4 17.524941264738253

Node 5 -0.268217928006194

Node 6 -15.69786641196598

Sigmoid Node 3

Inputs Weights

Threshold 7.72978523021964

Attrib P1 -11.698292773666923

Attrib P2 -4.796585835949385

Attrib P3 22.29758616110523

Attrib P4 -7.965837437162937

Attrib P5 21.239280023619646

Attrib P6 24.623343898716197

Sigmoid Node 4

Inputs Weights

Threshold -26.51472811845105

Attrib P1 0.49914609884105976

Attrib P2 1.9354300306975871

Attrib P3 -7.633429587144298

Attrib P4 13.29814564317826

Attrib P5 -19.570666393350827

Attrib P6 -48.006511218252754

Sigmoid Node 5

Inputs Weights

Threshold 13.431278695996594

Attrib P1 6.0693263945342375

Attrib P2 2.10646609764955

Attrib P3 53.542584995024036

Attrib P4 -23.32121054466619

Attrib P5 12.33589227891096

Attrib P6 -24.519208368505797

Sigmoid Node 6

Inputs Weights

Threshold 26.00831697869943

Attrib P1 -8.376349735430326

Attrib P2 9.371008413228056

Attrib P3 49.05966593354248

Attrib P4 -15.474705726268073

Attrib P5 29.175124929933087

Attrib P6 -23.954276761667806

Class A

Input

Node 0

Class B

Input

Node 1

Class C

Input

Node 2

=== Evaluation on test set ===

=== Summary ===

Correctly Classified Instances	157	51.4754 %
Incorrectly Classified Instances	148	48.5246 %
Kappa statistic	0.1789	
Mean absolute error	0.3619	
Root mean squared error	0.5244	
Relative absolute error	91.2221 %	
Root relative squared error	102.8869 %	
Total Number of Instances	305	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.876	0.679	0.513	0.876	0.647	0.632	A
	0.036	0.081	0.143	0.036	0.058	0.473	B
	0.4	0.073	0.68	0.4	0.504	0.643	C
Weighted Average	0.515	0.347	0.459	0.515	0.447	0.592	

=== Confusion Matrix ===

a	b	c	<-- classified as
120	17	0 |	a = A
64	3	16 |	b = B
50	1	34 |	c = C

Appendix V. Neural Network Model-RBF

Inputs: P1...P6

Outputs: A/B/C

Radial basis function network

(Logistic regression applied to K-means clusters as basis functions):

Logistic Regression with ridge parameter of 1.0E-8

Coefficients...

	Class	
Variable	A	B
pCluster_0_0	0.2036	0.4052
pCluster_0_1	0.0766	-0.0711
pCluster_0_2	-0.3067	-0.3568
Intercept	1.889	0.205

Odds Ratios...

	Class	
Variable	A	B
pCluster_0_0	1.2258	1.4997
pCluster_0_1	1.0796	0.9314
pCluster_0_2	0.7359	0.6999

=== Evaluation on test set ===

=== Summary ===

Correctly Classified Instances	137	44.918 %
Incorrectly Classified Instances	168	55.082 %
Kappa statistic	0	
Mean absolute error	0.3971	
Root mean squared error	0.5071	
Relative absolute error	100.108 %	
Root relative squared error	99.5054 %	
Total Number of Instances	305	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	1	0.449	1	0.62	0.711	A
	0	0	0	0	0	0.407	B
	0	0	0	0	0	0.662	C
Weighted Average	0.449	0.449	0.202	0.449	0.278	0.614	

=== Confusion Matrix ===

a	b	c	<-- classified as
137	0	0 |	a = A
83	0	0 |	b = B
85	0	0 |	c = C

Appendix VI. Combined Decision Tree Model

Inputs: P1...P6

Minimum Number of Records: 20

Outputs: A/B/C

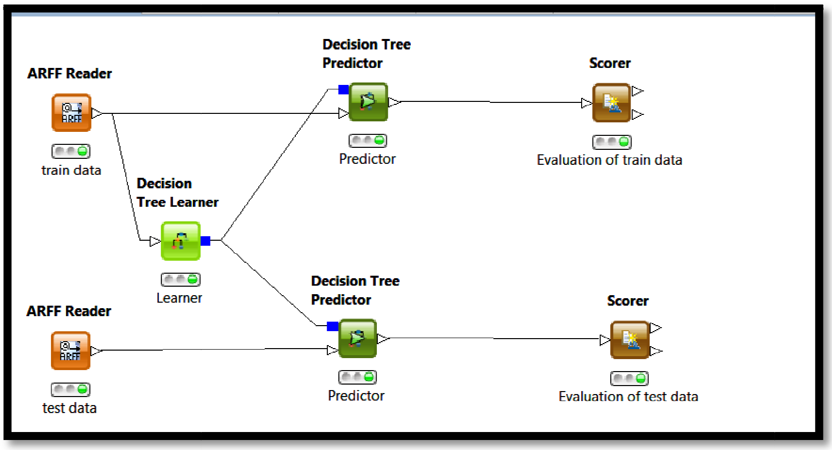
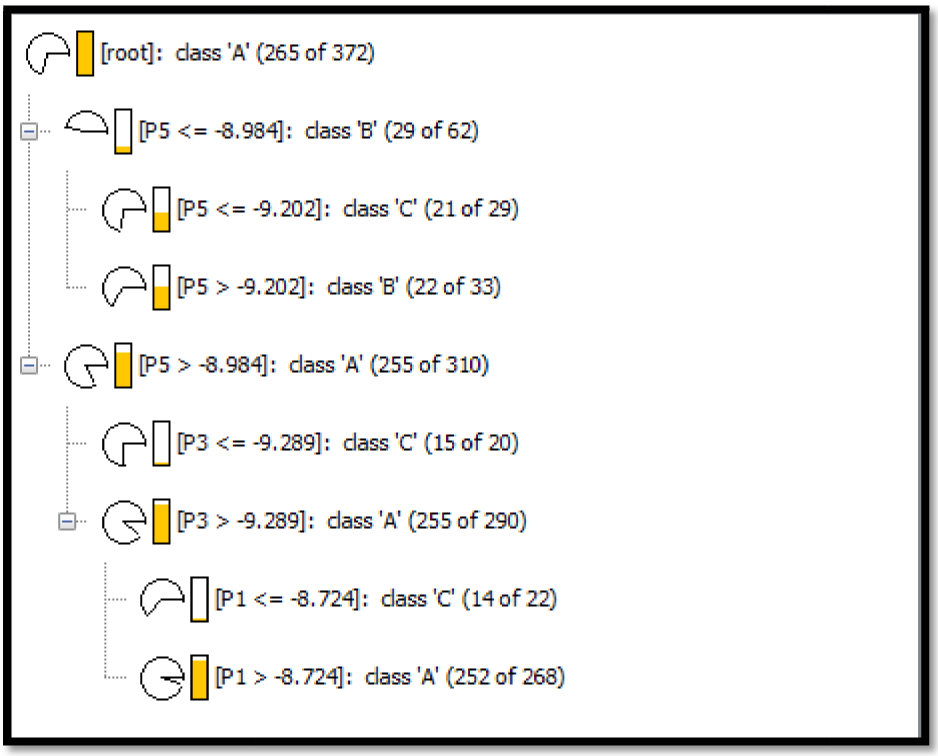


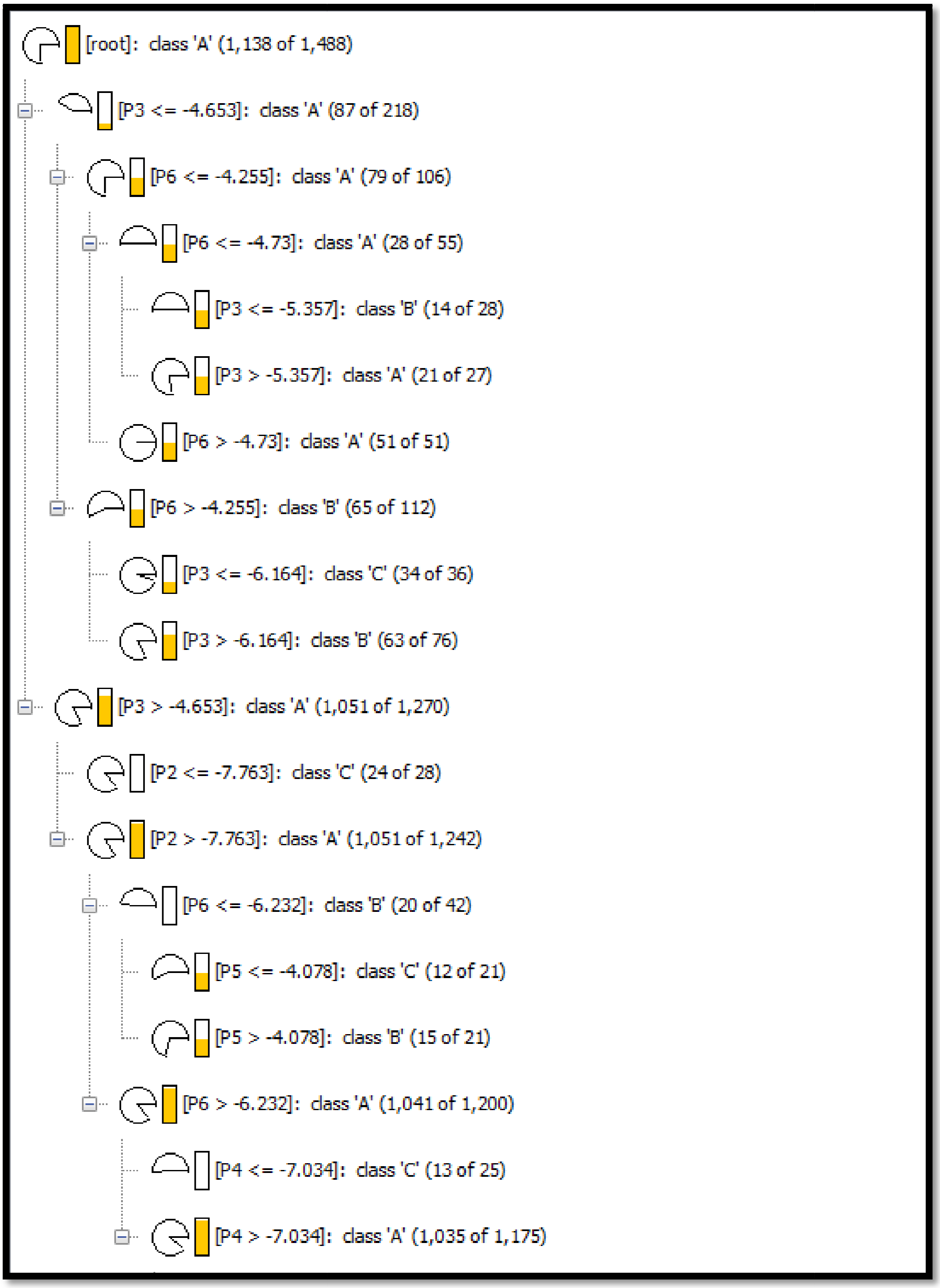
Figure 22: Screenshot of KNIME Decision Tree Model

Data sets were divided into two groups: The first group includes the hard failures of which threshold level is smaller than -9 and the second group includes the soft and medium failures. Separate decision trees were developed for each group.

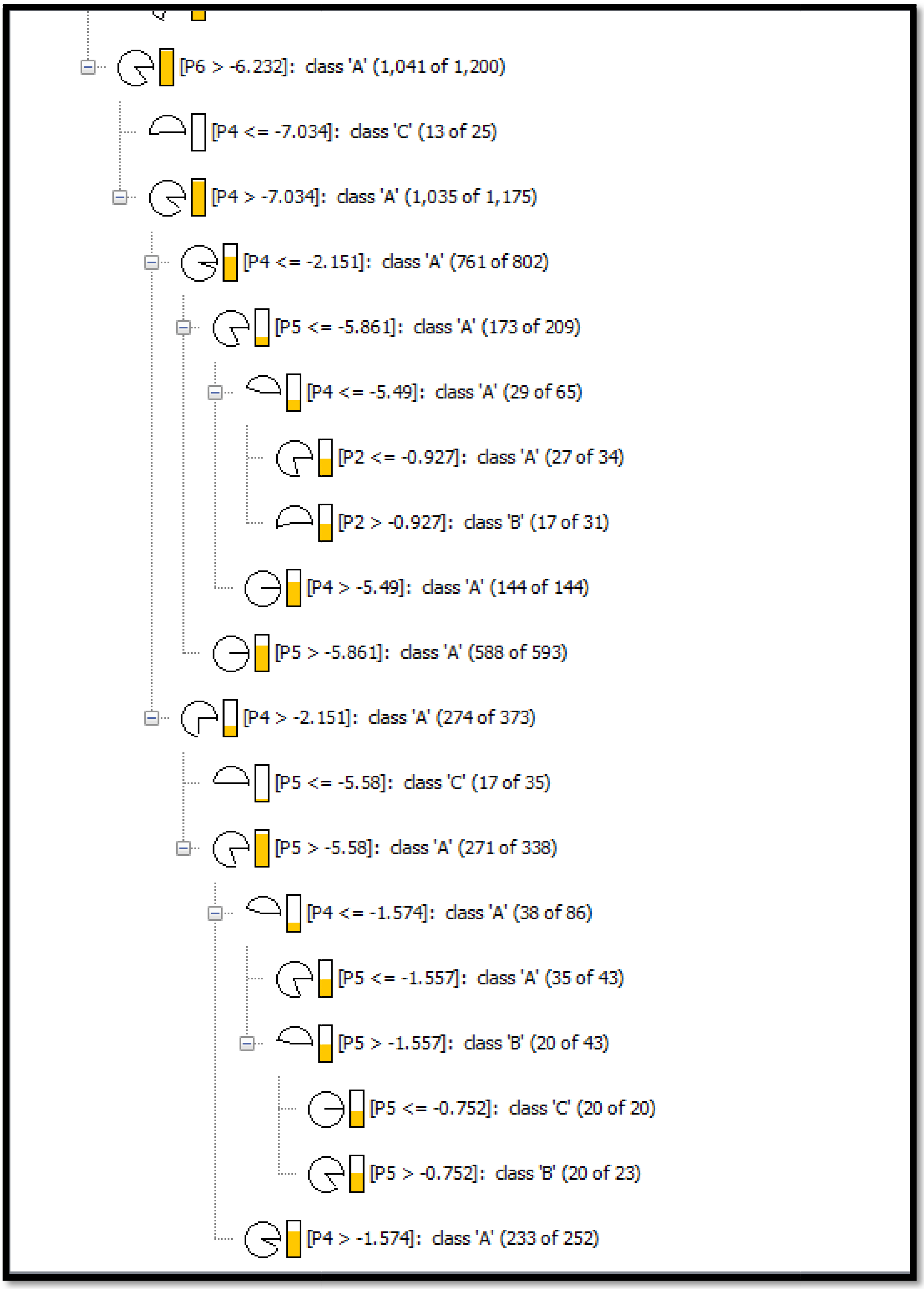
1) Decision Tree for Hard Failures



2) Decision Tree for Medium and Soft Failures



Decision Tree for Medium and Soft Failures (continued)



Appendix VII. Simple CART Model

Inputs: P1...P6

Minimum Number of Objects: 10

Outputs: A/B/C

```
P3 < -5.276999999999999
| P1 < -1.6444999999999999
| | P1 < -5.2735
| | | P3 < -9.362: C(14.0/1.0)
| | | P3 >= -9.362
| | | | P5 < -9.216000000000001: C(11.0/0.0)
| | | | P5 >= -9.216000000000001
| | | | | P5 < -8.8215: B(22.0/2.0)
| | | | | P5 >= -8.8215
| | | | | | P2 < -3.4185: B(21.0/22.0)
| | | | | | P2 >= -3.4185: A(28.0/6.0)
| | P1 >= -5.2735: A(60.0/4.0)
| P1 >= -1.6444999999999999
| | P3 < -6.0905000000000005: C(31.0/2.0)
| | P3 >= -6.0905000000000005: B(35.0/4.0)
P3 >= -5.276999999999999
| P2 < -7.763
| | P5 < -2.8935: C(24.0/0.0)
| | P5 >= -2.8935: B(13.0/1.0)
| P2 >= -7.763
| | P6 < -6.058999999999999
| | | P5 < -7.263: C(13.0/3.0)
| | | P5 >= -7.263
| | | | P6 < -6.5280000000000005: B(16.0/4.0)
| | | | P6 >= -6.5280000000000005: A(18.0/7.0)
```

```
| | P6 >= -6.0589999999999999
| | | P1 < -8.582: C(10.0/4.0)
| | | P1 >= -8.582
| | | | P5 < -9.175: B(7.0/8.0)
| | | | P5 >= -9.175
| | | | | P4 < -2.0945: A(960.0/90.0)
| | | | | P4 >= -2.0945
| | | | | | P3 < -2.3979999999999997: B(32.0/15.0)
| | | | | | P3 >= -2.3979999999999997
| | | | | | | P5 < -0.3635
| | | | | | | | P5 < -1.082
| | | | | | | | | P2 < 0.6719999999999999: A(117.0/10.0)
| | | | | | | | | P2 >= 0.6719999999999999
| | | | | | | | | | P1 < 1.102
| | | | | | | | | | | P1 < 0.806: C(10.0/0.0)
| | | | | | | | | | | P1 >= 0.806: B(9.0/1.0)
| | | | | | | | | | P1 >= 1.102: A(31.0/2.0)
| | | | | | | | P5 >= -1.082
| | | | | | | | | P2 < -2.709: C(19.0/0.0)
| | | | | | | | | P2 >= -2.709
| | | | | | | | | | P1 < 0.799: B(20.0/0.0)
| | | | | | | | | | P1 >= 0.799: A(11.0/0.0)
| | | | | | | P5 >= -0.3635: A(141.0/1.0)
```

Number of Leaf Nodes: 25

Size of the Tree: 49

=== Evaluation on test set ===

=== Summary ===

Correctly Classified Instances	152	49.8361 %
Incorrectly Classified Instances	153	50.1639 %
Kappa statistic	0.2178	
Mean absolute error	0.3369	
Root mean squared error	0.4965	
Relative absolute error	84.9225 %	
Root relative squared error	97.4202 %	
Total Number of Instances	305	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.737	0.286	0.678	0.737	0.706	0.76	A
	0.602	0.468	0.325	0.602	0.422	0.583	B
	0.012	0.005	0.5	0.012	0.023	0.722	C
Weighted Average	0.498	0.257	0.532	0.498	0.438	0.701	

=== Confusion Matrix ===

```
  a    b    c   <-- classified as
101   35    1 |  a = A
 33   50    0 |  b = B
 15   69    1 |  c = C
```

Appendix VIII. KNN Model

Inputs: P1...P6

K=4

Outputs: A/B/C

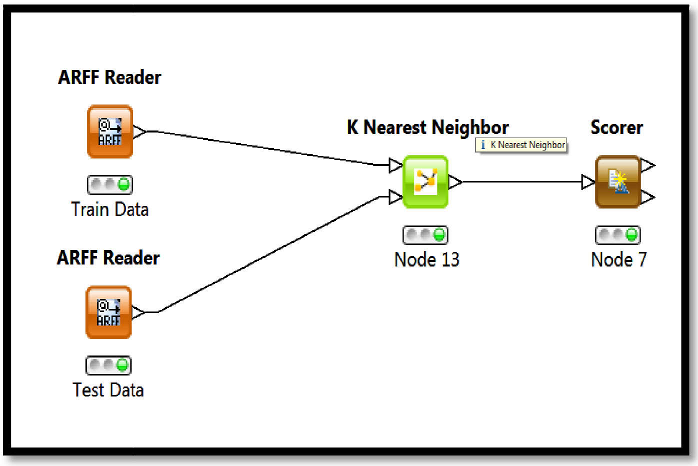


Figure 23: Screenshot of KNIME-KNN model

Appendix IX. Neural Network Model-MLP

Inputs: P1...P6

LR:0.4, M:0.1, NHL:a

Output: Remaining time in days

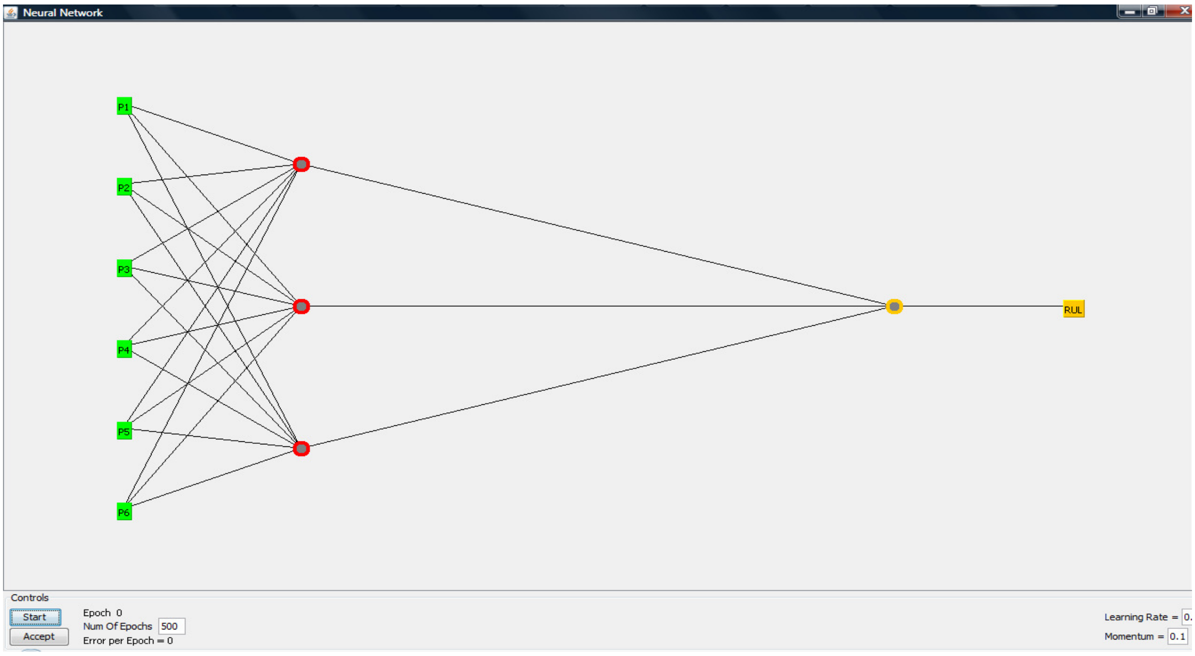


Figure 24: Screenshot WEKA NN-MLP model

Linear Node 0

Inputs Weights

Threshold 0.006480936361519446

Node 1 -1.107611043391414

Node 2 -0.7771267932603714

Node 3 -1.5184527261069822

Sigmoid Node 1

Inputs Weights

Threshold -27.56772289921536

Attrib P1 -20.299756378388704

Attrib P2 11.757303415103786

Attrib P3 14.15104798143839

Attrib P4 -1.8769185504764276

Attrib P5 -31.48160008262188

Attrib P6 23.132209678316983

Sigmoid Node 2

Inputs Weights

Threshold -13.626780162857127

Attrib P1 -3.8258135394781094

Attrib P2 5.4235318762705536

Attrib P3 4.185245202412306

Attrib P4 -6.684922811212388

Attrib P5 1.618173646549513

Attrib P6 -26.34678352783777

Sigmoid Node 3

Inputs Weights

Threshold -2.6938131303318196

Attrib P1 -0.5182833586246139

Attrib P2 -1.7289733470104356

Attrib P3 -3.012523523759215

Attrib P4 0.19482925777786173

Attrib P5 1.9575018737115757

Attrib P6 2.6106362640698935

Class

Input

Node 0

=== Evaluation on test set ===

=== Summary ===

Correlation coefficient	0.357
Mean absolute error	**68.3287**
Root mean squared error	80.3588
Relative absolute error	94.1922 %
Root relative squared error	95.7458 %
Total Number of Instances	305

Appendix X. Neural Network Model-MLP (given threshold level)

This model has been developed to analyze the effect of threshold level as an input on the remaining useful life prediction.

Inputs: P1...P6, Threshold Level

LR:0.3, M:0.2, NHL:a

Output: Remaining time in days

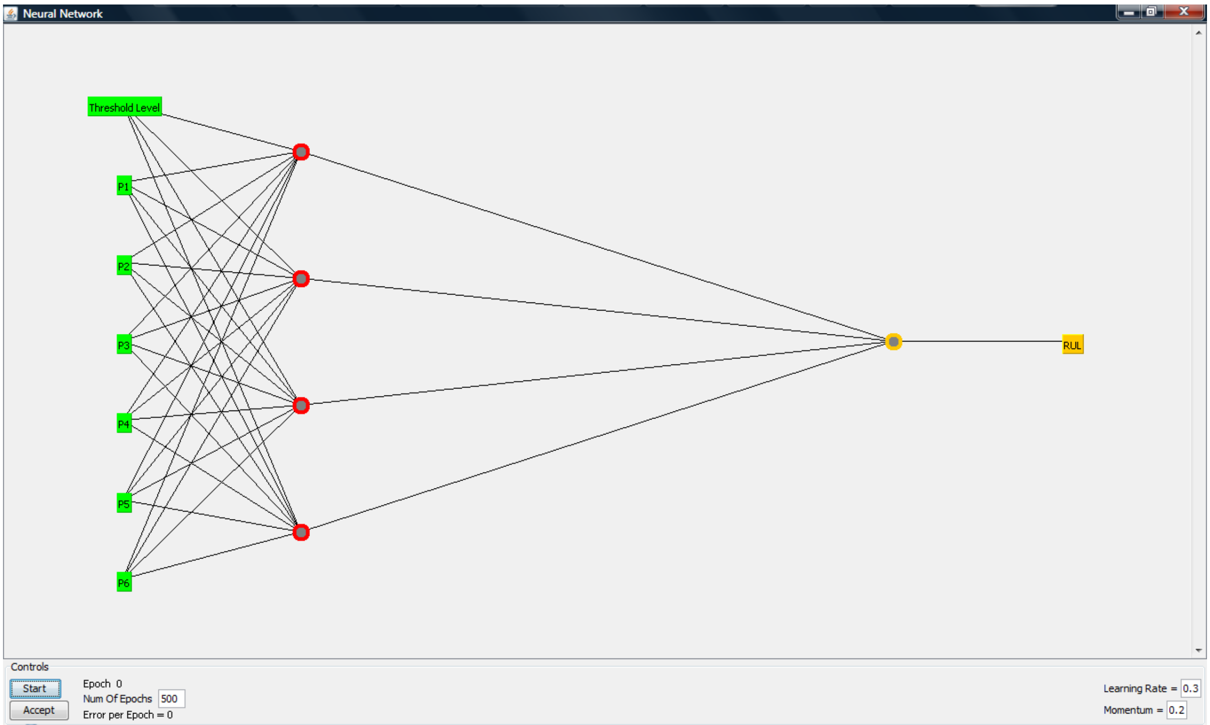


Figure 25: Screenshot WEKA NN-MLP model

=== Classifier model (full training set) ===

Linear Node 0

Inputs Weights

Threshold 0.6217563641164013

Node 1 -1.900795736043391

Node 2 -1.2149732515097276

Node 3 -1.5839876684844776

Node 4 -1.3164716815883042

Sigmoid Node 1

Inputs Weights

Threshold -2.297902554953836

Attrib Threshold Level 3.6258768839385147

Attrib P1 -0.06761636079786001

Attrib P2 -1.1378440995381143

Attrib P3 -2.819718163736291

Attrib P4 0.26856504230787626

Attrib P5 2.8246296610789634

Attrib P6 -0.13034164441028415

Sigmoid Node 2

Inputs Weights

Threshold -6.325628001580753

Attrib Threshold Level 1.9475374761518525

Attrib P1 -1.2032274924139537

Attrib P2 -0.32954057852118657

Attrib P3 4.2387383937142555

Attrib P4 2.907984572078976

Attrib P5 -5.254859808293894

Attrib P6 -13.890320584461264

Sigmoid Node 3

Inputs Weights

Threshold -10.69038541206472

Attrib Threshold Level 7.6540208490448265

Attrib P1 2.3402318693174458

Attrib P2 -0.7963118974934503

Attrib P3 7.37721168430417

Attrib P4 2.145597838689219

Attrib P5 -11.56354043575375

Attrib P6 2.886801807574954

Sigmoid Node 4

Inputs Weights

Threshold -7.86553023228846

Attrib Threshold Level -5.2563511593691405

Attrib P1 -11.34746626444243

Attrib P2 -6.184661814615738

Attrib P3 -8.032996082044976

Attrib P4 -9.663792372046991

Attrib P5 -1.5702259028481504

Attrib P6 14.427915918690703

Class

Input

Node 0

=== Evaluation on test set ===

=== Summary ===

Correlation coefficient	0.1925
Mean absolute error	**74.6245**
Mean absolute error in last 30 days before failure	**74.47**
Root mean squared error	105.487
Relative absolute error	102.8711 %
Root relative squared error	125.6855 %
Total Number of Instances	305

Threshold level does not improve RUL prediction in MLP model.

Appendix XI. Project Timeline

This appendix has been prepared to give an overview about project timing. Figure 26 shows the time spent on each project steps.

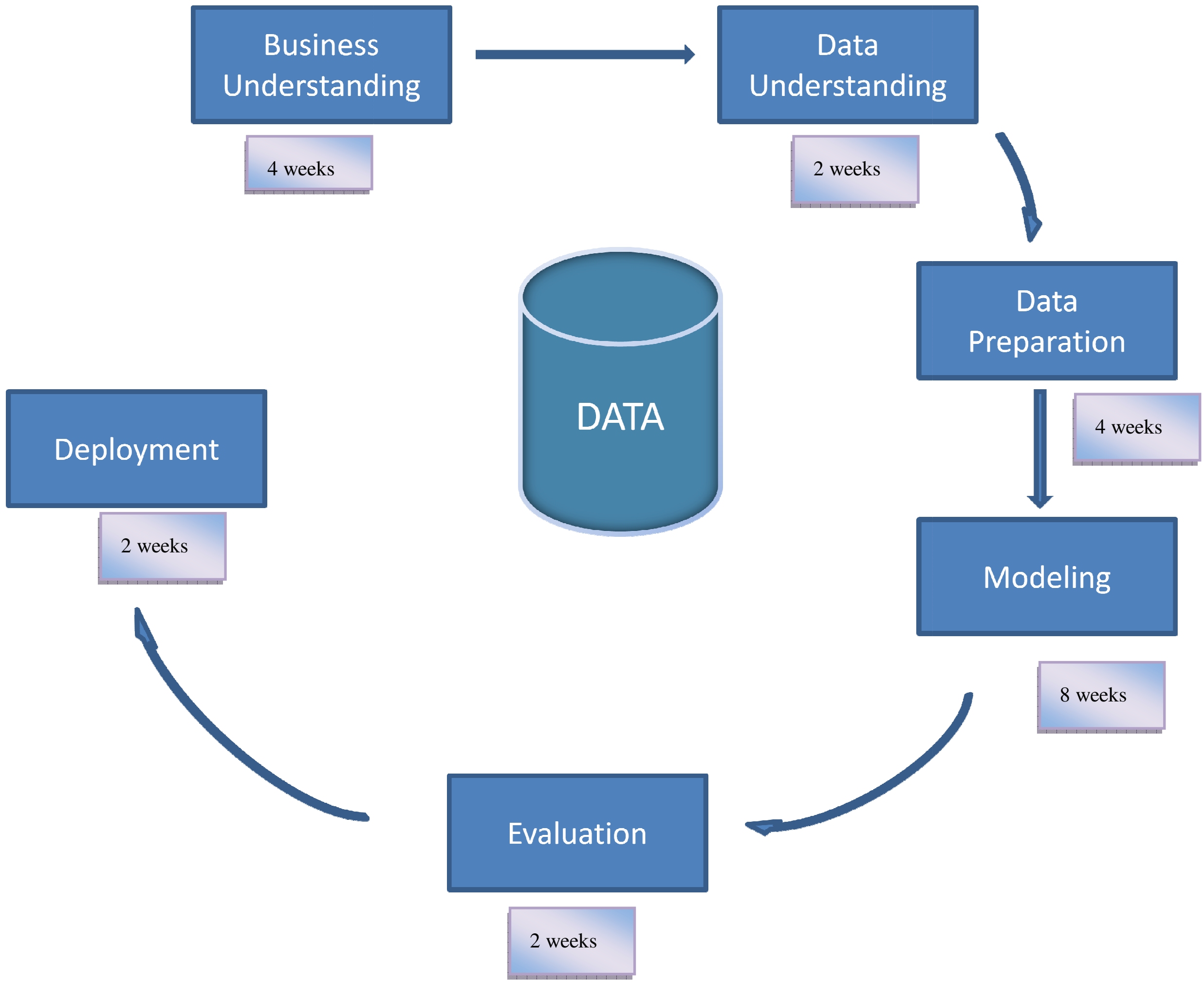


Figure 26: Project Timeline