# Analyzing the return of sound products with the use of data mining techniques

Breukink, T.G.

*Award date:*
2009

Eindhoven, September 2008

# Analyzing the return of sound products with the use of data mining techniques

By:

Ing. Thomas G. Breukink

0617081

Student identity number: 0617081

in partial fulfilment of the requirements for the degree of

**Master of Science**

**in Operations Management and Logistics**

Supervisors:

Dr. A.J.M.M Weijters, TU/e, IS

Dr. Y.Lu, TU/e, ID

TUE. Department Technology Management.
Series Master Theses Operations Management and Logistics

Subject headings: Data mining, Quality and Reliability, No Failure Found, Soft Reliability

# Acknowledgement

Writing a master thesis requires a lot of endeavour from the student, consecutively his name is written on the cover of the thesis. It is an illusion that the student alone is responsible for the work present in the paper. Without the help of the people around me, my master thesis would have never been completed. I would like to use this opportunity to thank those who have, in one way or another, contributed to this work.

In particular, I would like to express my gratitude to my mentors and correctors at the University. Ton Weijters who was able to get me back on track using such little pointers that left me to think that I invited the road by myself. Lu Yuan and Renate de Bruin, who kept surprising me with interesting insights from the diverse world of academic literature.

I would also like to thank to company for letting me handle the confidential data and the aid from within the company that was available to me. When asking a simple question to an employee the replies were normally long and all-inclusive. There were times I felt people were leaving their own work to help me out with the smallest details. Interesting is that everyone in the company was this helpful.

I have experienced aid from outside the education and business community. Discussions regarding the main topics and the world surrounding it, normally were of a more customer view on the topic. My neighbour, girlfriend and friends most likely remember these debates.

Thomas Breukink

September 2009

# Summary

*Many scientific summaries are quite lengthy. Because summaries are intended to be a short simplification of the project, the researcher decided to limit to the core and fit it on one page.*

The increase in returned products with no failure found (NFF) has been noticed by the business and academic world. This increase has been examined by many researchers and a multiple trends have been identified. Apparently, the increase in product complexity, increase in consumer expectation, consumer power and the decrease in time to market play a role in this process.

This paper examines one particular product in one particular company with the use of data mining tools. Data mining was chosen as a path because of the abundant availability of data and the limited possibilities with existing quality and reliability tools. The company under consideration experiences a high amount of NFF returns at a particular part. The situation is a bit different from the ones described in the literature. First it is a part of a machine, not an entire product and secondly the company operates in the business market, not the consumer market like is general in the literature.

In order to get a rich dataset, data from three different sources was merged. These three sources are production, helpdesk and service engineers and the collection is called the fused database. The merging itself is done based on machine serial numbers and part serial numbers. All products sold are supposed to return to the company, the same holds for parts. From this stream of return products, 50 parts are sampled and analysed on a monthly basis. These analyses form a centrepiece for the research. From the combination of these 750 sampled parts, and the fused database the causes for product return are distilled.

The fused database contained 18 features that could possibly explain the return of a product. The ‘decision tree miner’ produced a fault tree that clearly displayed three main causes of product return. These three causes explain roughly 80% of the NFF returns.

The first cause is the unnecessary replacement of a module. In the normal situation, the machine will notify the user when the module almost reaches end of its life. The consumer orders this module and replenishes it before the machine stops functioning. The problem occurs when there is an error in the machine and it does not produce high quality prints. The consumer contacts the welcome centre who advises him to replace a certain module; this could well be the wrong module. The data pattern that triggers this cause is two replenishments close to each other with the same cause and of course, the module needs to be replaced in this time. This cause explains about 1/3 of the NFF returned products.

The second and smallest cause is the difference in contracts. Roughly, two kinds of contracts exist, one where the customer pays for the consumables, and one where the consumables are included in the contract. Research revealed that the customers who pay for their consumables have a 16% lower NFF return rate. Because these customers are rare, only 4% of the NFF returns are explained by this.

The third is the most promising for future improvement. The machine returns to the company at the end of the contract. The machine is next broken up and the parts are disposed trough the same return stream. This is the same product stream where the samples are drawn. This research reveals that 48% of the NFF returns are caused by this reason. By filtering the returns, the company is able to set these little used modules back on the market and make a considerable profit.

Data mining techniques were able to reveal these causes of NFF quite easily. In this case, data mining proved a strong tool in revealing underlying causes of NFF product returns. There are a few limitations to the data mining techniques currently available. First, it is not suited for free text second the process to setup the dataset and the mining tools it is quite labour intensive. The researcher found that in this case generically collected data proved more useful than data collected for a specific goal.

This document does not include any numerical details or information regarding the company. Because the company operates in a highly competitive environment and this information could damage their image. This deteriorates readability of the thesis; the author is of course aware of these details.

# Samenvatting

*Veel wetenschappelijke samenvattingen zijn in lengte een rapport op zich. De onderzoeker heeft er voor gekozen om zich in de samenvatting daadwerkelijk te beperken tot de kern van de zaak.*

De toename in geretourneerde producten waarin Geen Fout Gevonden kon worden (zogenaamde GFG producten) is niet onopgemerkt gebleven onder bedrijven en academici. Deze toename is door verschillende onderzoekers beschreven en ligt mogelijk in de behoefte van consumenten, de complexiteit van producten of de ontwikkelingssnelheid.

Dit onderzoek probeert de oorzaken van GFG retour producten te verklaren met behulp van data-analyse. Het onderzochte bedrijf heeft al sinds het op de markt brengen van onderdeel X een groot aantal GFG retours. Deze situatie onderscheidt zich van die in andere onderzoeken omdat het ten eerste een module betreft uit een machine en niet een geheel product en ten tweede in een zakelijke markt plaatsvindt, en niet in een consumentenmarkt.

Om een zo rijk mogelijke dataset te verkrijgen is data uit drie verschillende bronnen, productie, helpdesk en reparateurs, aan elkaar gekoppeld. Het koppelen van de data is gedaan aan de hand van het serienummer van de machine en het serienummer van het onderdeel. Het beleid van het bedrijf is om alle machines en onderdelen retour te laten komen. Van deze retourstroom worden maandelijks vijftig onderdelen willekeurig geanalyseerd. Die analyses zijn gebruikt in dit onderzoek. Door onderzoek van de ongeveer 750 geanalyseerde onderdelen is geprobeerd om, met behulp van de samengevoegde dataset, de redenen voor retour te destilleren.

De samengevoegde database bevatte achttien attributen die de retour zouden kunnen verklaren. Door middel van een ‘decision tree miner’ kon een foutboom gemaakt worden. In de data-analyse werden drie grote onderliggende oorzaken voor de retour van GFG producten gevonden. Deze drie oorzaken verklaren ruim 80% van de GFG producten.

De eerste oorzaak is het onnodig vervangen van de module. In de ideale situatie geeft de machine aan dat de module aan het einde van zijn leven zit, waarop de consument een nieuwe module bestelt en hem zelf vervangt. Het probleem ontstaat op het moment dat de machine niet correct werkt, en geen oorzaak aangeeft. De consument neemt dan contact op met de helpdesk, waarna de helpdesk de klant adviseert om iets te vervangen. Dit kan de verkeerde module zijn, waardoor de oude module GFG retour komt, dit verklaard ongeveer een derde van de GFG retours. Het datapatroon dat aangeeft dat er een dergelijke situatie plaatsvindt, laat meerdere oproepen zijn rond dezelfde datum en reden.

De tweede en tevens kleinste oorzaak is het verschil in contract tussen klanten die wel, en klanten die niet betalen voor hun onderdelen. Klanten die betalen voor hun onderdelen retourneren minder GFG producten, namelijk ongeveer 16% minder. Omdat er weinig klanten zijn die betalen voor hun onderdelen wordt louter 5% door deze factor verklaard.

De derde oorzaak, en meest kansrijke voor een toekomstige verbetering, zijn onderdelen die terugkomen met machines die het einde van hun leven hebben bereikt. De machines worden aan het einde van hun leven opgesplitst en afgebroken. Het onderzoek toont aan dat binnen de retourstroom 48% van de GFG retours afkomstig is van deze bron, recycling is een mogelijkheid.

De data-analyse heeft in dit bedrijf een goed beeld laten zien van de oorzaken van GFG retour producten. Dat toont aan dat er veel potentie zit in data-analyse voor kwaliteitsdoeleinden. Hierbij zijn wel een paar kanttekeningen. Het opzetten van de data-analyse is vrij tijdrovend en het is niet mogelijk om te werken met vrije tekst. Daarnaast zijn generieke gegevens waardevoller dan velden die voor een specifiek doel verzameld worden.

Zoals u zult zien zijn er in dit document geen numerieke details en geen gegevens van het bedrijf aanwezig. De reden hiervoor is dat het bedrijf in een zeer competitieve markt actief is, waar deze informatie zou kunnen leiden tot een slecht beeld. De gegevens zijn uiteraard wel bekend bij de schrijver.

# Table of contents

# Table of figures



# Table of tables



# Table of equations

# 1 Introduction and hypothesis

## 1.1 Problem description

Manufacturing industries experience a number of conflicting trends that result in an increasing return of products where no technical failure can be found. There is a large variety in terminology no problem found, no fault found, no trouble found or no defect found are regularly used. In the rest of the document, these products will be described as technically sound products or an abbreviation of No Failure Found (NFF). Paragraph 2.1.3 will dive into this terminology and explain the differences between the different terms. Brombacher et al. (2005) first described the increase of NFF product returns. The most referenced figure of his research is Figure 1-1 this clearly shows a rapid increase in the number of returned products where no fault could be found.

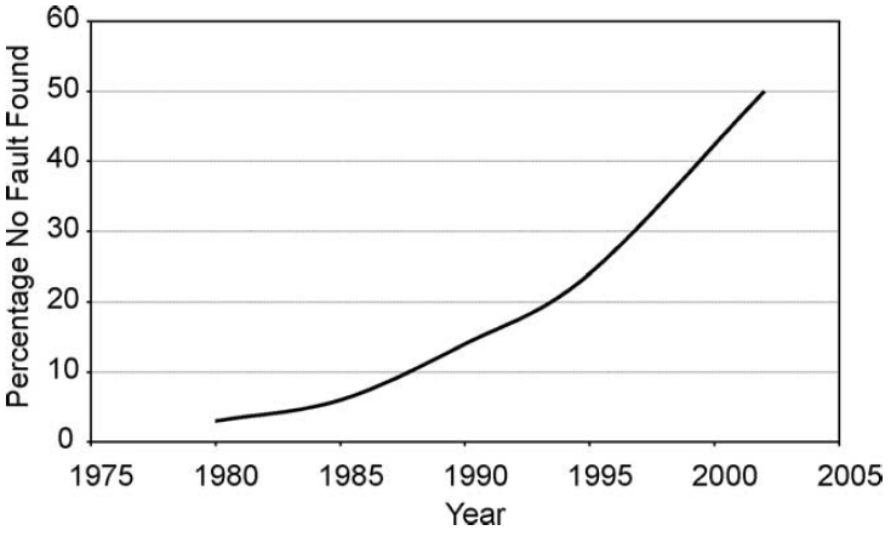


**Figure 1-1; Percentage NFF cases identified by Brombacher et al. (2005)**

The research of Brombacher et al (2005) was conducted on innovative technical products at a well-known manufacturer. In order to deal with these NFF returns two projects are initiated at the TU/e: soft reliability and data fusion. The soft reliability project is concerned with the classification of these failures and focuses on the gap between consumer requirements and technical specifications. The distinction is made in soft failures and hard failures *Hard failures* are failures where the product does not meet the technical specifications, while *soft failures* are products where the product confirms to specifications but is nevertheless returned to the manufacturer. The data fusion project is setup with the main goal to “support product development with prioritised information related to mismatches between product specifications and customer requirements” (Lu.Y, 2007). The next paragraph describes the project in more detail.

### 1.1.1 The data fusion project

The data fusion project is setup with the main goal to “support product development with prioritised information related to mismatches between product specifications and customer requirements” Lu.Y (2007). To generate this, different feedback databases need to be merged allowing extraction of more structural richer and prioritised field feedback information. This data fusion focuses on 6 different data sources; Service, Helpdesk / Call centre, Internet, Forum, Trade and Test this can be abbreviated in to SHI(F)TT fusion.

The data fusion approach is documented in the project proposal that describes four different phases shown Figure 1-2. The first phase discovery consist of the following parts; understanding the organisational structure and feedback mechanism, understanding the products, evaluating the

effectiveness of the current field feedback mechanisms, investigating the impact of data collection procedures regulations policies and cleaning/pre-processing. The second phase is concerned with the development of a generally applicable tool that is next evaluated and used. It is within the researcher's intention to answer part of the questions in the first phase of the project. How much of these questions are answered can be found after the formulation of the research questions in chapter 1.2 on Page 2.

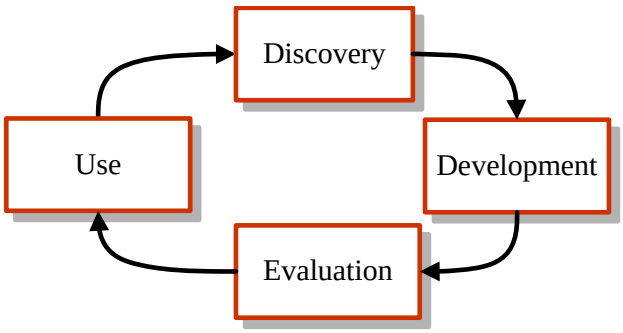


**Figure 1-2; Structure within the data fusion project**

It is clear that the data fusion project strongly relates to practice and therefore the project is connected to four manufacturing companies. These industrial partners are not new; there have been ongoing research projects with them for many years. At one particular company, a case study presented itself related to the NFF problem. Next to investigating the NFF problem at the company there is an added dimension to produce constructive results that assist the discovery phase of the project. It is now interesting to discuss the requirements from the company and to reveal the problem they face.

## *1.2 Research questions*

The company's requirements on this project ensured relevant research questions. This does not pose a problem with regard to the rigidness of the research because of the solid academic background provided by the 'IOP data fusion' and the 'Madame Curie' project. The research questions are setup in such a way that they prove useful for the exploratory part of the data fusion project.

The company captures large amounts of data regarding product return. Measures about the quality and reliability, return rates and usage are essential to control the quality and usage of the product. It is wise to use these datasets for three reasons. First, while it is a rich dataset the company is yet unable to use this extensive dataset to explain the NFF problem. Secondly, combining the dataset and the NFF issue could result in models generalisable over different modules and products, because logging happens in similar ways. Thirdly, the use of the large dataset results in a statistically stronger model than a model based on customer interviews and product analysis because of the large variety of customers.

Data mining techniques are especially strong in dealing with large amounts of data. Hidden patterns can be extracted form these datasets. Data mining is a explorative tool suited to give relations in databases. This exploratory nature of data mining and its ability to deal with large dataset are strong motivation for the use of data mining to reveal the causes off NNF product returns. This leads to the following problem statement:

> **How to apply mining techniques in field feedback data to improve understanding of the No Fault Found product returns.**

In order to answer the problem statement it is sliced in five more concrete research questions. The research questions are formulated and set up in such a way that it will prove useful for the data fusion project and are well related to the academic literature. The CRISP-DM methodology has proven useful to distinguish different phases in the project. Primarily knowledge is required on the business and feed feedback systems in place. This is all captured in the first research question:

1. **What field feedback systems are currently in place?**

The field feedback system uses several databases to store knowledge. From the SHI(F)TT databases the SHI(F)TT (Service, Helpdesk, Internet, (Forum), Trade and Test) databases are available. In depth knowledge needs to be gained about their content and information quality. Next to this, an open view is needed because other databases might contain valuable information as well. This information can next be used to generate different hypothesis and create a data mining problem definition. This leads to the second research question:

2. **What information regarding the No Failure Found returned products is available, and what features/attributes are possibly relevant.**

To analyse the dataset different mining techniques exist. These techniques need to be judged on their applicability within this project and a valid conclusion needs to be drawn on what techniques can best be applied. Therefore, the following research question is mainly literature oriented and must result in one or multiple techniques that can be applied to the data;

3. **Which data mining techniques are relevant to the case study?**

After knowing what data is available and what technique will prove the best results, the data from different data sources can be merged and analysed. The quality of the data largely determines the result of the following research question:

4. **What kind of relations can be found using these data mining tools?**

These data mining results are of no value if they are not validated with practice. This can be done by establishing a feedback of the results to the source and see if they are true and more importantly what improvements are possible leading to the decrease of the NFF returns. The next research question refers to validation and examination of the possible improvements:

5. **Validate these results with practice and find possible improvement for the NFF returns.**

One aim of researchers is to improve the current way of working not only in this situation but also for the entire world. In order to do so the final research questions is regarding the potential of the techniques used. Wheatear these techniques might prove useful in other situations.

6. **To what extend are these methods and techniques used in this case study applicable for other NFF problems.**

Now that the research questions are set and a clear direction is given, feedback is needed from the different parties involved.

## *1.3 'Hypothesis'*

The title of this chapter is largely misleading because, the research is not hypothesis driven but data driven. Hypothesis driven research requires hypothesis and the appropriate data. Here researchers look at relations between the data and see if the expected relation is present. With this technique, the hypothesis can be confirmed or rejected.

Another way to do research is data driven. This type of research does not require hypothesis setting because in the beginning no hypothesis is drawn. The computer makes the hypothesis with use of the data. For this reason you do not need hypothesis, these are included in the dataset. The results are strongly dependant on the dataset. The dataset is setup in such a way that it shows characteristics of the cases in multiple areas and one dependant variable. With the use of data mining techniques, these features are combined into a model that can predict the dependant variable (Reichardt, 2008). These features determine the direction of the research; the third chapter is dedicated to creating these features.

Outside the domain of IT the difference between data-driven and hypothesis driven research is often called exploratory research or confirmatory research.

## 1.4 Validation of research questions

Due to the large number of stakeholders, the research questions are validated with the requirements that stakeholders have on the project. It is important to note that there is a distinction between *exploratory science* and *design science*, according to van Aken (2007). Exploratory science is concerned with describing, explaining and prediction empirical phenomena and design science is concerned with developing valid knowledge that can be used in the field. This project balances on both types or research the aim of this project is to generate knowledge that can be used in the field nevertheless, explanatory science needs to be conducted before the design science part can start. The role of this case study is mainly exploratory the design part is ensured by the data fusion project.

From Table 1-1 it is clear that all requirements from mentor, student and company are covered in the research. From the data fusion project only part of their requirements are treated ,this is because the project is still in it's orienting phase.

| Stakeholder | Requirements on the project | Included |
|---|---|---|
| The Student | Executable in 21 weeks | **Yes** |
| | Design orientation ensured | **Yes** |
| The Mentor | Topic should be within the domain of mentor (process mining) | **Yes** |
| The data fusion project | Topic should be within the domain and capabilities of the data fusion project? (field feedback, data merging) | **Yes** |
| | • Understanding the organisational structure | **Yes** |
| | • Understanding the feedback mechanism | **Yes** |
| | • Understanding the products | **Yes** |
| | • Evaluating the effectiveness of the field feedback mechanisms | **Yes** |
| | • Investigating the impact of data collection, procedures, regulations and policies | **Yes** |
| | • Cleaning/pre-processing | **Yes** |
| The Company | Explanation of the high percentage of sound product that return | **Yes** |

**Table 1-1; Primary requirements of different stakeholders related to the questions**

## 1.5 Document outline

This first chapter describes the problem situation in detail. The research context of quality management and a description of the company is given. Preliminary investigations are included to explain the current situation these are used to give direction to the research. This results in concrete research questions that are answered throughout the research.

The second chapter provides an aggregated view on the relevant academic literature in the different fields. This includes the field of ‘quality and reliability’ and ‘information quality’. This encompasses literature regarding customer feedback, maturity index of reliability, recent trends and information on data mining. Little new information is presented here it is mainly an overview of relevant scientific knowledge. In the third chapter ‘research method’, the approach of solving the research questions is treated. The forth chapter is more practically oriented and starts with an explanation on the present data flows regarding customer feedback. The dataflow resulting from these workflows are analysed in chapter five and the most relevant features are extracted. Chapter six covers outlier removal and data preparation. After these steps the dataset is finally ready for the mining, this is done in chapter seven.

The seventh and last chapter also contains the most interesting findings. Besides these findings, there is a discussion on the used methodology. Multiple follow-up projects are noted here as well, most of these are interesting for the company but some are also interesting for the academic society.

# 2 Theoretical Framework

The main aim of this chapter is to embed the research into the academic literature, and complement on the work done by previous researchers in this field. Besides this, background is provided on the research question limiting the amount of double work.

The problem at hand is specific for the company and industry, while the literature is generic. In order to use the available literature a conversion is required from specific to generic. The research is strongly related to both data mining and field feedback. The literature review is split in two main themes; the feedback literature and the data mining literature. Both themes require a different conversion from specific to generic. Figure 2-1 provides a picture of the setup of the literature review, showing the different topics.

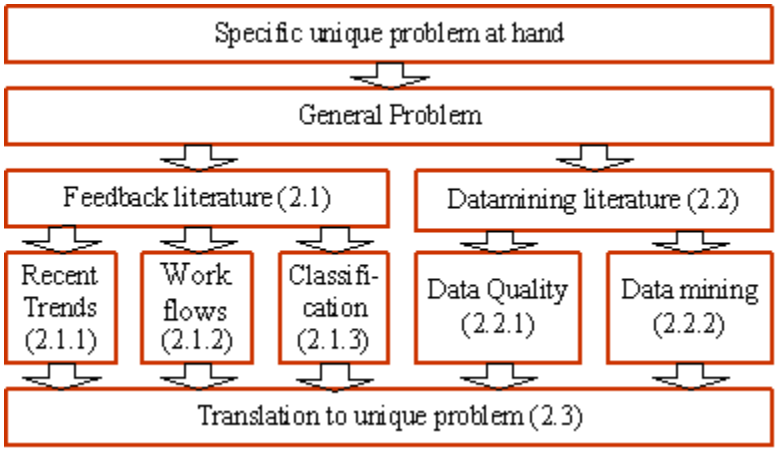


**Figure 2-1; Setup of the literature review**

From the feedback literature, three themes are included. Researchers sought causes for the high amount of NFF returns and came up with a number of trends. These trends occur in the manufacturing industry leading to the return of NFF products. This is interesting because it gives insight in the possible reasons for product return and this could lead to testable hypothesis when mining the data. Secondly, Magniez (2007) discussed the information flows using the Maturity Index of Reliability (MIR) method of Brombacher and Sander (1999), this is a logical and proven method to judge feedback flows. This is useful because it gives insight in the information flows over the different departments. Thirdly, classification between the different types of failures is discussed based on a paper by Koca (2006), this has benefit because it allows classifying errors when mining.

The data mining literature is less company specific and therefore better related to the problem at hand; less translation is required. The data mining literature is split up in two themes the data quality and the data-mining domain. The data quality domain concerned with rating and judging the quality of data. This is interesting in the light of the project because the results depend on the quality of the data that is put in the data mining tools. The data-mining literature is evidently included because the analysis will be done using these tools.

## *2.1 Field Feedback literature*

“It is not the strongest species that survive, nor the most intelligent, but the ones most responsive to change” (Charles Darwin). A similar statement can be made for companies in a competitive environment. According to Baker and Sinkula (1999), “The ultimate competitive advantage is to learn in order to respond and change faster than competition”

Three themes are discussed from the area of feedback literature. First, the recent trends that are estimated to be at the bases of the NFF return. Secondly, an elaboration is given on the research done by Magniez (2007) Thirdly, failure classification is covered; this tries to identify different models for analysing returns.

### 2.1.1 Trends causing No Fault Found

Before elaborating on the trends that cause the return of sound products, two important discrepancies between the problem at hand and the literature require explanation. First, the business model of the products is different from that described in the literature. The literature mainly discusses consumer products that follow the normal sales pattern. The product at hand is a business-to-business product and works with lease contracts. There is a large variety in lease contracts but generally, they are setup very much like a mobile phone contract. A number of copies per month are included in the contract, when the customer exceeds this amount extra payment for supplies must be made. The lease contracts generally also include service. In case of a machine defect the customer gets in contact with the Call Centre (CC). The CC will try to solve the problem over the phone in case this is not possible a service engineer is sent over. Secondly, the examined product is different from the products described in the literature. The literature mainly discusses consumer products that have one user. The product at hand has multiple users with different needs. Besides this, no one is owner of the product and the product is maintained by the working population.

During the 1990s, manufacturing industries were confronted with the return of working products Brombacher et al (2002). Those working products now carry the label NFF. The amount of returns that cover this description rapidly increased over time as indicated in Figure 1-1 on Page 1. Quality and reliability methods normally examine product with regard to their specifications. In the case these NFF returns, Quality and reliability methods do not apply because according to the specifications there is no technical failure. Nevertheless, these products return on the basis that they do not perform as expected. Several researchers show the pain that consumers face leading to product return: DenOuden (2005) indicates that the largest proportion of customers returned the product on basis of "Didn't work as thought it would" as displayed in Figure 2-2. A research done by WDS global indicates that the largest proportion is due to the fact that customer is not able to work with the product. We must note that experiments are conducted in different consumer markets.

| Reason for return (WDS) | Percentage | Reason for return (DenOuden) | Percentage |
|---|---|---|---|
| **Technical** | **36%** | **Technical** | **52%** |
| Hardware fault | 24% | Product broken | 52% |
| Software fault | 12% | | |
| **Non Technical** | **64%** | **Non technical** | **48%** |
| Handset configuration issue | 31% | Didn't work as thought it would | 28% |
| Struggling with functionality/usability | 24% | No explanation | 7% |
| Device miss-sold/ does not meet expectance | 8% | No Longer wanted | 3% |
| | | Had already one | 2% |
| | | Wanted a different one | 2% |
| | | Other | 5% |

**Figure 2-2; Reasons for product return by WDS global (Jul 2006) and DenOuden (Oct 2002)**

Several reasons are the basis for the return of sound products. Many researchers have found contradictory trends explaining these returns. An overview of the most influential trends is provided in the subsequent paragraphs.

#### 2.1.1.1 Increase of consumer expectation,

An explanation resides in a relative old model developed by Kano (1984). The model is shown in Figure 2-3. The model of Kano explains the perception of products to the consumer. The model is best explained using examples of consumers that are looking for products. The green 'basic' line refers to functions that must be present in a product. For a car, such features are brakes or a door. Without these attributes, the car would not be safe and the consumer will not buy them. A car with an extremely good braking system does not draw a large amount of extra attention. Therefore, the green basic line is limited to only a certain level of satisfaction. The blue line marked performance can be explained by an example of digital cameras. The customer satisfaction gets higher with the number of megapixels that are present on the device but the increase is linear. For this reason, these satisfiers are called linear satisfiers. The last form of satisfiers are the, noted in red, excitement or delighters. This includes surprisingly new features that consumers did not expect. For example, the feature that pictures rotate automatically on the camera when the consumer turns his digital camera or for example, a light that switches on when you open the glove compartment in your car. These functions are new and exiting to consumers. After a while, these functions will be seen as normal and new satisfiers need to be introduced.

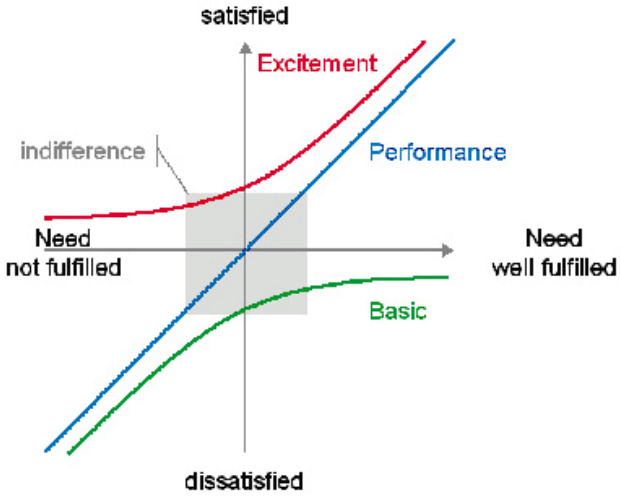


**Figure 2-3; Consumer satisfiers by Kano (1984)**

Oliver (1997) found that the satisfaction level of the product is dependent on the consumer's prerequisites of the product. Table 2-1 displays an example of two customers, buying a car. At the store, two customers are given different expectations of the gasoline usage of a car. Afterward the car performs better or worse than expected.

| | Customer A | Customer B |
|---|---|---|
| Expectation | 15 km/L | 25 km/L |
| Performance outcome | 20 km/L | 20 km/L |
| Numeric difference | +5 km/L | -5km |
| Subjective disconfirmation | Better than expected | Worse than expected |
| Expressed | "Great" | "Mad as hell" |

**Table 2-1; Example of satisfaction levels by Oliver (1997)**

Consumers that expect better specifications from a product are unsatisfied while consumers that expected worse from the product are extremely satisfied. When marketing promotes articles as small

wonders it might increase the number of sales but it will also increase the number of unsatisfied customers DenOuden (2005).

You might start wondering what all these marketing characteristics have in common with the field of quality and reliability. Berden (1999) found that there is a trend going on that might explain the large number of NFF returns. The expectations that consumers have on products are increasing. Customers are expecting better products that last for a longer period of time this expansion is displayed in Figure 2-4 leading to returned products.

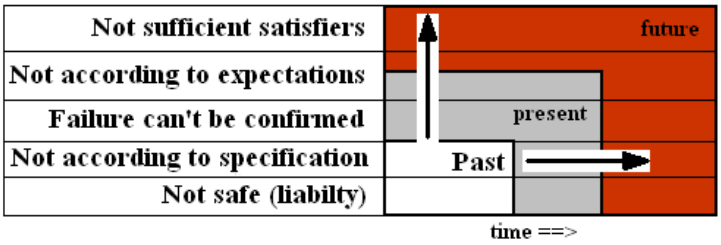


**Figure 2-4; Trends in warranty coverage by Berden et al (1999)**

#### 2.1.1.2 Increase in product complexity

In order to keep up with the market manufacturers keep improving their products. This is completely logical after the explanation in the previous paragraph. The expectations of the consumer are increasing and therefore products need to contain more satisfiers. This extra functionality makes products technically more complex.

The increased functionality will complicate the prediction of quality and reliability problems. With each extra feature that is implemented the number of interfaces needs to be considered. Magniez (2007) states that these interfaces are presented on three intersections; internal to the product (subsystem level), Product-user interaction and product to product (interfacing and connections). This indicates that adding features to a product has the tendency to quickly complicate matters.

#### 2.1.1.3 Outsourcing, globalization and segmentation

The study from Ragatz (1997) indicates that when developing innovative products depth and early integrations are necessary to reduce cycle times, improve quality and reduce cost. Globalization and segmentation make business processes increasingly complex. The main tread with regard to quality and reliability is the bad deployment, loss or delay of information. This lack of adequate information and the segmentation of the development teams have a negative influence on the depth and early integration in development teams.

Companies operate on a global market and their aim is to optimise performance on a global level, in order to do so outsourcing has large advantages. The main results in the field of quality and reliability are once again the loss or delay of information. To prevent this from occurring well-structured collaboration is needed between outsourcer and supplier (Magniez, 2007). The company under consideration does produce certain parts in lower cost regions, for this reason this is an interesting trend.

#### 2.1.1.4 Time to market

In development, the time to market is the time it takes to develop a product and get it ready for the market. Being first on the market with a new product has large competitive advantages such as product knowledge and market share. This trend is visualised by an example from the media player industry in Figure 2-5. Three media player technologies are covered in the research by Zia, O (2005) the VCR, DVD and DVD+R. The increasing line is the number of features included in the product and the decreasing line is the price of the products. The price drop and the number of features on the

DVD+R evolved in a faster pace that the VCR indicating that the technology runs on a increasing pace.

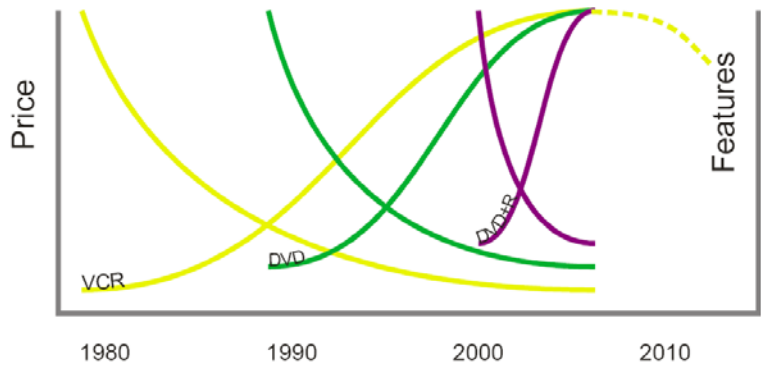


**Figure 2-5; The increasing pace of technology by Zia, O (1995)**

In order to be the first on the market or even keep up with the market tremendous pressure is put on the product development team to develop the product as quickly as possible. This has its risks with regard to quality and reliability there is for example no time to conduct fully factorial testing. Besides this, the product is more than often placed on the market before all the initial errors are removed from the product. The company accepts that a not fully tested product that may contain errors is put on the market, this results in a larger number of product that return faulty.

#### 2.1.1.5 Online buying and consumer power

New customers buy a large variety of their products online. Online buying results in buying products without having seem them before or testing them in practice. There are a rich variety of forums of previous users that share their experiences about products that helps making up your mind about the quality of products. There is a large difference in buying your product in a store or buying the product online. In the store, you can really see, feel, hear and touch the product versus online where you can only see the product. The consumer will try the product at home instead of in the store. The next trend, consumer power enforces the possibility of the consumer to return the product when it is not satisfied.

The consumer is aware of their power and uses this when it is not satisfied about products. Even websites are setup encouraging people to complain when they are unhappy about products. The combination of both trends might result in people who buy the wrong products online and have the rights to return these products because it does not fit their needs. Pitt et al. (2002) also describes this in a paper called “The internet and the birth of consumer power”. The online buying is not so much relevant for this research however, the consumer power is essential. Because this might lead to access replacement.

It is useful to look at the processes concerned with the return of products. Multiple techniques are developed to examine these processes. The next chapter discusses some of these techniques.

### *2.1.2 Workflows at the company*

The model of Sander and Brombacher (1999) is used to judge information flows. A high MIR level corresponds closely to continuous improvement. This is relevant because this enables companies to react swiftly to market changes and to deal with consumer complaints. First the model of Sander and Brombacher is explained, and next the quality flows within the company are treated.

The model by Sander and Brombacher differentiates between different phases; In the *uncontrolled phase (<MIR1)* the company is unaware of the field behaviour of products. There is no feedback from the field to the company; no systems are in place to ensure improvement. In the subsequent *measure phase (MIR1),* the company receives feedback from the field however there is no knowledge on the origin of the failures of products. The higher *analyse phase (MIR2)* includes the previous phase and

next to the feedback this information is analyzed and the manufacturer is aware of the root causes of the problem but he is not able to do something about it. The *Controlled phase (MIR3)* includes MIR2 and next to this, the manufacturer is able to take action and improves the situation. Yet the manufacturer is not able to prevent similar events happening in the future. The *Learning (MIR4)* phase indicates that the manufacturer has quantitative knowledge on the problems that occur in the field. The manufacturer is able to find the root causes, take preventive actions and prevent similar actions of occurring in the future.

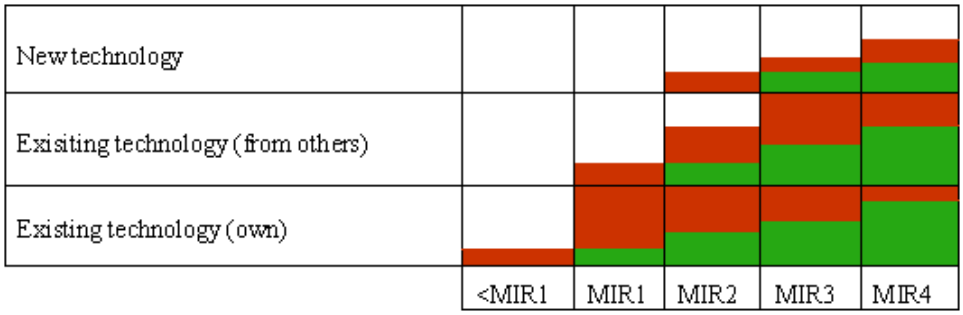


**Figure 2-6; The MIR level by Sander and Brombacher (1999)**

Figure 2-6 is a graphical representation of the latter described in words. The red blocks indicate that some of the problems can be solved. The green blocks indicate that the problems can be prevented. From the distinction between the different technologies, it is clear that innovative companies that deal with new technologies are required to have a high MIR level. The higher the MIR level the better a company is able to prevent future quality issues in new products. The described model is used to evaluate the feedback mechanisms in place. On a department level, the way feedback is handled and used can be measured.

This model explains why it is important to be able to respond quickly to market changes, and how this feedback is measurable. Figure 2-7 shows the results of the analysis by Magniez, C. (2007). The link between the Customer and the Call center is on a MIR level 1 because this information is not used further in the company. The MIR levels go up as problems become more escalated. When the problem gets escalated to a specialist the MIR level is 4 meaning that this information is used for design improvement, even for future designs.

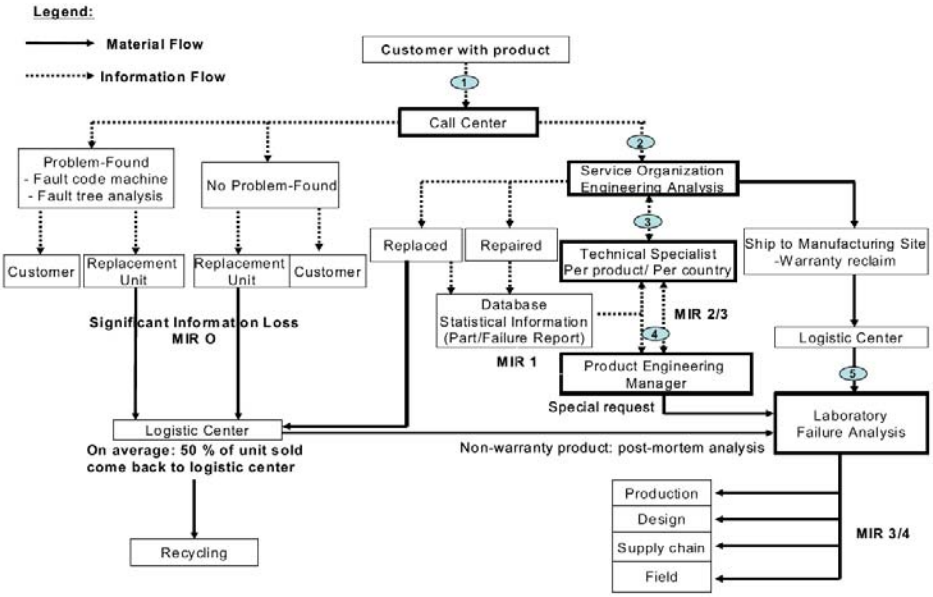


**Figure 2-7; The workflow created by Magniez, C (2007)**

### 2.1.3 Classification of failures

Creating different classes with similar symptoms is useful in order to do analysis on them because they have similar characteristics. There is no generally accepted model to classify these errors. There are a few models proposed in the literature Koca et al (2006) made an overview between two models.

Te first model under consideration is depicted in Figure 2-8 the model is created from the field and the categories are determined by service organisations. This type of distinction incorporates a problem when a failure could not be found it does not indicate that a failure is not present. This is depicted by the arrow from "Error in service diagnostics" to "fault found". According to Koca et al (2006) the coupling with the two categories yields a highly undesirable fuzziness when it comes to applying it in the field. This model closely resembles the categorisation used by the company under consideration however; there are more causes to No Fault Found than only "product does not meet expectations". For example, the processes in place can very well result in the return of sound products. Another difference is that products with a diagnosed fault could be in fact "no fault found" products. This is due to for example; handling damages or testing sequences that reveal faults the consumer did not.

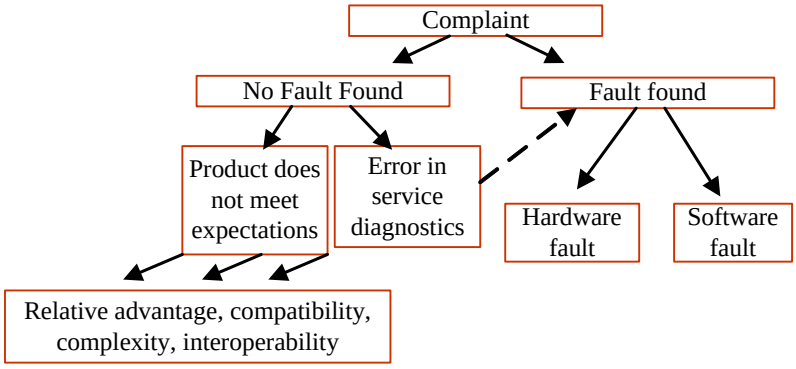


**Figure 2-8; Field Call Categories from a service perspective by Koca et al (2006)**

The second model is a model proposed by the academic society and shown in Figure 2-9, this model has more theoretical grounds but is less applicable in industry because the detail is often missing. The model separates two main causes. *"hard failures"* are failures where the product is not able to meet the explicit technical specifications and hence the consumer requirements. *"Soft failures"* are failures where the consumer complains on the (lack of) functionality of the product, despite the product meeting its specifications. Bare in mind that soft failures is not a abbriviation of software failure and there is no link between the both of them.

Koca (2006) indicates that "this categorization is no adequate for recognizing and classifying failures whose symptoms cannot readily be diagnosed an/or whose root causes cannot be readily identified. This is what researcher call the gray area." The classification is sooted for the academic society but for the industiral world the model is not usefull because identification is difficult if not impossible. The company under consideration in this research seems to suffer from other type of failures besides the two classifications of Geurdens.

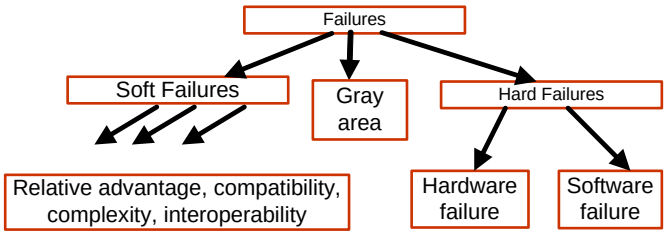


**Figure 2-9; Failure categories from the academic perspective by Geudens et al (2005)**

The reason for these different models is clear however, it does not improve the interpretability of failures. The NFF diagnose often drawn by service agents cannot be related to the academic literature because NFF is not the same as soft failures.

The classification model proposed by Koka (2006) seems to relate better to practice and relates better to the current way of working within the company. For these reasons, this model is used throughout the research. Koca (2006) classified the returned cases by hand. Especially with large returns flows this proves problematic. Data mining software might be able to deal with great amounts of data and thus with large amounts of product return.

## 2.2 Data mining literature

On the combination of data availability and exploratory nature of the research, a data mining approach is chosen. This chapter sets the basis for this analysis by looking into information quality and data mining tools.

### 2.2.1 Information Quality

Throughout the project there is major emphasis on knowledge, a lot of research is done in the area of knowledge management and some concepts are important to be aware of.

#### 2.2.1.1 Different sorts and types of information

Hislop has an interesting perspective on data, information and knowledge. According to Hislop (2005) *data* are representations of observations and measurements. *Information* is data arranged in a meaningful pattern. *Knowledge* is data or information that is structured and linked with existing systems of beliefs and bodies of knowledge. All the three terms contain information; the distinction is the way the information is presented to the world. Within our field this would imply that feedback information might contain a lot of useful knowledge but it needs to be properly mined before it can be used.

The three types of information noted by Hislop (2005) are explicit. Explicit knowledge is propositional, articulated knowledge. The opposite of explicit knowledge is tacit knowledge, best explained by a famous statement of Polanyi (1966) “we can know more than we can tell” and tacit knowledge is knowledge that is used but cannot be articulated. Examples of this are perceptions, but also skills, discovery etc. Explicit knowledge is rooted in tacit knowledge. A sequent step is to write down this explicit knowledge this knowledge is called codified knowledge. Generally the following sequence is used: Tacit (only in thought) => Explicit (the tacit knowledge is made verbal) => Codified (the knowledge is stored in a database).

Another aspect of knowledge is the validity; whether knowledge is true or false. There is a factor that makes this complex, and there are a lot of gray areas. To illustrate these areas a small thought example; Statement: “Water boils at 100 $^0$C” This is generally speaking true, but if we look at the statement more specifically there are cases imaginable where this is not the case. If we increase the pressure, water has the property to boil at lower temperatures. Thus this statement has its boundaries,. We should improve the statement by changing it to “water boils at 100 $^0$C under atmospheric pressure”. This is closer toward the truth, although not exactly right; because when salt is solved in water it tends to boil at a higher temperature. We need to improve our statement once again: “H2O boils at 100 $^0$C under atmospheric pressure”. It requires a lot of effort to get statements written down so that they are correct under all circumstances.

Data warehouses are able to store only codified knowledge. This means a lot of information is lost in comparison to the available tacit knowledge. The information in the data-warehouse is even submitted to jet another factor that decreases the information more. At helpdesks, customers are asked to describe the problem over the phone. This problem is then interpreted by the employee and recorded in the system. A lot of information gets lost in this process.

Research on the performance of businesses regarding the available knowledge have been executed, Haas & Hansen did such research. Haas&Hansen (2005) explain in their paper that knowledge is often necessary, but not a sufficient condition for performance, task units are not always better off obtaining and using more knowledge. Next to this, relying on codified knowledge may result in work that lacks innovation or customization. If we translate these finding we see that it is essential that

relying only on a data warehouse is not a wise option. Besides the data warehouse sources of reliability information other sources should be present. This seems logical looking at the concept of tacit knowledge, not everything can be recorded in codified knowledge, and therefore interaction will stay essential.

#### 2.2.1.2 Four dimensions of information quality

The paper of Wang and Strong is included because they refer to intrinsic data quality. Wang&Strong (1996) state that data needs to fulfil some basic requirements; Accuracy, Relevancy, Representation and Accessibility IQ. These measures have been identified using factor analysis on the ratings of data consumers on different data quality aspects.

The first measure *data accuracy* refers to the intrinsic quality of the information. In other words, is the data correct, objective and does de data come from a reputable source. The second measure *relevance* is regarding the relevance and timeliness of the data and referred to by Lee (2002) as Contextual IQ. The third measure *representation* refers to how the data is represented, the data should not be in a foreign language, the length should be appropriate etc. The last measure *accessibility* refers to the way the consumer can get to the data.

The research by Wang and Strong (1996) suffers from low internal validity, the researchers only surveyed data consumers, and there was not a conceptual model at the basis of their findings. At the time, such a conceptual model was not present, a few years later the meta analysis by Lee (2002) changed this. Lee (2002) reviewed multiple conceptual models and statistical findings, summarizing this he came with a more valid model. With the background of Wang and Strong and the verification of their research by Lee (2002) the information quality criteria can be reflected at the field feedback information. Within the research, the information quality criteria can be used to validate the field feedback data.

### *2.2.2 Data mining*

In order to make valid analysis on the datasets different data mining software packages are analysed. Data mining is the process of extracting hidden patterns from large amounts of data (www.wikipedia.org) often also called knowledge discovery in databases (KDD). In terms that are more scientific this can be described as transferring information to knowledge. According to Fayyad et al (1996) data mining is related too many other information technology fields including machine learning, statistics, AI, and databases. This chapter will not go into these fields and stick with the tools available to mine the data. The mining tools under consideration are a selection of open source software from a known data mining community http://www.kdnuggets.com. This list is not supposed to be complete, but a sample of the available dataset.

This chapter consists of three parts; first, the data mining process explains that steps that are required before the analysis can be done. Second, the different mining tasks are grouped. Third, the different mining tools are analysed on the basis of the mining tasks defined in the second part.

#### 2.2.2.1 Data mining process

The rapid growth of digital available data consequently is followed by a rapid growth of tools dealing with this data. The process of generation knowledge out of this data is called ‘data mining’ or ‘knowledge discovery in databases’ (KDD). Figure 2-10 provides an overview of the steps that a mining process is composed about. Nevertheless, the process is interactive and iterative, involving numerous steps with many decisions made by the user Brachman and Anahad (1996). This includes many iterative steps, the overview is therefore an indication of the project flow while in fact deviation from this flow is practice.

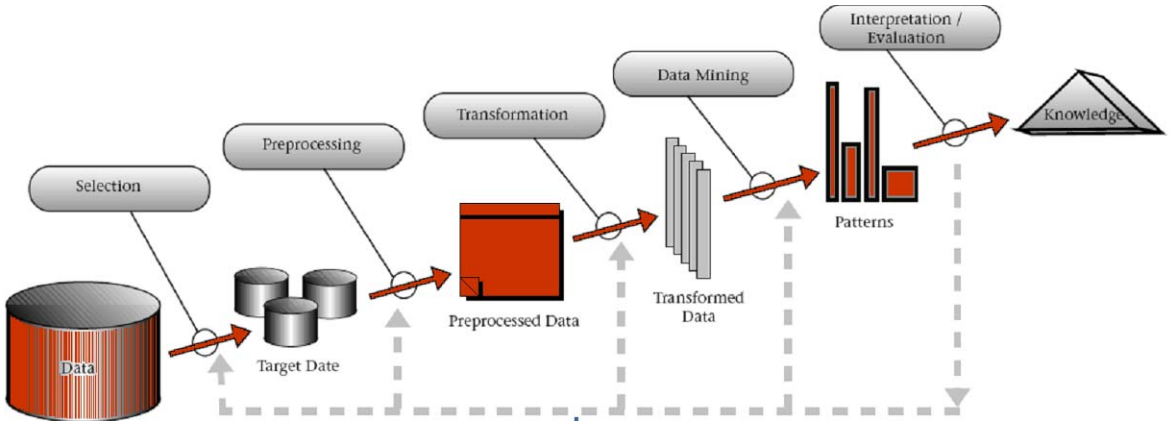


**Figure 2-10; An Overview of the steps that compose the mining process by Fayyad et al (1996)**

The data mining sequence starts with gaining access to the required databases and an understanding of the application domain, goals and targets. Next, a selection can be made from the larger dataset. In order to do so the scope and goals of the project need to be well defined. The smaller extracted set of data can then be cleaned and pre-processed in our case this also includes the merging of different data sources. This will lead to a dataset that after transformation can be used for mining. A mining tool next shows patterns that can be interpreted to knowledge. This step has received most attention in the literature nevertheless Fayyad et al (1996) stress that the steps leading toward this analysis are as important (and probably more so) as the data mining itself. The next paragraph will describe three data-mining methods classification, Clustering, Regression that are used in the mining step.

### 2.2.2.2 Different mining methods

Fayyad et al (1996) define three different mining methods; Classification, Clustering and Regresion. The Authors state that most (if not all) methods can be viewed as extensions of a these basic techniques. The actual underlying technique can be one or a composition of mathematical techniques such as polynomials, splines, kernel and basis functions, threshold-Boolean functions etc.

The first method classification maps the cases into predefined classes. An example of this can be found in the area of banking where loans are given to customers that belong to a certain class. Customers that have not defaulted on their loans and have a high salary are allowed to take another loan. The chance of customers in this segment to not default on their lawns is lower than, customers not belonging to this class.

The second method Clustering is a more descriptive task where a finite set of clusters is identified that describe the data. Rules explain the clusters cratered by the software. Important to note is that this are not predefined groups, these groups are created by exploratory data analysis. The task Probability density estimation is closely related to clustering but it tries to find the joint multivariate probability density function of all of the variables. Defining different target customer groups is an example for this use.

The third method Regression is a method to fit a function in the dataset; this function can be linear and non-linear. The method tries to minimise the error when the line is fit trough the data. Applications are many for example the ability to flirt with women and the number of beers consumed has a strong relation.

The data mining software packages often include the functionality described above. More than often the names differ and the underlying techniques are more or less sophisticated. Besides this basic functionality, extra functionality might be added in the data mining software packages. A few examples of these are decision trees and rules, nonlinear regression and classification methods, example based methods, probabilistic graphic dependency models and relational learning models. The following paragraph analyses the different data mining software packages on the inclusion of the different foundation mining tools.

#### 2.2.2.3 Data mining tools

The fifteen data mining software packages under consideration are a selection of open source software from a known data mining community (http://www.kdnuggets.com). This list is not supposed to be complete, but a provides a sample of the available data mining software.

Adam (Algorithm Development and Mining system) is a software application developed by the University of Alabama. Originally designed for mining and image processing it has gained popularity in analysing images of the universe. Remarkable conclusions about the universe have been drawn with the aid of this tool.

The Alpha miner is a Java based application developed by the University of Hong Kong. According to their website, the software is suited for analysing data from E-commerce. The software allows setting patterns that can be used on different datasets allowing drag and drop support. This makes it easy for manager to do analysis they do not know the background about.

The Databionic ESOM tool is developed at the university of Marburg in Germany. The generally applicable tool has a rich set of analysis such as Clustering and Classification.

The Gnome Data Mining Tool is created by a company named Togaware. The package is best suited for a Linux operating system and it can easily work with LaTeX. It has a rich set of Bayes classifiers and decision trees. However, regression is not really implemented in the software and it is mainly code driven.

The KEEL (Knowledge Extracting based on Evolutionary Learning) tool is developed to assess evolutionary algorithms. It includes regression, classification, clustering, pattern mining and so on. The tool has been developed for two purposes research and education. Seen the purpose it is of no surprise that the origin of this software is a collaboration of research groups in Spain. The GUI is quite limited and most functionality is gained using the coding.

The University of Konstanz in Germany created a mining tool KNIME. The base version includes over 100 processing nodes for data. Besides this rich functionality, additional plugins from WEKA can be executed. The package has a user-friendly interface and supports a rich set of input files.

The Machine Learning in Java package is, as the name already suggests, highly text driven there is a limited GUI. The functionality is limited to Categorization and tree inductions.

Germany is contributing a lot to the open data mining software, the university of Dortmunt created MiningMart. The aim of this package is to analyse customer profiling.

The MLC++ package is mainly code driven. The algorithms beneath originate from the field of machine learning. The Orange software is also mainly code driven but it has the ability to add Orange Widgets, these are components to provide a GUI. It is originally developed to work with another package called Python.

Formerly called Yale Rapid Miner is a mining package that is also concerned with the collection of data. The data is collected by users on a website, this data can be immediately analysed. It is written to operate on a server. This package is not really suited because the dataset is already available.

Togaware developed ‘Gnome data mining tools’ and also the Rattle package. The Rattle (the R analytical Tool To Learn Easily) package is based on the R software and has limited GUI. The main objective of this software is to build supervised and unsupervised models.

The StarProbe package is originally designed to run on a server and provide real time analysis on a website. This package is able to run the analysis online without the need of installing software. However a server is needed to run the package on.

TANAGRA is the successor of SPINA and is developed by the University Lumière in Lyon. The main objective of the software is to do exploratory data analysis. It includes the three basic components but extensions are limited.

The Weka includes both a user interface the ability to work with code. However, the functionality offered in the coding is far superior to the ability offered by the GUI. This package is recommended by the IT specialist from the data fusion project.

| Tool | Language | Gui | Data cleaning | Classification | Clustering | Regression | Developed by |
|---|---|---|---|---|---|---|---|
| ADaM | C++ | Y | Y | Y | Y | Y | University of Alabama |
| AlphaMiner | Java | Y | - | Y | Y | Y | University of Hong Kong |
| Databionic ESOM Tools | Java | Y | Y | Y | Y | Y | University of Marburg (Germany) |
| Gnome Data Mining Tools | C++ | N | N | Y | Y | N | Togaware |
| KEEL | Java | Y | Y | Y | Y | Y | Collaboration of research groups |
| KNIME | Java | Y | Y | Y | Y | Y | University of Konstanz |
| Machine Learning in Java (MLJ) | Java | N | Y | Y | Y | N | Kansas State University |
| MiningMart | Java | Y | Y | Y | Y | Y | University of Dortmunt |
| MLC++ | C++ | N | N | Y | Y | Y | Stanford University |
| Orange | C++ | Y | N | Y | Y | Y | University of Ljubljana |
| RapidMiner (formerly YALE) | Java, html | Y | N | Y | Y | Y | Rapid I |
| Rattle | R | Y | N | Y | Y | Y | Togaware |
| StarProbe | HTTP | Y | Y | Y | Y | Y | CMS |
| TANAGRA | VB | Y | N | Y | Y | Y | Univerisity Lumière Lyon |
| Weka | Java | Y | Y | Y | Y | Y | Pantaho |

After reviewing the different software packages, the following observation can be made; There are many software packages available with overlapping functionality. Most of the software packages contain the three types of data mining techniques. However, the underlying calculation can differ from one package to another. Besides this the ease of operation is different over the packages some of them contain a more detailed Graphical User interface while others only work with code. Besides this the different software packages can work with different types of databases. Some have unique features and work on specific platforms. All together, there is a lot of mining software available; the question is weather all this functionality is needed in the project. Most likely, if the simpler software packages are not able to find something the more sophisticated will not either.

## 2.3 Summary of the theoretical framework

This literature review addresses two types of literature: field feedback and the data mining.

The first has a relation with the project because trends in the return of NFF products are researched for consumer products. The product in this project however is a Business-to-Business product nevertheless the literature is interesting to review. Interpretation is required to transform the information to the product at hand. Another difference is that the consumer does not directly interact with the module but with the entire product. The trends described by the literature do resemble the activities happening at the company and industry. This includes the increase in product complexity, outsourcing, globalization, segmentation, time to market, online buying and consumer power.

The Second topic data mining describes the process to get from the raw data to knowledge. This process involves five steps: Selection, Pre-processing, Transformation, Data mining and Interpretation/ evaluation. Although data mining seem to be the most essential, the antecedent steps of the data mining are important (and maybe even more so) than the data mine itself. The process seems linear in nature however, practice shows that it is iterative and many decisions need to be made. Looking at the data mining there are in essence three data mining techniques; Classifications, Clustering and regression these techniques have numerous extensions. Fifteen freely available software packages are analysed and judged on their functionality and ease of operation. Most of the packages include all three types of analysis however ones are more sophisticated than others. Nevertheless, conclusions found in less sophisticated tool will also be present in other more sophisticated tools. The KNIME tool seems like a good option, it has a nice user interface, large functionality and works with many data-types.

# 3 Research Method

In the research proposal of the IOP project Y.Lu (2008) states: “In order to deal with these increasing and often conflicting trends, the development organization requires that field feedback information be in detail and much earlier available in the development process so as to deliver the products to meet customer requirements.” Within the company a lot of information is present, but no one is currently in the possession of crafts to deal with this information. According to Lyman (2000):”The amount of data doubles every three years, data mining is becoming an increasingly important tool to transform this data into information”

The direction of data mining is chosen to solve the NFF problem occurring at the company. The CRISP-DM reference model is an often-used method to do data mining; this tool is explained in the following paragraph. Besides this data mining tool, each research question requires different techniques for answering. Clarification on these techniques is given in the second paragraph.

## 3.1 The Methodology

The proposed project is a data-mining project. The CRISP-DM reference model describes this type of project in detail. This reference model will be used to setup the research. The authors stress that the model is not scientifically setup but originated from practice. Figure 3-1 is a graphical overview of the different phases of the CRISP-DM reference model. It is important to know that a data-mining project does not structure this easily and many jumps from one to another phase not noted in the graph are required. The iterations happening more often are noted with the double arrows like between ‘business understanding’ and ‘data understanding’. The different phases of the CRISP-DM reference model are up for discussion next.

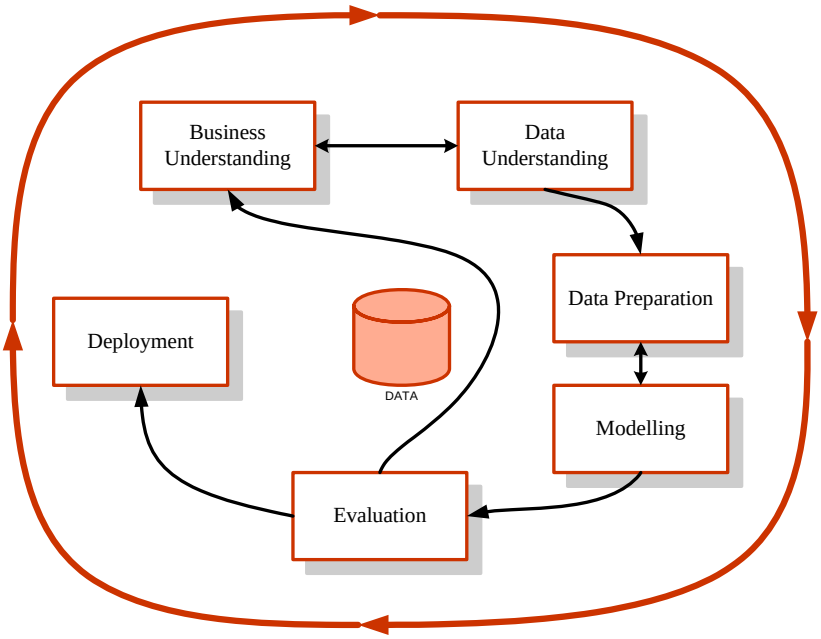


**Figure 3-1; Phases of the CRISP-DM reference model**

### 3.1.1 Business understanding

The main objective of this phase is to understand the business objectives and the need for data mining, this part is largely covered in the first chapter of the research proposal. This knowledge is used to create a data mining problem definition.

### 3.1.2 Data understanding

The second phase 'data understanding' is to collect, understand and analyse the dataset. The analyses done are to get insight into the quality of the data and to identify data quality problems. For more information about this please read Chapter 4. In this chapter, the sources and content of the databases are discussed.

### 3.1.3 Data preparation

In the 'data preparation phase' a final dataset is constructed that will be fed into the modelling tools. This will require combining data from different sources called data fusion. The CRISP-DM reference model does note that fusion might be needed to get data in the right format. However, not to such a extend as is required in our case.

### 3.1.4 Modelling

In the modelling phase various modelling techniques are applied on the final dataset and analysis are done. Most likely stepping back and forth from data preparation to modelling is required in order to deal with different mining tools. Fayyat et al. (1996) elaborated on the approach required by data mining; the two methods are quite similar with the distinction that CRIPS-DM provides a more global view. Fayyad et al. focus more on the details of performing data mining; they can be used side by side.

### 3.1.5 Evaluation

This phase is concerned with the evaluation of the results. In order to check if predictions made by the data mine match the reality a new sample is used. The sample of 100 units is first analysed using the data model. Next, the modules are analysed in the tradition method. The predicted and the real value are next compared.

### 3.1.6 Deployment

The deployment phase is not included in the project, because it is irrelevant to the research besides there is a time constraint. In the conclusions, a model explaining the causes of NFF returns is present. The model will point to directions for future projects that can emerge form this project. These follow-up projects can deploy the gained knowledge into the company.

| Business understanding | Data understanding | Data preparation | Modelling | Evaluation | Deployment |
|---|---|---|---|---|---|
| • Determine business objectives<br>• Assess Situation<br>• Determine Data mining Goals<br>• Produce Project Plan | • Collect Initial data<br>• Describe data<br>• Explore data<br>• Verify data quality | • Select data<br>• Clean data<br>• Construct data<br>• Integrate data<br>• Format data | • Select modelling technique<br>• Generate test design<br>• Build model<br>• Assess model | • Evaluate results<br>• Review Process<br>• Determine next steps | • Plan Deployment<br>• Plan Monitoring/ Maintenance<br>• Produce Final report<br>• Review Project |

**Table 3-1; Tasks and deliverables of CRISP-DM reference model**

## 3.2 Actions per research question

From each question, the methodology and the resources that are needed to carry out the analysis will be described.

*What field feedback systems are currently in place?* The methodology is an unstructured interviewing technique in the departments that are noted by Magniez C. (2007). This includes the following departments; Call Centre, Service Organization Engineering Analysis, Technical Specialist, Product Engineering manager, Laboratory failure analysis.

*What information regarding the NFF returned products is available, and what features/attributes are important.* Make a data diagram using the Unified Modelling Language (UML) of the relevant information from the different SHI(F)TT databases. The information from these databases is required in order to do so. After defining the UML model this can be used to define the most interesting data attributes that need to be incorporated in the model that explains the NFF returns.

*What data mining techniques are relevant to the case study?* This research question requires a multi step approach. The first step encompasses a literature review resulting in different techniques. The second step is a comparison between techniques on the grounds of applicability to the problem at hand. This comparison is next used to assess the most applicable tool.

*What kind of relations can be found using these data mining tools?* This research question is dependant on the selected mining techniques from the previous question. Most effort will be on the merging of different data sources before relations can be found. The results of this analysis heavily depend on the work done in the first three research questions.

*Validate these results with practice.* Bring the results back to the company and the involved parties. Discuss weather this results are logical and what the implications are. Next, find possible improvements.

# 4 Business understanding

A data mining project does not follow a linear pattern and is known to have many iterations between data collection and analysis. However, in the following chapters a linear path is described from problem to solution. This is possible because the iterations that did not contribute to the result are excluded from the documentation. Whenever questions remain, please contact the author for clarification. This chapter describes two steps from the CRISP-DM methodology namely: business understanding and data understanding.

## 4.1 Feedback systems currently in place

The company under consideration has multiple quality feedback mechanisms that flow from the customer to the different departments. Figure 4-1 displays the information flows regarding product quality in the field. This figure requires some clarification as we can distinguish three main flows. Each of these flows involves different parties and different databases.

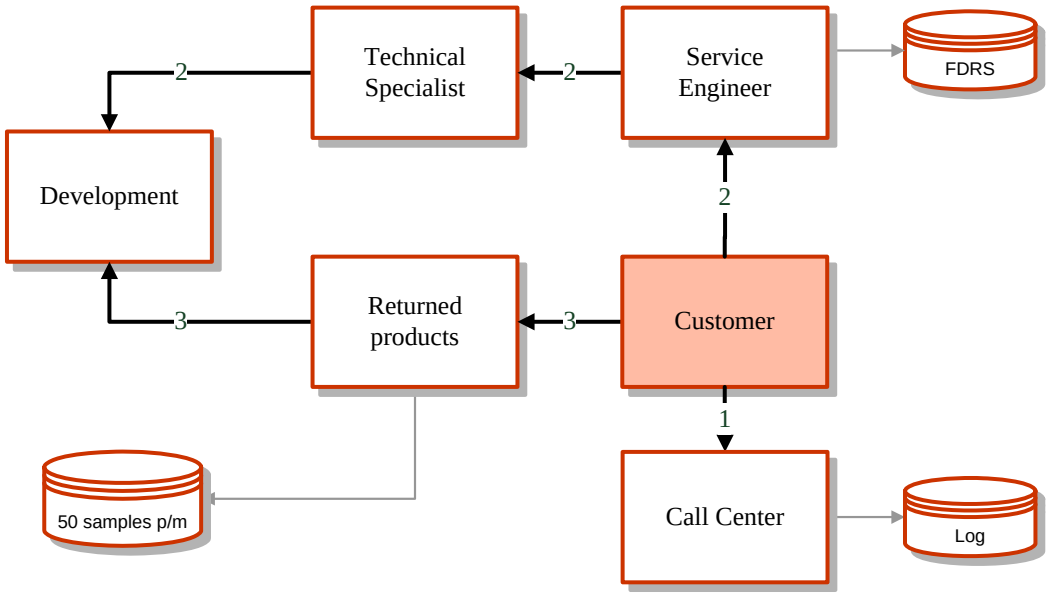


**Figure 4-1; Data flow diagram**

### 4.1.1 The Call Centre

The call centre (CC) is the place customers contact when problems occur or when supplies need to be ordered. This call centre is organised around cost efficiency, this involves moving the site over the world to most cost effective locations. The CC has two main targets. First: call centre employees are evaluated on the number of calls they handle per day. Secondly, call centre agents are evaluated on their ability to solve 30% of the calls at once. In order to meet these criteria call centre agents will try to solve calls quickly over the phone.

Calls are evaluated with the aid of a flow chart. Because each call is evaluated individually, there is no link between an earlier call. This means that there is no memory in the call centre. When a customer calls again, the call centre agent does not remember the previous call.

Conclusions made by the call centre have a large effect on the systems downstream. The first target; handling a large amount of calls might conflict with the quality of handling the calls. The already present reduction in information quality due to the translation from tacit to codified knowledge is enforced due to the urge of call centre agents to handle the call quickly. The second target states that

call centre agents are required to solve a certain proportion of the calls at once. This means that call centre agents are required to solve the problem over the phone this results in less funded conclusions and might lead to excess replacement of modules. The excess replacement of modules is also supported by the paper from Magniez (2007). It is essential that the targets posed on the Call Centre are defined carefully.

Excess replacement of modules does not pose a financial problem as long as it is cheaper than sending a service engineer to the customer. However, it proves difficult to measure the financial effects because no knowledge is available on how many modules are needlessly replaced. Besides the financial effects, there are also the indirect effects of customer service. According to J.Croning (2000) these effects include customer satisfaction.

### 4.1.2 The Service engineer

In the case that problems are not solvable over the phone, a service engineer is sent to the customer. The costs of a service engineer visit exceed the cost of sending a new module to the customer. When sending a service engineer the problem is always solved. When the engineer is not able to solve the problem, he will visit again the next day. This is in contrast with sending a module to the customer; in this case, there is no guarantee that the problem is solved.

The service engineer captures his observations in a Field Database Reporting System (FDRS). In the FDRS database, information is logged regarding problem description, type of failure and material use. Besides the FDRS database the service engineer extensively uses a knowledge repository called Eureka. This knowledge repository is maintained by the engineers and is used by the engineers. Engineers report solutions for problems ranging from very basic often-occurring problems to unique one of kind problems. Engineers that use the solutions rate the information, indicating what information can best be used. With the field visit, it became clear that engineers use both systems extensively and have a lot of contact with each other over the phone.

Two types of engineers can be distinguished brand specific and brand neutral engineers. As the name suggests brand specific engineers only operate on one brand while neutral engineers service different brands of machines. Both type of engineer service a wide range of machine types. For each product type there is a technical specialist that on a quarterly basis has feedback sessions with the development department regarding the quality of the products.

### 4.1.3 Returned products

In the market under consideration, it is normal that supplies are returned for refurbishment. The third stream of customer feedback concerns the return of physical products. The development department analyses a batch of these returns every month. This information is used by product developers to improve future designs, and to be aware of the quality issues in the field. The information and analysis are stored in a local database only accessible by the development department. In the beginning of the product life these analyses are more extensively done.

These analyses are done by a test engineer according to the following sequence. First, visual inspection is performed to ensure that all the parts are present and see if there is a visible mall function. Secondly, the memory of the device is loaded into the database (more about this data in sub-paragraph 0). The modules that have reached EOL are excluded from the analysis. Thirdly, the modules are tested in live conditions and the quality of the results are analysed. The found results are logged into the database. The errors are next analysed in a Pareto chart, and the most occurring errors are discussed and if needed improved in the design.

# 5 Data understanding

## 5.1 Data sources

It is evident from the previous chapter that different departments are occupied with the handling of different types of customer feedback. This information is consecutively spread over different databases that have a different setup. In this chapter, the different databases are discussed and a relation between these databases is described.

### 5.1.1 Call Centre database

The call centre database is an immense database containing over millions of logs. The machine serial number is the unique identifier in all this data. The call centre logs the data over the phone this makes it more vulnerable for errors since it is dependent on the observations from the customer and the call centre employee. The following attributes are logged by the call centre: Date, Machine serial, Call causal, Agent number, Replenished parts.

The knowledge that is captured in this repository does not state anything about the surroundings of the problem only the solution is logged. This information is originally logged to know the usage of supplies by the customers and guard the costs customers make within the contract.

### 5.1.2 Service engineer database

The service engineer database is setup in the same matter as the call centre database. A large difference is that the service engineer has more fields to enter such as the machine copy count, travel time, working time, customer complaint etc. The service engineer database, shows the history of the machines, especially what service was provided on them. This only includes machines positioned in European countries.

### 5.1.3 Returned products data

This relatively small database contains around 1400 case of which half can be excluded because these modules originate from countries outside the EU. These cannot be analysed because the machines are not included in the other databases. After selection around 750 cases remain.

The data is the basis from the analysis it is the only dataset where the cause of the problem is included. This categorical variable is either Technical, EOL or NFF. The database further contains all the information that is extracted from the chip in the module. The database is not shared over different departments and is only available within the development department.

### 5.1.4 Production Test data

After production modules are tested, these tests are stored in a database. This is the forth database and contains measurements of the modules. Around 25 items are tested automatically. This database is not part of the SHI(F)TT fusion project. However, we include them anyway because they can only make the results stronger.

## 5.2 Data structure

One of the questions within this project is weather it is possible to combine the different databases. Besides this, the question remains if it is beneficial to combine the different databases. This chapter ends up with a simple “data warehouse” that is used for the analysis. The term data warehouse is not entirely correct because the fused dataset is static. Normally data warehouses contain data that is updated on a daily basis with information from the ERP-systems. Besides this data warehouses contain an immense dataset including all information from the company. The dataset under

consideration is only a fraction of this information. For this reason, we will not use the term ‘data warehouse’ but refer to the ‘fused database’.

### 5.2.1 UML

Before an in depth explanation on the fused database the tools that are used are shortly described. The Unified Modelling Language (UML) is a standard general-purpose modelling language in the field of software engineering. The UML was first developed in 1996 to summarise the large amounts of models present at that time. After a rough ride, the UML 2.0 has proven a useful method to develop software. UML consist of multiple tools regarding the structure and the behaviour of programs. One particular tool will prove useful when combining different databases this is the Class Diagram. A class diagram describes the different tables and the relations among those tables. Because it is a centrepiece, short explanation is given on the language.

### 5.2.2 Basics of a Class diagram

To explain the working of a class diagram a simple example is used from E-commerce. In this example, there are customers that place orders at an online store. Keep in mind that this example is highly simplified.


**Figure 5-1; UML Example**

Figure 5-1 displays the UML diagram continaing three classes. Customers contains details about the customers of the online store such as Name adress details. Using a unique ID they can be separated from each other. With this unique ID it is possilbe to relate customers to orders. The table they are stored in will look like something similar to Figure 5-2. Orders and Products are setup in a similar manner.

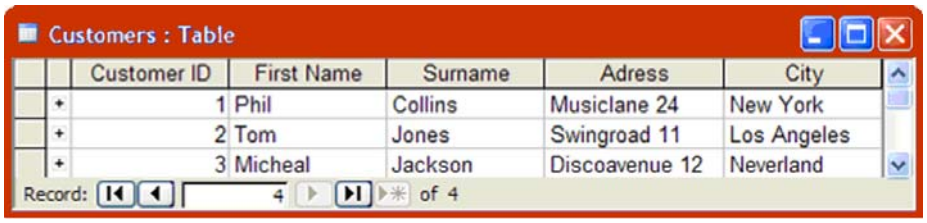


**Figure 5-2; Example of customer table**

The lines between the tables indicate that there are links between these tables. The line between Customers and Orders indicates that a customer can have multiple orders attached to him. This is illustrated in Figure 5-3 where the sub datasets of Tom Jones are expanded. He paced two orders at different times for different products. There is also the possibility to drill down the database and see what products where on these orders. These tables are called relational databases and you can jump trough the tables obtaining the information you require.

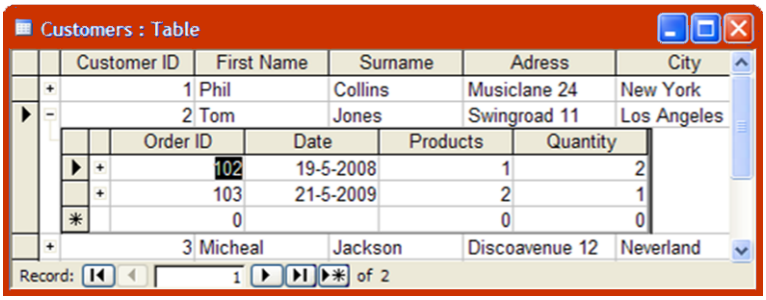


**Figure 5-3; Customer table with expanded sub dataset**

With queries, new tables can be created with the information from these tables. For example, if the manager would like to know how much of each product is sold over a certain period. A query on the orders would aggregate the information into a table where analysis can be done. There are mining tools available that are able to deal with this type of queries form live situation this is called Online Analytical Processing (OLAP) techniques.

Now that the Class diagram is explained, the next sub-paragraph gives an overview of the fused database using this exact tool.

## 5.3 Building the database

The database is setup in 'Ms Access' this simple tool has proven sufficient functionality to merge the different tables. **Error! Reference source not found.** displays the relations between each of the datasets. The 'XCRU Analysis' class is the centrepiece of the database, in this class the data transformations are stored.

The XCRU analyse class contains two types of serial numbers; the XCRU serial number and the Machine serial number. The XCRU serial number is linked to both classes of test data drawn from production. The Torque and V-high test data contain each around three Million cases. In order to speed up the analysis the non-relevant serial numbers were removed.

The machine serial numbers are linked to the Machine History class. The machine history class contains all the calls made on all machines. This is an immense dataset containing a few billion calls. Once again all non-relevant serial numbers were removed to speed up the analysis.

Besides these databases, a few classes were defined; the _Cause contains the dependant variable with three categories explaining product return; NFF, EOL, Technical failure. The _Machinetype class contains nine classes with distinctions between the machines where they operate in.

## 5.4 Feature selection

The main goal of the data mining is to understand what triggers the NFF product returns. To answer this question interesting features are defined from the dataset. In the next paragraphs, explain the selected features.

The choice of features is essential because this choice determines whether significant results can be found. The features are evidently dependant on the content of the dataset. You cannot add features if you do not have this information. With aid of the development department, service engineers and Professors at the university features were extracted from the dataset. These features can be mapped in two different categories technical and non-technical. Besides the improvement in readability, the defined categories allow comparison with the paper of Koca et al (2006). Koca et al (2006) propose two models one scientific and one practical. The practical model distinguishes between technical and non-technical failures. This problem with this separation is that there is no guarantee that the no fault found group does not contain a fault. In other words, testing methods might be unable to reveal the technical fault and a product would faulty end up in the NFF group.

The first two element of the dataset are the serial number of the module and the machine serial number. These are used as a case identifier. This information is essential to reference between the different data sources but it contains no value for the analysis.

### *5.4.1 Technical failures*

#### 5.4.1.1 Module related features

The first measure extracted is the copy count and is an exact measure of the number of copies made by the module. This count is read directly from the memory chip present in the module. This information is only available from the modules that are analysed at return. Because we are dealing with MODULE-A and MODULE-B that have a different life a extra measure is required. This measure will prove confusing for the mining software; we need a measure that is equal for both modules. By dividing the number of copies over the print stop, the life of the module regardless of the type is calculated. This measure is depicts the percentage of the total life.

Secondly, the MODULE type is logged. Previously only two different types of MODULE's were defined, however this is not entirely true, within these two types there are some smaller differences resulting in eight different types. The module type is read from the chip present in the module.

With aid of the module serial number, the test values captured during production can be recovered. During production, a number of critical voltages and the torque measures are done. These 21 measures are normally used to eliminate products that fall outside the specification limits. In light of this research, these measures are used to predict the type of failure. When these measures prove to be significant, this is a strong indication that the specification limits are not setup correctly.

When a machine is delivered at the customer, the machine contains a MODULE. With aid of the serial number of the module, the origin of the module can be revealed. It is either "delivered with the machine" or "replenished by the consumer". This distinction between modules is interesting because it reveals if the errors with regard to the life of the machine.

#### 5.4.1.2 Machine related

Using the machine serial number extracted form the module, a trace is put on the machine the module was used in. With this trace, the machine history can be recovered from the database used by the field engineers. This database contains general information about the machine such as the country and type of the machine. The modules return from countries all over Europe and come from 20 different types of machines.

Besides these basic measures, machine details close to the occurrence of module replacement can be revealed. This includes the machine copy count and the customer complaint. These are both relevant; with the machine copy count we can see if the problems occur at a certain life of the copier. The engineer logs the problem using set coding. However not many engineers obey this coding, resulting in a poor measure, however we can include the measure. It can be spotted if the NFF happens with a certain type of failure or in a certain place of the machine. To do this the code is split in two parts the place of failure within the machine where the error originates and the type of failure.

The date that the module is replenished is know; with aid of this detail, the number of calls 10,20 and 30 days surrounding the replenishment can be calculated. Within this detail, we can also calculate the service engineer visits 10, 20, 30 days around replacement. These measures are useful when we are looking replacements that are caused by complexity of the situation. According to Campbell (1988) there is a positive relation among task duration an task complexity. In other phrases: "complex tasks acquire more time than simple tasks". This is essential because we can distinguish between complex and simple problems. Keep in mind that no correction is done for the skills of the engineer and some other surrounding factors.

Other researchers have identified a trend that complex products result in higher NFF returns. Following this argument would motivate that the NFF return rates for complex problems is higher than the "NFF" return rate for simple problems. The dataset allows extraction of the number of calls

around the replacement, the average call length. With aid of these to measures, the total call duration around the replenishment is calculated. Because of the relation between task duration and task complexity duration of the visit could well prove to be a significant measure.

Another way to analyse the historical data is to calculate measures over the life of the machine such as the total number of calls, call centre/service engineer ratio and the money spend on supplies.

Machines return because contracts end or other reasons, evidently the module is returned as well. These modules contribute tot the NFF dilemma, currently no knowledge is available on what the influence of this factor is. To calculate this feature a combination of the final reporting date and the replenished date is required. When both dates occur within a short time interval the module is returned with the machine. This results in a Boolean measure called return with machine.

### *5.4.2 Non-technical related failures*

The database contains relatively little details regarding the customer requirements, or the usage of the machine. It is for future researches to look at these features. If available, an extra dataset from for example sales might be required to get this picture complete. From the available data, some measures could be extracted that explain the differences between customers.

#### 5.4.2.1 Customer related

The FDRS database contains the type of contract, this could well be a significant factor. There are a immense number of different contracts. The mayor difference in these contracts is the cost structure, some costs are included and others are excluded form the contract. In order to mine the data correctly a boolean value is created measuring if the MODULE expenses are in or excluded in the contract. A large difference between these types of contacts is expected by the development department.

Besides the contract details the number of prints/month states something about the use of the machine by the customer. The Number of prints per call might very well be an indication about the usage of the customer.

## *5.5 Summary of the features*

The dataset has become quite a large set of features. This is excellent because the bigger the number of features the larger the chance that the factor causing the NFF returns is included. Table 5-1 shows an overview of the features used for the data mining.

| General | Details | Type |
|---|---|---|
| Module serial number | | Text |
| Machine serial number | | Text |
| Reported cause | The dependant variable | 3 Categories |
| **General** | **Details** | **Type** |
| Copy count | Copy count of the module | Number |
| Module type | 8 different types of modules | 8 Categories |
| Production test values (voltages) | Contains 5 fields | Number(s) |
| Production test values (torque) | Contains 15 fields | Number(s) |
| Type of replenishment | Rental or sales | 2 Categories |
| **General** | **Details** | **Type** |
| Program | MODULE-A/MODULE-B | 2 Categories |
| Machine type | 20 different machine types | 20 Categories |
| Machine copy count | Copy count at time of replacement | Number |
| Total number of calls_repl | At time of replacement (correct for time and copy) | Number |
| Total number of calls | Over machine life (correct for time and copy) | Number |

| | | |
|---|---|---|
| Reported cause | A code entered by the engineer explaining the cause of the call | Text |
| Call centre/service engineer ratio | | Number |
| Number of supplies | The number of supplies used (correct time and copy) | Number |
| Scrap machine | Days between scrap of machine and removal of module | Number |
| Av # of prints per Module | is a small percentage of customers returning a lot of nff products | Number |
| Number of calls around replacement | Distinguish between both types | Number |
| Call centre/ service engineer ratio_around replenishment | | Number |
| **General** | **Details** | **Type** |
| Country | 13 european countries | 13 Categories |
| Type of contract | Lease or buy contract | 2 Categorical |
| Average number of prints per day | | |

**Table 5-1; interesting features**

A question often asked when starting this data mining: ”are all relevant factors included in the model?” It is difficult to answer this question because of so called “unknown unknowns” a phenomena described by Mullins(2007). A known unknown is a something that you know is not included in the analysis, you can estimate this and keep in mind that there is a unknown factor there. More tricky are the unknown unknowns (unks unks), you do not know what you do not know. There is no way to find out what the effect of these factors is because you do not know what these factors are.

The dataset has been rotated and cut multiple times to get the richest set of information that could be extracted from the original dataset. The researcher is aware that themes regarding customer satisfaction and customer requirements are missing. These factors could result in a more predictable model, and these are known unknowns. Unfortunately, this data is not available.

It is difficult to deal with the unknown unknowns, Mullins(2007) suggests long interviews with open ended questions. These interviews were conducted before merging the dataset; however, this is no guarantee that all factors are included. Because the project is data driven the worst that could happen is that these factors that are not included are the factors that cause the NFF returns. Resulting in a non-significant model. On the other hand factors could be included that have a strange relation to the dependent variable. The consequence of missing certain factors, or including the wrong factors are a non-significant model or a strange un logical model.

## 5.6 Data richness

In the literature review the paper by Lee(2002) was extensively treated, this paper set certain criteria on the richness of data. The richness of data is evaluated on four different criteria (noted vertically in Table 5-2). The databases used for the analyses are analysed on the basis of the criteria set by Lee(2002). At the analysis, data accuracy and data relevance proved most essential. Data representation and data accessibility were not as important.

With regard to data accuracy the databases the Production test database performed best, this is logical seen as it is entered by a machine. The quantitative data did merely show the measurement results, but these were of course accurate. The Service Engineer database performed worse, this was concluded due to a lot of missing fields and some inconsistencies within some cases.

With regard to data relevance the best database is the returned product database. This is of no surprise seen as the data that is in this database is collected for this purpose. The data presented in the other databases is collected for different purposes and therefore less relevant. However, the Call Centre database and the service engineer database contained some generally collected fields that can prove useful.

On the rating for representation, the data scored quite high because all data is in English. However some free text fields were sometimes entered in local languages. The production test data is only know by production, this is because this data is not really interesting for the rest of the company. The rest of the databases are accessible over the global network. Therefore, the ratings for data accessibly were medium and high.

| | Call Centre Database | Returned Product Database | Production Test data | Service Engineer Database |
|---|---|---|---|---|
| **Data accuracy** | Medium | Medium | High | Low |
| **Data relevance** | Medium | Medium | Low | Medium |
| **Representation** | Medium | High | High | Medium |
| **Data accesibility** | High | Medium | Medium | High |

**Table 5-2; Data richness of the different databases**

Overall, the researcher rates the data in the databases as medium. Nevertheless, major improvements are possible. A first improvement is in the area of data accuracy, the dataset contains a lot of missing values. Asking employees to enter data that are more complete will definitely improve the analysis. Due to the size of the dataset, these missing values could be excluded from the analysis still leaving sufficient data. However, this process might be biased because the not entered data could differ from the entered data.

A second improvement is in the field of data relevance. The data that is present in the databases is collected for another purpose than quality analysis. Adding some fields for this purpose will show more detail and improve future analysis.

# 6 Data preparation

This chapter is practically oriented en describes the analysis done in the company.

## 6.1 Merging of data

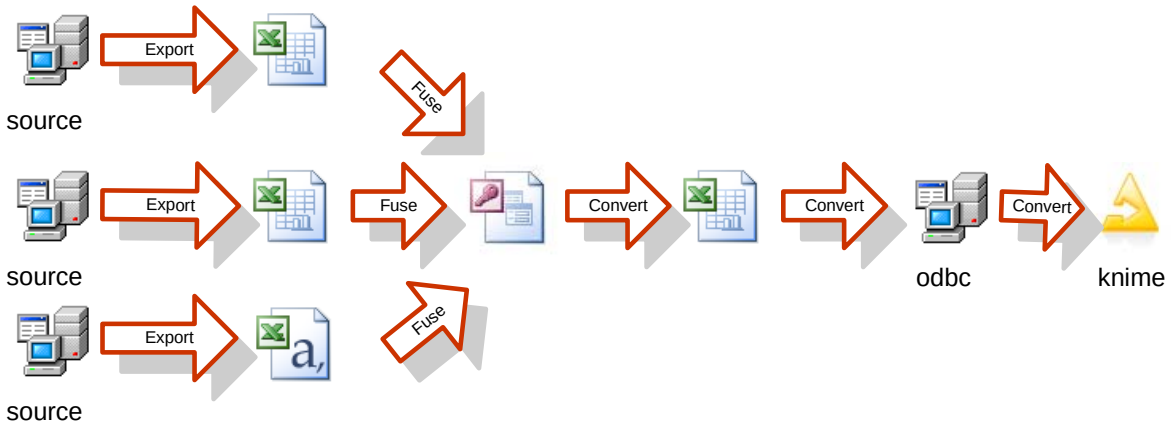


**Figure 6-1; Graphical representation of the data fusion**

The databases come from different sources transformations are required to get the dataset suited for data mining. Figure 6-1 is a representation of this transformation. The export and conversion steps are quite straight forward, interesting is the place where the databases are fused. This is done using in a ms access database.

The developed access database allows easy merging of the different data sources. Figure 6-2 displays the access menu where the datasets are managed. In the developed tool, three source databases can be selected and edited by clicking on the buttons. The 'Merge and save to excel' button merges the source databases into a new table containing the features described in Chapter 0. This table is stored in the software for analysis, and exported to a excel sheet for analysis with external programmes. It is important to note that this tool is extremely specific to the problem at hand, however modifications are possible to change the tool for different datasets.

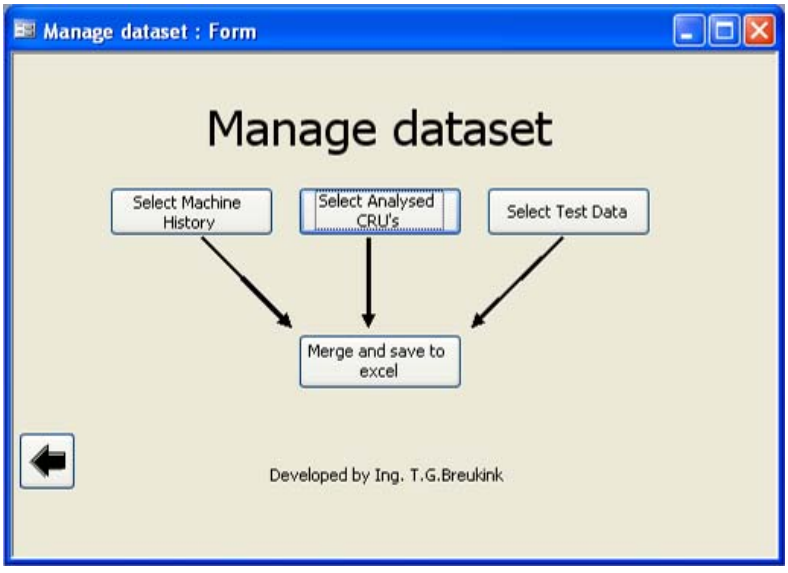


**Figure 6-2; GUI of the manage dataset menu**

## 6.2 Cleaning the data

The fusion resulted in a database containing 800 analysed modules suited for the analysis. These 800 modules are a random sample over a period of 5 years collected in 13 European countries. This dataset is not yet suited for data mining. The data requires transformation before it can be used in a data mining tool. First the outliers need to be removed, next a solution must be found for the free text values and finally the data must be set in the right format.

### 6.2.1 Removing the outliers

In statistical analysis, outliers have a large impact on the results. With a simple explanation the difference in statistical analysis and data mining is explained.

Medina, Washington has 1206 households and an average net worth of $44,253,482 per household, compared to other cities this is an extreme difference. Examining the dataset reveals that three outliers are present in the dataset: Bill Gates (46 billion) Jef Bezos (5.1 billion) and Craig McCaw (2.0 billion). Treating Bill Gates as an outlier decreases the average net worth per household to $6,115,934. Removing the top thee results in a more acceptable net worth of $224,189, this matches other cities. As illustrated, outliers in statistical analysis can easily lead to wrong analysis. The figures originate from the Forbes (2008) website.

What are the effects of outliers in data mining projects? These described type of outlier has a smaller effect in a data mining project. With mining methods such as classification, prediction, clustering and association rule learning optimal cut-off points are calculated. These cut-off points are calculated based on a learning model that attempts to get the largest difference between two groups. In our example when predicting the number of ‘Hummers’ that are sold in the Medina Washington area the learning dataset indicates that with a net worth of over $198,456 the difference in the two groups is largest. This cut-off point is far less sensitive for outliers because instead of looking at the actual values, it compares to two groups created.

Besides the traditional outliers, data mining projects suffer from an extra type of outliers that originates from the method of data collection. The method collecting the data is vulnerable for virus attacks and attempts of intrusion. This is noted by Petrovskiy(2003). The database under consideration is well protected, and no indication is found that attacks or bugs have created outliers.

Another difference between the typical statistical project and a data-mining project is the number of people that build the dataset. In a statistical project, the dataset is typically collected by the researcher or accepted sources are used. In case of a data-mining project, the data is often collected by a large number of employees. These employees have a different view on the situation and report the information in different ways. Next to this, the dataset is originally not developed for the purpose that we are going to use it for. This makes that the dataset might contain inconsistencies because of employee differences.

Analysing the dataset for outliers resulted in removing 33 cases, based on the following reasons:

- Some of the modules never entered a machine, and therefore lacked a lot of information.
- Some of the machines were used to order supplies for an entire machine park, and therefore they did not represent the overall picture.
- Some of the machines there were no replenishment calls made, indicating that these are done on another machine.

After removing these cases the final dataset contains 777 cases.

### 6.2.2 Dealing with free text

To deal with free text the information entered by the engineers requires recoding into categorical or numerical values, with some fields this proved possible. Engineer report a error code containing two elements, the place where the error occurs and the type of error. This field could be easily re coded

into two fields. A clear input structure helps dealing with free text, entirely free text fields unfortunately needed to be excluded from the analysis. These free text fields could not be recoded or interpreted into better predictors, because the notes made by the users relate to different topics. Looking at them individually does however give a picture of the situation, but automating this seems, in the researcher's opinion, an impossible task.

## 6.3 Overview of the data

The complete overview of the data is available in the confidential version of the master thesis. There are 750 cases analysed.

De data collection occurred over a large period, the first sample was taken in March 2002 and the last sample is taken in January 2009. The large time span allows comparing the NFF rates over time.

# 7 Analyses and result

This chapter is the result of the fourth phase of the CRISP-DM reference model, as described in Chapter 3.1. Throughout this chapter, several mining techniques are applied in order to get one solid model.

## 7.1 Modelling technique

In the literature review, 15 tools were considered for mining a dataset. After removing the tools that did not prove sufficient functionality, only five remained. After installing these, examining the user interface and a discussion with experts the KNIME tool was selected as the most appropriate one. This tool is to a based on the algorithms present in the WEKA system, but the User Interface is more attractive and easier to understand.

Of course, the ability to properly mine models is dependent on the algorithms and functionality offered by the mining tool. The WEKA system seems to be the most extensive and reliable. However, the WEKA system is not easy to work with because it is mainly code driven. The KNIME tool has a user-friendly interface and is easier to work with. This makes it the best tool to use in light of this project.

## 7.2 Modelling method

There are multiple methods to mine the database. Fayyad et al (1996) define three different mining methods Classification, clustering and estimation. In this research, we would like to predict the dependent variable. The cause of return is a categorical variable that can have one of three values; Technical, EOL or NFF. Because of the predictable nature of the question rule induction is chosen. Rule induction builds a decision three. These rules result from the data and predict the dependent variable. You could also get a similar result using regression analysis however; regression analysis does not work that well with categorical variables.

Another advantage of a decision three over other techniques, such as neural networks is that it clearly displays the causes. Besides just displaying the causes you can see what the influence is and how reliable. In other words, the resulting model is easy to understand.

Experts in the field also state that the more complex methods not always result in better models. If there is a relation present, the simpler techniques will most likely also display this relation. This is a motivation to choose for less complex method of decision threes.

## 7.3 Model layout

Partitioning data into training and test sets is an import part of evaluating the model. With aid of these partitions, the proposed model can be tested and the accuracy of the model is confirmed. This will reduce the risk of over fitting the data. Minimising the difference between the learning set and the predicted set reduces the risk of over fitting. Over fitting occurs when the model is accurately describing the noise of the dataset instead of the underlying model. Testing the model with the remainder of the dataset and comparing both these model accuracies will reveal if this is the case. A large difference between the accuracy of the learner set with the testing set indicates that the model might over fit the data.

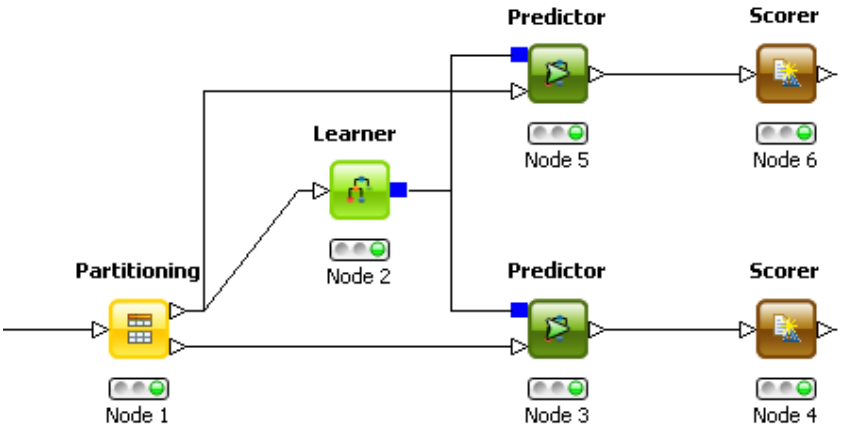


**Figure 7-1; Split training and test data**

The process of training and learning is set in the software like the screenshot in Figure 7-1. The separation between training and test dataset is important. Typically, most data is used for training and a partition is used for testing. Several methods exist to make this separation two are discussed. The fist approach is to randomly split the set in two parts often the separation 90/10 or 80/20 is used. A second method is to extract only 1 case, use the rest of the cases for the model and then test the model with this one case. This is done multiple times, normally the number of cases that the dataset contains. The second approach is effective with smaller datasets. Because of the large number of features present in the dataset and the number of cases, the first approach is used to make a separation in the data. In order to do so a loop needs to be inserted around this model.

## 7.4 Setting the parameters

The model allows some parameters to be set. First off, we set everything to the standard values, and start changing one essential value. This value is the “minimal number of records per node” this is the minimal number of records required that support a leaf. Setting this value to one will result in a extremely accurate model on the learning data because it will simply create a node for every case in the model. However, the test data will have a low accuracy because the model is over fitted to the training data. When making this value extremely large, say 300 it will have little nodes and the model will be too generic and not extract the subtle differences in the data.

Looking at the graph ‘minimal number of records per node’ versus the ‘test data accuracy’, we expect a graph like the one shown on the left hand side in Figure 7-2. At a low number of records per node, the accuracy of the training data is high and the test data is low. Increasing the number of records per node will decrease the training data accuracy but increase the test data accuracy. This indicates that the over fit is being removed. After the optimal point both models perform equally and there the train data is not over fitted.

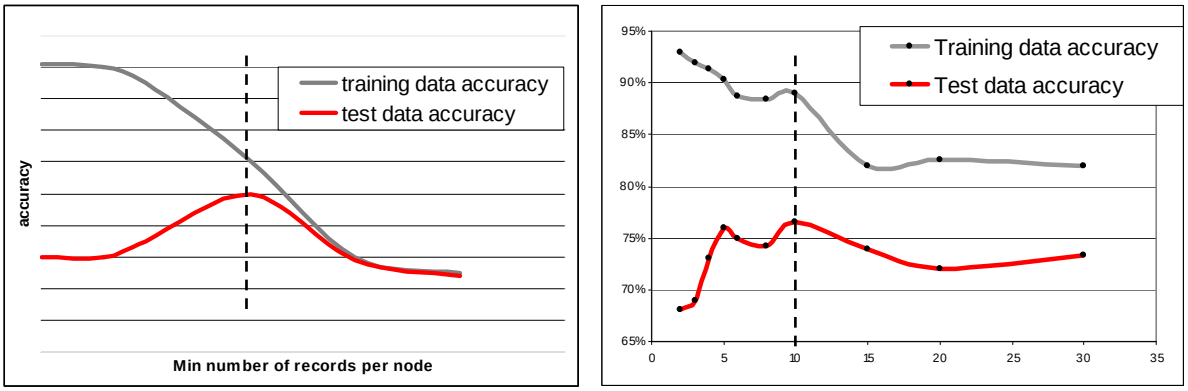


**Figure 7-2; Training data accuracy vs. Test data accuracy**

When looking at the graph build with the real data in Figure 7-2 it is surprising to see that what we expected to happen really is true. Beyond five records per node, the test- and train-data perform similarly. Therefore, we can state that we require a min number of records per node of at least five. The difference of 10-15% between the training en the test data seems an acceptable value. We might choose to go beyond these minimum of five records per node because this will result in less rules and a better understandable model.

## 7.5 Evaluating the model

In the data-mining phase, the 'decision tree learner' came up with the model shown in **Error! Reference source not found.**. The overall model is accurate in 80.0% of the cases. In other words, 155 faults occurred when classifying 776 cases, this corresponds to a sigma of 2.4.

The 'Percentage of life' is the most dominant predictor in the dataset. This is of no surprise because this factor is used in the test sequence. The testing sequence is designed in such a way that no in depth analysis are done on modules over 75% of their life. By analysing the almost EOL modules little is gained, and therefore the testing is setup in such a way. That the data-analysis proves this as the first best predictor is of little added value besides proving that the methodology works. Yeeh!

### 7.5.1 Calls around replacement

The second most dominant factor the "# of calls 10 days around replacement" shows that two or more calls around a replacement trigger a NFF return. Before depth is provided on this, the factor must be further explained; the factor contains two types of calls, namely calls over the phone and engineering visits. The factors containing either engineering or phone visits did not prove as influential as the combination of both.

From the total 145 NFF cases in the model, 50 cases are classified as such. This factor is able to explain 34% of the causes of the NFF cases. This is quite a large number! There are two other factors not noted by the mining algorithm because they are not implemented in the dataset. They are established on a more theoretical basis.

### 7.5.2 Machine returns

Besides the data mined results, another result has arisen when examining the return process. There is a theoretical error in the return of MODULE-A and MODULE-B. The machine is set at the customer at a certain contract, at the end of this contract the machine including supplies and modules is returned to the manufacturer. The returned machine contains a module that is most likely not at it end of life. This by itself is a useful conclusion; unfortunately, the dataset under consideration does not allow looking into this specific problem.

The only option left is a mathematically estimation of these effects. First, we can safely assume that ¾$^{th}$ of the modules from this return stream returns as a NFF, because the life is uniform between 0 and 1 and above .75 the modules are considered as EOL modules. Throughout the life of the machine on average eight modules are consumed. Assuming that the last module has 75% change of returning with the machine, the process will explain 9.4% of 19% NFF returns.

$$\frac{\text{Chance of no NFF return at machine EOL}}{\text{Number of modules consumed}} = \text{Chance of NFF return due to machine return}$$

**Equation 7-1; Calculating Theoretical NFF returns**

Because the return of machines is generally done by an engineer that only reports one call. This cause cannot overlap with the cause mentioned by the number of calls around replacement. This 9,4% is a

part of the total number of returned modules that returns NFF. In other words from the 776 cases that returned 776*9.4%=70 modules should theoretically return NFF. This means that from the 145 cases that did return NFF, 70 cases are explained. From the 145 NFF returns, 48% is due to a wrongly established flow in the process.

### 7.5.3 Contract differences

Within the company a large number of different contracts exist, these contracts are split in two groups. In the first group, the consumables are included this means that the consumer does not pay separately for the replenished parts. In the second group, the consumables are excluded from the contract, and the consumer pays a fee when replacing the module. Because the second group is relatively small the effects of this measure are limited, and not noted by the 'decision tree miner'. However, differences between these groups are interesting, because they illustrate something regarding customer behaviour.

t is clearly visible that customers who pay for consumables have a lower NFF- and Technical- return rate. The EOL return rate is higher with these customers. It is safe to state that customers who pay for supplies are more careful with replacing them. They tend to be economic and less likely to replace the modules. Due to the Boolean nature of the variables and the size of the sample, we are not able to calculate confidence intervals or state something about the statistical significance of the effect. The effect however is of no surprise and this is what we expected to see with the analysis.

## 7.6 Causes underlying the rules found

From the analysis thee causes were clearly found, this paragraph comes with possible explanations of why these relations are present.

### 7.6.1 The number of calls around replenishment

When developing the features, the purpose of this particular feature was to predict if modules were replaced incorrectly, let us elaborate on this more. When the customer calls to the call centre it might be wrongly advised to replace the MODULE. If this is the case, the problem is not solved and another call will follow in the near future. We found that in 50/155= 34% of the NFF returns there were more than 1 calls around the replacement of a MODULE. These indicate that the replacement of the MODULE was a incorrect action.

The subsequent question among my readers should be: "Why are these modules replaced without proper investigation?" The answer to this question is two fold. The first reason is on a financial basis, sending a module that can be replaced by the customer is about a factor 2 cheaper than sending a engineer to replace the module. This cost benefit analysis has been calculated and the current tradeoffs resulted in the best ROI. The second reason for the customer replaceable units is because the machines need to look well-built. When a engineer visits this image is reduced, when a consumable is replaced this does not change this image. The machine just required some supplies, this is logical and not seen as a defect of the machine.

Do these both reasons still hold if we incorporate the new knowledge gained by data mining on the NFF returns? This question, together with two other business cases is answered in Paragraph 7.7 and the management decisions made years ago are tested with this new knowledge. It proves to be a interesting dilemma!

### 7.6.2 The effect of machine return

This is the most influential factor is the effect of return machines unfortunately this factor proves to be least useful for the company, this contradiction is unfortunate. The high value of NFF rates is cuased by the following flow: machines that reach their end of life are returned to the company, some of them are refurbished and sold as second hand machines others are ready to end up on the scrapheap. Parts are separated and they enter the return stream. From this return stream, the modules are sampled and

analysed. Apparently a large part of the NFF sampled modules originates from these returned of machines.

### 7.6.3 The contract type

Customers with a contract including supplies had a lower NFF rate as customers paying for their consumables. This is an indication that the type of contract influences the customer's behaviour. It could also be that the behaviour is initially different and customer chooses a contract that fits their preferences.

## 7.7 Follow-up projects

Interesting is that the rules found in the analysis leave room for improvement. The new knowledge can be used to evaluate and improve business rules this is done in three business cases. Normally a business case captures the reasoning for initiating a project or task but it can of course also be used to evaluate the current way of working.

### 7.7.1 Business case 1

This business case evaluates a prior decision made by management. Management decided that (seen as sending a customer replicable unit is cheaper than sending a engineer) the welcome centre should try to send a module instead of sending a engineer. In this business case, we distinguish two different cases. The current way of working (A) and another situation (B) where an engineer is sent when the welcome centre is in doubt.

The conclusions from this business case were that the current way of working is the correct way of working. The analysis is excluded from this document for confidentiality reasons.

### 7.7.2 Business case 2

A similar business case can be conducted with regard to the contracts. Contracts that do not include supplies have a 16% lower NFF return rate than contracts that do include supplies. However, the company encourages the second type of contracts; consecutively these contracts are more common.

The researcher is not aware of the margins that are made on these contracts. Using this information in combination with the newly gained knowledge an optimum can most likely be found. This enables the company to reanalyse the current way of working. Because the researcher does not have sufficient information and knowledge about these contracts to do these analysis. This is left as a future project for managers at the corporation.

### 7.7.3 Business case 3

The business case regarding the return machines is quite simplistic, but has large potential. Reusing modules that return from the field will generate profit right away. The benefit for the company is substantial.

These modules require extensive testing and after cleaning they can be repacked and resend to the consumer. The process will look according to Figure 7-3. Preliminary analysis on the cost found that this would prove a profitable way of working. A more rigid business case is required to calculate the exact profit and some testing is required to see if the process proves rigid enough. This is a nice spinoff of this project.

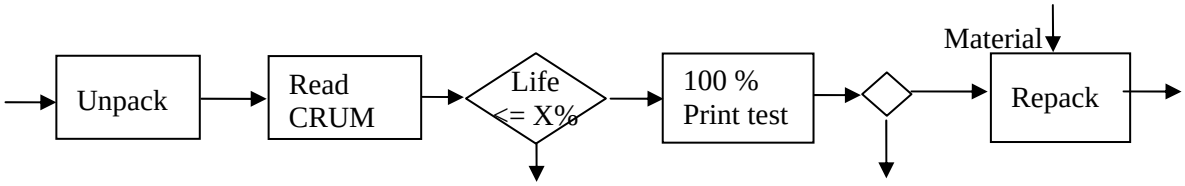


**Figure 7-3; Proposed process of re-use**

The idea described above is currently judged by the company. When it proves financially attractive and there are no technical hiccups it might be implemented in the near future. This also proves interesting for other product lines.

# 8 Conclusions and recommendations

The probably best-read chapter of this document is separated into three pieces. First, the results form the analysis are accumulated and summarised. Secondly, three business cases are presented that re-evaluate decisions made by the company using the newly gained knowledge from the data mining. Thirdly, light is shed on the applicability of data mining in the field of quality and reliability, and the possible future applications for the tools.

## 8.1 Results from the analysis

The company holds mainly data from three sources the machines, the modules and the service engineers. These databases are setup in such a way that although they are enormous they can be connected to each other using module and machine serial numbers. Broadening the scope might ask to also use information sources from other departments, these might not be ordered according to serial numbers. For example the CRM management system most likely orders customers on their customer number. These links are more difficult to establish, however there was no need to do so.

Examining the data on the data accuracy, relevance, representation and accessibility as proposed by Lee (2002) shows that the databases are not extremely rich in their content, however they prove sufficient for the purpose proposed, later in the conclusions possible additions of the dataset are discussed. It is interesting that incomplete data, and data collected for another purpose can still prove valuable for these analysis. Now it is time to look at those results.

### 8.1.1 Initial findings

Aggregating the available data on the returned products showed interesting insight. Quite early in the project the different return rates of both modules could be explained.

The differences between the both products are explained by the larger life of the MODULE-B module. The EOL is carefully designed with the module at hand; there is no confusion on when the product reaches it EOL. The module logs the number of times the product fired and after a certain number of fires the module needs to be replenished. MODULE-B has twice the life of MODULE-A, this forms no problem because of some design improvements. Initially the cause of higher NFF rates was thought to be in the different design. However, when we look at how the modules return this cannot possibly be the case.

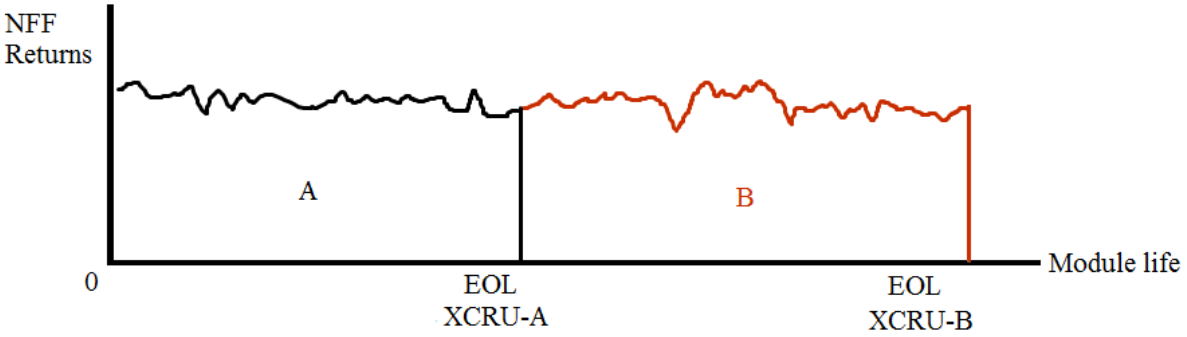


**Figure 8-1; Sketch of NFF returns over module life**

When looking at the return of MODULE-A (with the shorter life) it is clear that returns happen random over the life of the module. A sketch of these NFF rates is provided in Figure 8-1, horizontally the life of the modules is displayed and vertically the number of returns occurring at a particular time of the modules life. Consecutively area "A" depicts the total returns of MODULE-A. The total returns of MODULE-B are of course the sum of area "A" and "B". Seen as the life of MODULE-B is twice the life of MODULE-A this explains the factor 2 higher return rates. Apparently, there is something else triggering these NFF returns. The next paragraph will shed more light on this.

Keep in mind that the company operates with lease machines. In this market, it is normal that all supplies return. This means that besides the Technical also EOL modules return. The NFF return percentages are calculated slightly different from the literature conducted on consumer products. The NFF rates are calculated dividing the NFF returns over the total returns. In the paper by Brombacher (2005) these total returns include NFF returns and Technical returns. Therefore, the NFF rates found here are lower, but to the researcher's opinion more accurate.

### *8.1.2 Causes of NFF*

The previous paragraph explained that there is little difference in both modules, and it is well possible that they both suffer from the same root causes that trigger the product return. Therefore, in this chapter both modules are treated as equals. Comparisons between both modules did not show large differences.

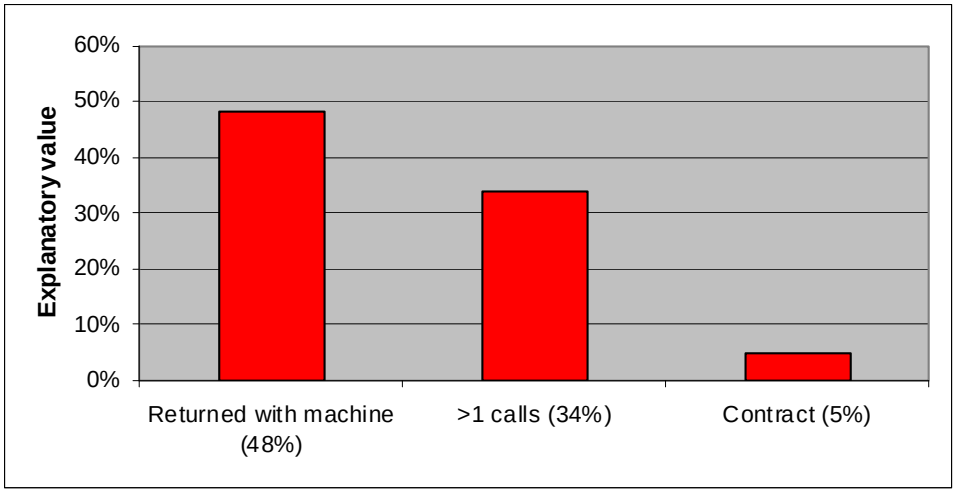


**Figure 8-2; Cause of return**

Figure 8-2 is without a doubt the most important figure in this document. It reveals the mysteries surrounding the NFF returns.

Let us first consider the numbers of calls around the replenishment. The purpose of this particular feature was to predict if modules were replaced unnecessary. When the customer calls the welcome centre, it might be advised to replace the MODULE because of wrong judgement. If this is the case, the problem is not solved and another call will follow in the near future. We found that in 50/155= 34% of the NFF returns there were more than one call around the replacement of a MODULE. These indicate that the replacement of the MODULE was an incorrect action.

The relation between the welcome centre and NFF returns has been described by Magniez C. (2007). He stated; “it can be assumed that CC will tend to replace modules whenever it is not able to solve the problem, resulting in high NFF returns.” The main reason for this badly diagnosed replacement is on financial basis, sending a module that can be replaced by the customer is about a factor 2 cheaper than sending an engineer to replace the module. This cost benefit analysis has been calculated and the current tradeoffs resulted in the highest profit. Does this reason still hold if we incorporate the new knowledge gained by data mining on the NFF returns? For an answer of this question, please look at the follow-up projects in Paragraph 7.7.

The second cause incorporates the effects of machine return to the corporation. This is the most influential factor; unfortunately, the effect of return machines proves to be least useful for the company. The high value of NFF rates is caused by the following flow: machines that reach their end of life are returned to the company; some of them are refurbished and sold as second hand machines others are ready to end up on the scrapheap. Parts are separated and they enter the return stream. From this return stream, the modules are sampled and analysed. Apparently, a large part of the NFF sampled modules originates from these returned machines.

Thirdly, the contract type proved a small effect in explaining the NFF. However, the differences between contracts are relatively large, due to the small proportion of contracts that do not include supplies the effect of this factor is limited. Contracts that do not include supplies return 16% less NFF modules as contracts that do include supplies. It is quite straightforward that customers easier change parts if this is free. This is an indication that the type of contract influences the customer’s behaviour. It could also be that the behaviour is initially different and customer chooses a contract that fits their preferences.

## 8.2 Discussion and implications

This final paragraph highlights the results on three different levels namely: Data mining, data fusion and company specific.

### 8.2.1 Reflection on data mining

The researcher is quite surprised about the potential of data mining. The large amount of data can, relatively easy, be used for analysis purposes. It requires some endeavour to build the fused database but after this the analysis are relatively easy.

Because of the exploratory nature and the scope of the data, the results were outside the scope of the department where the research was conducted. The results have impact on different parts of the organization. Changes are required in different departments than where the research was conducted. Difficulties with this are that other departments might be less participating in the project and less likely to change.

With regard to the content of the data ware house the researcher found that some measures proved more results than others. Besides this some measures were not present and would add value to the research. Customer details and customer complaint are measures that would largely improve the model. Production test data is an example of data that did not prove useful for this research. It is of the researcher’s opinion that the more general measures (collected for no specific purpose) are more useful then measures collected for a specific purpose. However, this is not supported in the literature and is merely an observation.

Important to note is that this fused database is developed with a certain goal. The data is setup in such a way that certain hypothesis can be tested. Therefore a clear goal is defined before starting the build. For this reason the research doubts weather the tool is able to detect problems by itself. However, it is proves possible to test hypothesis that managers have and draw funded conclusions on this. In order to facilitate this, the data warehouse should be setup dynamically and allow changes in the data structure. This is a known trade-off with software engineering on one hand software provides structure. Due to this structure, the software functions effectively but lacks the flexibility to be suited for every application. The right tradeoffs need to be considered for the system to function effectively.

### 8.2.2 Reflection on Data Fusion project

As indicated in Chapter **Error! Reference source not found.**, this project fits a larger data fusion project. This chapter aggregates the most important findings for this larger project. As seen by the conclusions in the previous paragraph it proves to be useful to merge data from different sources. The Data Fusion project will be fond of this result, it indicates the potential in the merging in different data sources. Throughout the project, the researcher became aware of possible problems in the future development of such a system;

First, the project is based on the use of data sources to determine the origin of NFF failures. This information is captured directly at the customers and a lot of information is lost interpreting and digitalizing the information. Why not focus on adequate systems to deal with the consumer complaint and use their statements as the main source of information.

A second problem arises due to the wide applicability of such a system. The system should be able to work with a wide variety of data sets and able to fit different goals. It seems a very daunting task to create such a system, almost impossible. However, we should not underestimate the possibilities of information technology. The progress that has been made with CRM and ERP systems is enormous. Some time ago the wide applicability of these systems and the different goals that they severe would not have been though of.

Throughout the research the researcher experienced that data collected for a specific purpose was not useful, the more generally collected data did proved a better use. For example, the test data collected particularly for rejection of modules did prove of little use. Information regarding contracts type and usage proved more useful. This might be an interesting research objective for future studies focussing on what data to use in a data fusion projects.

The research by D.Norman(1993) notes that Computers are in particular able to deal with repetitive work while humans are not so good at repetitive work. Humans however are good in creative work, computer do lack the ability in this area. The system should be build in such a way that it is capable of collecting the data and presenting it in such a way that the more creative human is able to use and read this data. Keep in mind that the computer is not able to interpret this data. In the field of Artificial Intelligence (AI) attempts are made to give computer this creative capably, a particular part of AI is occupied with data mining and prediction.

### 8.2.3 Implications for the Company

Prior to this research, there were merely hypothesis from employees on the causes of the NFF returns. Thanks to this project, these hypotheses were turned into facts. This helps the company in several ways; first, as shown by the business cases in paragraph **Error! Reference source not found.** the new knowledge can be used to evaluate the current way of working. Second, the uncertainty regarding product quality is to a large extend explained, managers know the main underling causes. Thirdly, pointers are created to monitor and intervene when the NFF rates rises to a critical level.

Keep in mind that the company initially cooperated in this project to improve the design of the modules, and see where its weakness lies. The data-mining project however showed that the main problems occur in different departments of the company. The processes in place trigger the high

amount of NFF returns. Due to the size of the organization, it is difficult to make changes in another department.

Finally, improvements are possible with the collection of data. The test data proves little value for quality and reliability analysis, indicating that the specification limits are correctly defined in the light of NFF returns. However, other data might be included to make the dataset richer. This includes first customer details such as size, industry etc. second the return path and thirdly the return reason. Adding these measures allows analysis at another level. The customer details allow better comparisons in the area of customer satisfaction. The research at hand proved contract type as a significant factor, other customer details might prove significant and eventually lead to product redesigns. The return reason allows for a faster response to trends occurring in the market.

# Abbreviations

| | |
|---|---|
| AI | Artificial Intelligence |
| BPS | Business Problem Solving |
| CD&MG | Consumables development and manufacturing group |
| CRU | Customer Replaceable Unit |
| DMAIC | Define, Measure, Analyse, Improve and Control |
| CRM | Customer Relations Management |
| EOL | End of life |
| ERP | Enterprise Resource Planning |
| ESC | global Equipment Supply Chain |
| GSSC | Global Service Supply Chain |
| IS | Information Systems |
| IT | Information Technology |
| KDD | Knowledge Discovery in Databases |
| NFF | No Fault Found |
| SHI(F)TT | Service, Helpdesk/Callcenter, Internet, (Forum), Trade and Test |
| TOC | Table Of Content |
| UML | Unified Modelling Language |
| MODULE | Xerographic Customer Replaceable Unit |
| MODULE-A | Xerographic Customer Replaceable Unit product-type A |
| MODULE-B | Xerographic Customer Replaceable Unit product-type B |

# References


van Aken et al.(2007)

van Aken, J.E., Berends, H., van der Bij, H., “Problem Solving in Organizations”. Cambridge: Cambridge University Press 2007.

Brombacher et al. (2005)

Brombacher A.C., Sander P.C., Sonnemans P.J.M., Rouvroye J.L. “Managing product reliability in business processes ‘under pressure’”. Reliability Engineering and System Safety 2005; 88; p. 137-146

Campbell J.D. (1988)

Donald J. Campbell, “Task Complexity: A Review and Analysis”, The Academy of Management Review, Vol. 13, No. 1 (Jan., 1988), pp. 40-52.

Cronin J. (2000)

J.Josehph Cronin Jr., Michael K. Brady and Tomas M. Hult, ”Assessing the effects of quality, value and customer satisfaction on consumer behavioural intentions in service enviroments.” Journal of retailing, Vol 76, Issue 2, Summer 2000, P 193-218

DenOuden (2005)

Ouden, P.H. den, “Development of a design analysis model for consumer complaints revealing a new class of quality failures”. Technische Universiteit, Eindhoven. Eindhoven: University Printing Office.

Fayyad et al (1996)

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., “From Data Mining to Knowledge Discovery in Databases”. American Association for Artificail Intelligence fall 1996.

Geudens et al (2005)

W.H.J. Geuden, P.J.M. Sonnemans, V.T. Petkova, A.C. Brombacher, “soft Reiliabiliyt, a New Clas of Problems for Innovative Products: “how to approach them”; IEEE Annal reliability and Maintainability Symposium 2005, pp 374-378, Alexandraia, VA USA, 2005.

Hislop (2005)

Hislop, D,. Knowledge Management in Organizations: A Critical Introduction. Oxford: Oxford University Press, chapter 2, 2005.

Magniez, C (2007)

Magniez, C., “Combining information flow and physics-of-failure in mechatronic products”. Technische Universiteit, Eindhoven. Eindhoven: University Printing Office.

Norman. D(1993)

Donald Noman, “Things that make us smart: defending human attributes”, Perseus books (1993). ISBN:0-201-62695-0

Pitt at al. (2002)

Pitt, F.L., Berthon, R.P., Watson, T.R., Zinkhan, G.M., “The Internet and the birth of real consumer power”, Business Horizons Volume 45 Issue 4 August 2002, pages 6-14.

Reighardt J (2008)

Reighardt Jörg, “Hypothesis- v.s. Data Driven Research”, University of Wurzburg august 20th 2008. http://videolectures.net/cvss08_reichardt_ddhd/

Sander and Brombacher (1999)

Sander, P.C. & Brombacher, A.C. 1999. MIR: the use of reliability information flows as a maturity index for quality management. Quality and reliability engineering international 15: 439-447.

Kano (1984)

Kano, N, Attractive quality and must-be quality", The Journal of the Japanese Society for Quality Control, April 1984, pp. 39-48.

Lu Y (2007)

Lu Y., "Merging of incoherent field feedback data into prioritized design information (S.H.I.T.T. fusion)". Project proposal at TU/e

Petrovskiy (2003)

Petrovskiy, M.I.,"Outlier Detection Algorithms in Data Mining Systems", Programming and computer software, Vol. 29, No4, 2003, pp. 228-237.

Polanyi (1966)

Polanyi, M., The tacit Dimension, Published by Butterworth-Heinemann pag 135, 1966.

WDS global (2006)

White paper by WDSglobal.com to download see :www.wdsglobal.com/news/whitepapers/ 20060717/MediaBulletinNFF.pdf

Zia, O (1995)

Zia, O, "Continuing engineering education, the answer to the ever increasing pas of technology renewal and global economy". IEEE Annual Reliabliyt and Maintanability symposium 1995, pp2c2 14 vol 1