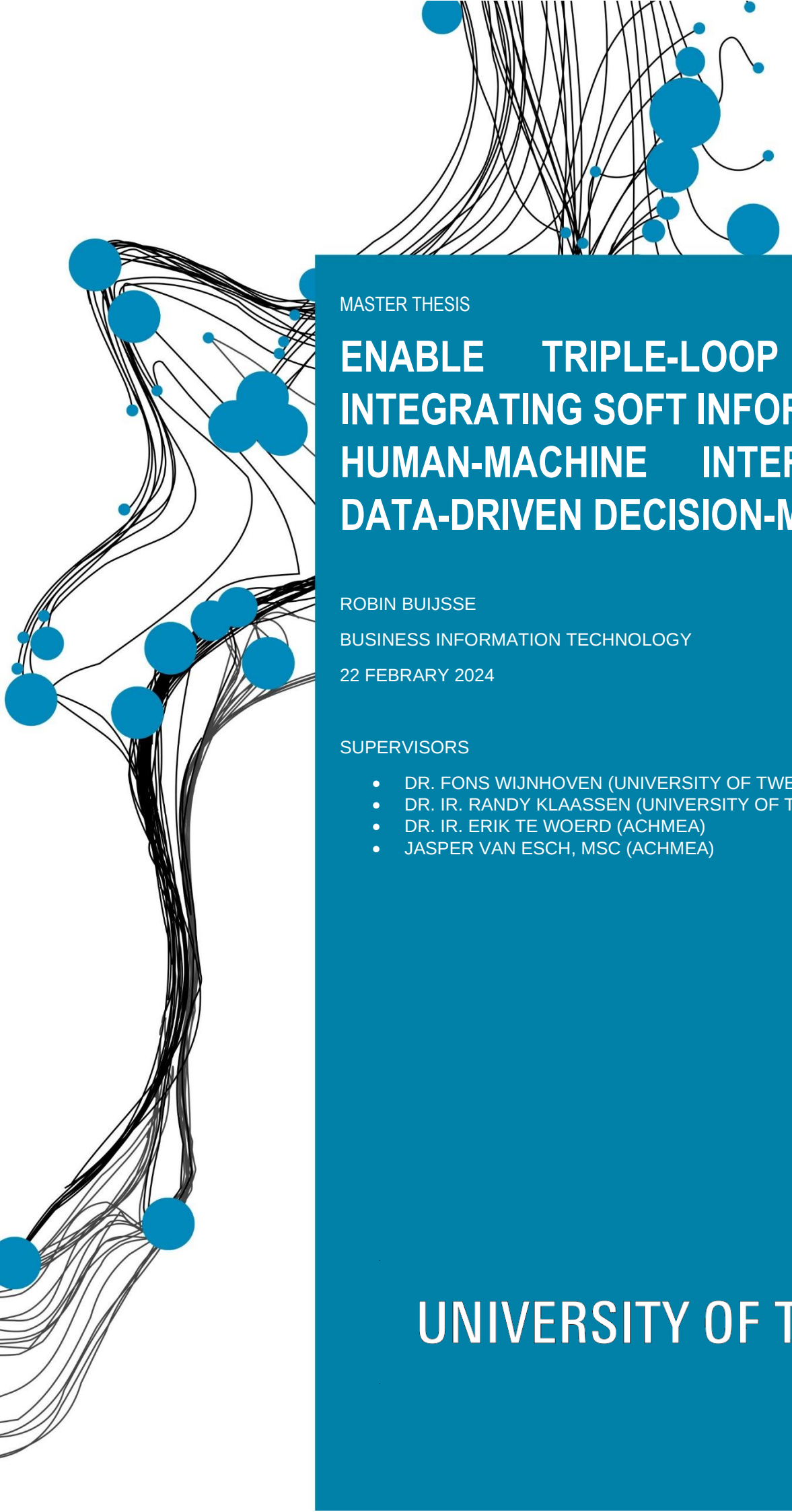

MASTER THESIS

ENABLE TRIPLE-LOOP LEARNING: INTEGRATING SOFT INFORMATION AND HUMAN-MACHINE INTERACTION IN DATA-DRIVEN DECISION-MAKING

ROBIN BUIJSSE

BUSINESS INFORMATION TECHNOLOGY

22 FEBRARY 2024

SUPERVISORS

- DR. FONS WIJNHOVEN (UNIVERSITY OF TWENTE)
- DR. IR. RANDY KLAASSEN (UNIVERSITY OF TWENTE)
- DR. IR. ERIK TE WOERD (ACHMEA)
- JASPER VAN ESCH, MSC (ACHMEA)

UNIVERSITY OF TWENTE.

ABSTRACT

This research presents a methodology that integrates an organizational learning approach for developing data-driven decision-making (DDDM) to successfully integrate mutual human and machine learning (triple-loop learning). Validated in a non-life insurance case study, the methodology addresses the ineffectiveness of an existing DDDM tool due to the lack of human-machine interaction. The existing DDDM tool is further developed by triple-loop learning. The case study demonstrates how triple-loop learning enables the DDDM tool, consisting of predictive models, to implement human decision norms and values, and enables people to gain new insights from working with the DDDM tool. The research contributes to design science theory by offering a methodology and guidelines to enable triple-loop learning in the development of a predictive model for DDDM in general. Within DDDM, triple-loop learning should lead to the alignment of human and machine mental models so that decisions made by DDDM align with human norms and goals.

Keywords: triple-loop learning, organizational learning, human-machine interaction, data-driven decision-making, human-in-the-loop, predictive modeling

EXECUTIVE SUMMARY

The goal of this research is to **add a prediction of the usefulness of re-inspections to the prioritization method**.

Achmea aims to transform into a **data-driven insurer**, and this research signifies a significant stride toward achieving that goal.

Added value

Add **soft information** to the re-inspection prioritization method to give the right context and to make a distinction between impactful and non-impactful re-inspections.

Create a **feedback** effect to induct autonomous learning.

Increase the **acceptance** of the prioritization method among risk experts.

The implementation of the usefulness predictive model establishes a feedback loop, where:

- the prioritization method prioritizes buildings for re-inspection;
- risk experts provide feedback by assigning a usefulness score to the re-inspections;
- the usefulness predictive model, integrated into the prioritization method, autonomously learns itself with the usefulness score.

The feedback loop leads to a prioritization method that is self-learning and self-improving over time, ultimately resulting in re-inspections that progressively become more useful and, consequently, more impactful.

Results

The most effective algorithm in this study predicts usefulness by utilizing **policy data** and making **predictions for the entire portfolio**.

A simulation compared this usefulness prediction-included prioritization method with the current prioritization method, revealing a **35% improvement in the average usefulness score** for the prioritized re-inspections for SMEs but no significant improvement for large enterprises.

Recommendations

The recommendation is to **implement the re-inspection usefulness predictive model** for its potential to give context to re-inspections and to introduce autonomous learning within the prioritization method. The re-inspection usefulness predictive model results in more impactful assessments and improved acceptance of the prioritization method.

☆☆☆ Limitations in the **data quality** result in restrained performance of the usefulness predictive model. Improvements through the SKB+ project and the linkage between policy and inspection data are necessary.

The usefulness of the re-inspections exhibits an **uneven distribution**. Consequently, the usefulness predictive model struggles to accurately assess the extreme usefulness categories. To mitigate, it is crucial to establish a **tuned definition** of the usefulness of re-inspections and to **communicate the importance** of accurate assessment of the usefulness to risk experts.

Due to the lack of transparency and interpretability in the prioritization method, decision-makers face challenges in making adjustments in the prioritization, and it remains unclear to risk experts why certain companies are prioritized for re-inspection. Consequently, **improvements in transparency and interpretability** within the prioritization method are deemed necessary.

The usefulness predictive model, trained on data from past re-inspections, fails to **consider re-inspections** in sectors **that haven't been conducted**, potentially leading to an oversight of their actual usefulness in the assessment.

ACKNOWLEDGMENTS

I would like to express my gratitude to Achmea for providing me with the opportunity to conduct this research. I am grateful for the freedom they afforded me to develop this research independently. I extend my gratitude to my daily supervisors, Erik te Woerd and Jasper van Esch, for their support, invaluable suggestions, and constant encouragement throughout my graduation at Achmea. Their guidance has been indispensable, and I am grateful for the pleasant collaboration. I am also very grateful to the various stakeholders who gave their time and participated with dedication in the interviews and evaluations for this study. Without their commitment and cooperation, this study would not have been possible. I am indebted to Fons Wijnhoven and Randy Klaassen for their academic guidance, insightful comments, and constructive criticisms, which have significantly enhanced the academic quality of this research. I would also like to express my gratitude to all those who, directly or indirectly, contributed to the completion of this work. Your support and encouragement have been invaluable. To you, the reader, I extend my thanks. I hope that the knowledge presented herein proves valuable to you, and I appreciate your time invested in reading this work.

TABLE OF CONTENTS

Abstract 2

Executive summary 3

Acknowledgments 4

1 Introduction 7

1.1 Problem analysis 9

1.2 Problem statement 9

1.3 Objective 10

1.4 Outline of the paper 10

2 Theoretical framework 11

2.1 Data-driven decision-making from an organizational learning approach 11

2.2 Predictive modeling 13

2.3 Predictive model development as an organizational learning process 16

2.4 Conclusion theoretical framework 17

3 Research design 18

3.1 Research questions 18

3.2 Human involvement 20

3.3 Machine learning 21

4 Externalization and combination of tacit knowledge 24

4.1 Stakeholder perspective on the prioritization method 24

4.2 Usefulness of a re-inspection 27

5 Machine learning 30

5.1 Data preprocessing 30

5.2 Model development results 34

5.3 Simulation of usefulness prediction on prioritization method 39

5.4 Conclusion machine learning 41

6 Combination and internalization of explicit knowledge 42

7 Discussion 46

7.1 Main findings 46

7.2 Practical implications 47

7.3 Theoretical implications 47

7.4 Limitations and future work 48

8 Conclusion 50

References 51

Appendix A: Case background 54

Appendix B: Literature search methods 57
Appendix C: Predictive modelling algorithms 60
Appendix D: Interview questions 62
Appendix E: Data understanding and preparation 63
E.1 Enterprise systems 63
E.2 Initial databases 66
E.3 Data connection between databases 69
E.4 Data preparation 74
Appendix F: Model development results 76
Appendix G: Interpretability scenarios 77

1 INTRODUCTION

As one of Europe's largest insurers, Achmea offers a comprehensive range of insurance products, which are economic protections from identified risks occurring or discovered within a specific period (Nissim, 2010). Within Achmea, the Non-Life Business to Business division focuses on non-life insurance for companies. Fire insurance is one of the most important products that the Non-Life Business to Business division offers to companies. With this insurance, buildings and various components on and in buildings are insured. Risk experts from the Risk Expertise Department (re-)inspect company buildings to assess and analyze risks associated with company buildings. They can only re-inspect approximately 200 buildings annually, so strict choices must be made as to which buildings to re-inspect or not to cope with limited re-inspection capacity. For that reason, a re-inspection prioritization method has been developed that prioritizes which companies should be re-inspected.

The prioritization method consists of a damage probability model and a damage burden model. The damage probability model and the damage burden model respectively predict the probability that damage will occur at a company building and what the impact will be if damage occurs. By combining the predicted probability and predicted burden of damage, the risk of damage for company buildings is predicted, as risk is uncertainty about and severity of the consequences of an activity that is typically assessed in terms of their likelihood of occurring and the potential severity of their impact (Aven & Renn, 2009). By prioritizing companies based on the predicted risks, this model serves as a data-driven decision-making (DDDM) tool.

Prioritizing and executing re-inspections is an operational procedure that is visualized by a Business Process Model in Figure 1. The data scientists trigger the prioritization method to predict the damage probability and damage burden and to make a prioritization based on the predictions. The resulting list of prioritized companies is forwarded to decision-makers, who are employees of the underwriting department, employees of the risk expertise department, and the insurance product manager. They make changes to the list where necessary, supplement the list, and authorize the list. The risk experts carry out the re-inspections according to the list. Additional details on the operational procedure for prioritizing re-inspections are in Appendix A.

The computation with which the re-inspection prioritization is made ('Make a prioritization based on the predictions' in Figure 1) differs depending on the scale of the customer's company. The prioritization method for small and medium-sized enterprises (SMEs) is created by multiplying the predicted probability and burden of the damage for each company and ranking the companies based on this priority score from high to low. The prioritization method for large enterprises is created by first ranking the outcomes of the predicted probability of damage and the predicted damage burden separately and then taking the inverse of this ranking number for both the predicted probability and burden. These two inverse numbers are added together to form a priority score. The priority score for large enterprises is then ranked from high to low to form the final prioritization for large enterprises. As a result, large enterprises that have a high score in the damage probability model or a damage burden model also rank high in the final prioritization method. The process of making the based on the predictions is visualized in Figure 2.

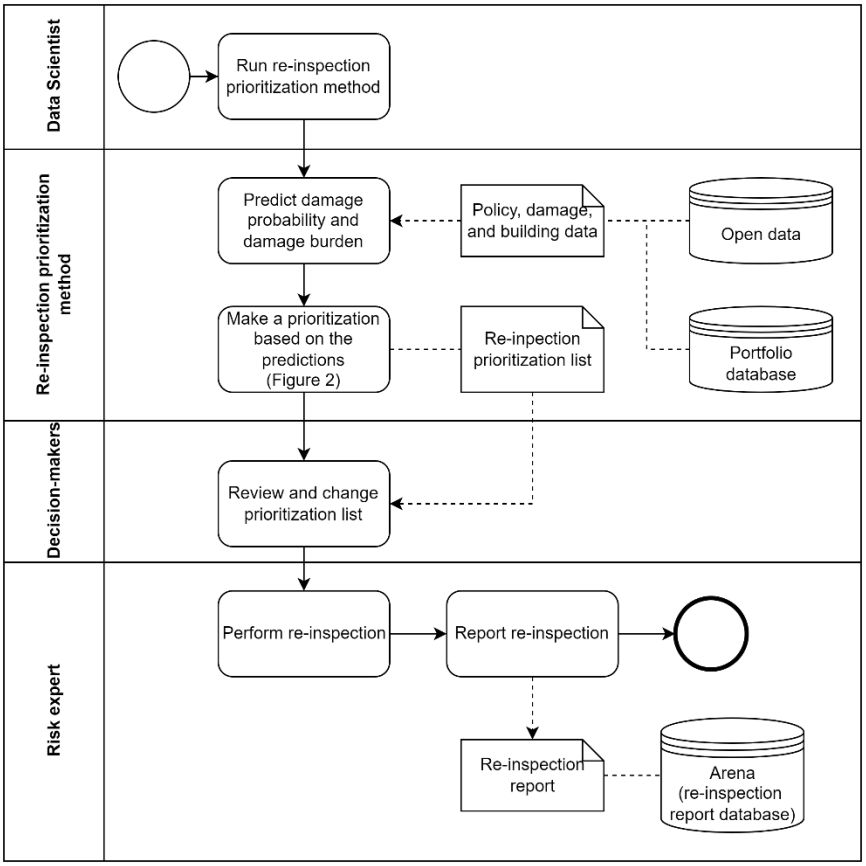

Figure 1: Business Process Model for selecting re-inspections using the prioritization method

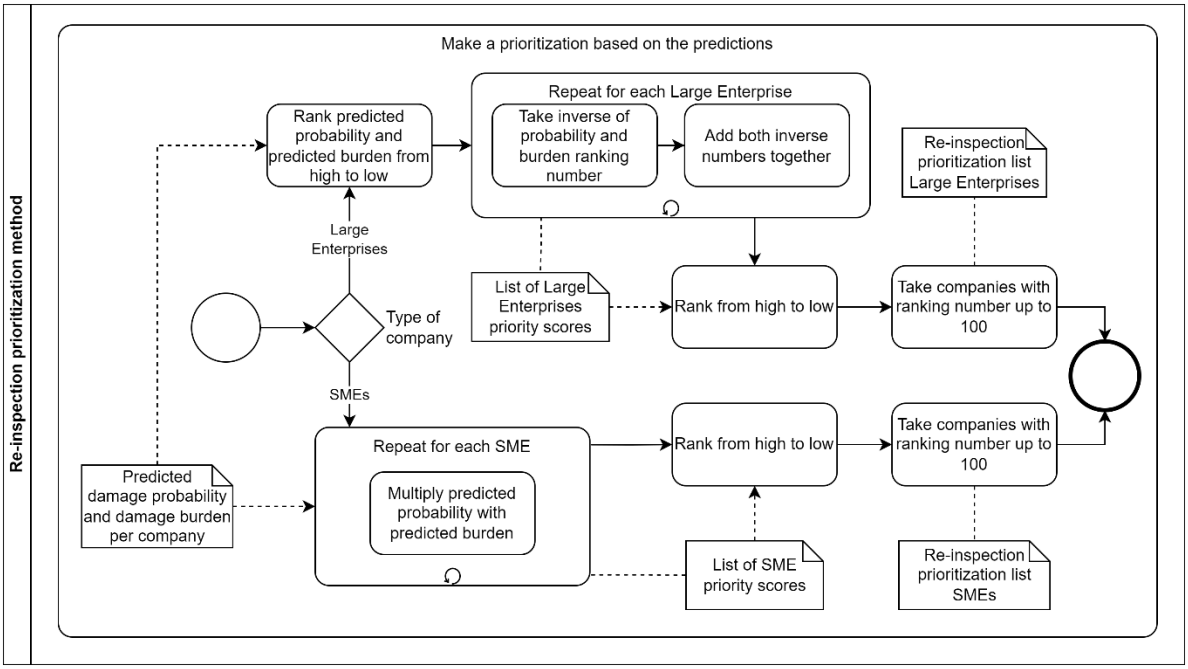

Figure 2: Business process model for the prioritization with the current re-inspection prioritization method, which is a sub-process of Figure 1 ('Make a prioritization based on the predictions').

1.1 PROBLEM ANALYSIS

In several cases, companies receive a high prioritization based on predicted risk but the actual value obtained from a risk expert's re-inspection is minimal. For example, there are cases where a company already faces substantial risks that cannot be further mitigated, resulting in a high predicted risk but no added value in re-inspecting the company. Moreover, the prioritization method does not consider earlier re-inspections, leading to companies being flagged for re-inspection even if they were recently inspected and perceived as not useful. This leads to redundant re-inspections.

The cause of the ineffective prioritization is that there is a lack of interaction between the risk experts and the re-inspection prioritization method because the method does not consider feedback and information provided by the risk experts. There are multiple consequences. On the one hand, the lack of feedback prevents the model from incorporating the experts' judgment, professional experience, and contextual understanding. As a consequence, the model's prioritization becomes misaligned with real-world re-inspection priority norms, meaning that companies are inspected unnecessarily and companies are not inspected that should have been inspected. On the other hand, the lack of feedback creates a sense of diminished confidence among risk experts, as they are unable to validate or refine the model based on their expertise. This lack of confidence hinders the adoption and acceptance of the prioritization method. In conclusion, the current (re-)inspection prioritization is inefficient due to the lack of interaction between the model and risk experts.

A potential solution is to develop a model that predicts the usefulness of a re-inspection according to risk experts. This predicted usefulness can be incorporated as a factor in the current prioritization method, enabling the feedback from risk experts to be integrated into the decision-making process. An organizational learning approach is essential to facilitate the interaction between risk experts and the model, allowing for the development of DDDM that adopts human norms. Implementing an organizational learning approach should lead to a situation in which the usefulness of a re-inspection is conceptualized by an interaction between human and machine, risk experts are involved in the development of the predictive model, the predictive model will be developed more effectively with the risk experts' knowledge, and the risk experts meanwhile will build trust in the prioritization method.

However, while retrospective research has been conducted on DDDM using an organizational learning approach, limited prospective research has been carried out into the development of a DDDM tool from an organizational learning approach. Moreover, there is no proven method to develop a DDDM tool that successfully enables triple-loop learning: organizational learning realized by an interaction of human and machine (Seidel, Berente, Lindberg, Lyytinen, & Nickerson, 2018). Therefore, there is a gap in scientific knowledge regarding how to effectively develop a DDDM tool, especially a predictive model, by triple-loop learning.

1.2 PROBLEM STATEMENT

The current re-inspection prioritization method lacks interaction with risk experts, resulting in inefficient prioritization and diminished confidence among experts. There is a need to develop a predictive model that incorporates risk experts' feedback to predict the usefulness of re-inspections. Facilitating triple-loop learning is essential, however, there is a lack of scientific knowledge on effectively developing a DDDM tool by which triple-loop learning will be achieved. This leads to the following main research question:

How can a model for predicting the usefulness of a re-inspection be developed by triple-loop learning?

1.3 OBJECTIVE

The objective is to develop a model that predicts the usefulness of a company re-inspection and incorporates this prediction into the current prioritization method to achieve triple-loop learning. This model should be updatable and capable of utilizing newly available data. The prioritization method for company re-inspections should consider the predicted probability and impact of damage occurring, as well as the predicted usefulness of the inspection. The practical aim is to increase the impact that risk experts can make with the prioritized re-inspections and to increase the acceptance of risk experts in the prioritization method. The scientific aim is to validate a method that enables triple-loop learning in the development of a DDDM tool.

1.4 OUTLINE OF THE PAPER

Chapter 2 gives a theoretical framework for predictive models and organizational learning, and combines both topics to arrive at a proposed methodology to develop a predictive model from an organizational learning perspective to enable triple-loop learning. Chapter 3 represents the research methodology for this study. Chapters 4, 5, and 6 contain the results of this research in chronological order. Finally, in chapter 7, the findings are discussed, and in chapter 8, conclusions are drawn.

2 THEORETICAL FRAMEWORK

In this chapter, a theoretical framework is outlined to arrive at a method with which a predictive model can be developed using an organizational learning approach. To arrive at this method, DDDM from an organizational learning perspective is first discussed. The concept of explainable AI is then reviewed because this is an important enabler of successful organizational learning. Then, the development of a predictive model is discussed. At the end of the theoretical framework, the knowledge from the theoretical framework is combined to arrive at the proposed method. The theoretical framework is built using a literature search. The literature search methods can be found in Appendix B.

2.1 DATA-DRIVEN DECISION-MAKING FROM AN ORGANIZATIONAL LEARNING APPROACH

This section explains what a decision is, covers the concept of DDDM and human in the loop, and discusses DDDM from an organizational learning approach.

2.1.1 Data-driven decision-making

A decision is a choice between two or more alternative courses of action (Hutton & Klein, 1999). Making a decision is a process where an individual or a group of individuals assesses various possibilities and selects one option to pursue. Individuals with the authority or responsibility to make decisions are decision-makers (Eisenhardt & Zbaracki, 1992). Making a decision involves several steps (Lunenburg, 2010). First, a goal has to be recognized and a problem in achieving that goal has to be identified. Then, relevant information is gathered and analyzed to generate and understand the available alternatives to achieve a goal. With this information, the decision-maker can evaluate the advantages and disadvantages of each alternative based on certain criteria. This evaluation involves considering the feasibility, satisfaction, and impact of the alternatives. The decision is made by selecting the alternative that appears to be the most favorable or promising according to the decision maker. Once a decision is made, the chosen course of action is implemented and executed. The decision-making process is iterative, as the evaluation of the outcomes of decisions and the impact of decisions lead to new decisions being made.

DDDM is the practice of making informed decisions based on the analysis of data rather than exclusively on intuition or personal judgement (Brynjolfsson, Hitt, & Kim, 2011). At the core of DDDM is the belief that data can provide valuable information, which is considered to be crucial for reducing risks, improving outcomes, and optimizing performance (Cech, Spaulding, & Cazier, 2018). Improvements in the collection and processing of data will also generate new insights for decision-making (Brynjolfsson et al., 2011; L. Wu, Hitt, & Lou, 2020). Instead of relying on gut feelings or assumptions, decision-makers use data to analyze complex problems in more detail, evaluate potential options, and predict scenarios. DDDM extends beyond data analysis by promoting evidence-based conclusions and facilitating proactive, informed choices. DDDM is rooted in different subdisciplines, such as machine learning (Fu, Xu, Xue, Liu, & Yang, 2021), business intelligence (Chen, Chiang, & Storey, 2012), and data science (Provost & Fawcett, 2013). Research has shown that DDDM is positively related to decision-making quality within organizations (L. Li, Lin, Ouyang, & Luo, 2022).

The presence of human decision accountability, decision complexity, problem ambiguity, and decisional uncertainty often prevents DDDM from resulting in a complete automation of the decision-making process in which computers replace human decision-making (Jarrahi, 2018). Instead, DDDM facilitates an augmentation of the decision-making process, which means that people collaborate closely with machines to make a decision (Raisch & Krakowski, 2021). People and machines should

combine their complementary strengths, enabling mutual learning and multiplying their capabilities (Kokina & Davenport, 2017). This emphasizes the need for human-in-the-loop (HITL), an approach where people are directly involved or integrated into a system or process that relies on automated or AI technologies (X. Wu et al., 2022). There are multiple reasons or conditions in which direct human involvement is needed in automated or AI technologies. Human involvement is needed as people can comprehend and interpret contextual information to make informed decisions in the broader context (X. Wu et al., 2022), people can provide clear explanations of decisions that ensure transparency (Mosqueira-Rey, Hernández-Pereira, Alonso-Ríos, Bobes-Bascarán, & Fernández-Leal, 2023), human intelligence and expertise can avoid potential pitfalls and biases that arise from purely algorithmic approaches (Mosqueira-Rey et al., 2023), and human feedback leads to refining algorithms, updating models, and addressing limitations of DDDM applications (Jarrahi, 2018). Since people and machines have to work together and understand each other in DDDM, an organizational learning approach for DDDM is important.

2.1.2 Organizational learning with DDDM

Organizational learning is the change that occurs as an organization acquires, creates, retains, and transfers knowledge (Argote & Miron-Spektor, 2011). Organizational learning consists of single-loop learning, double-loop learning, deutero learning, and symbiotic learning (Jarrahi, 2018; Wijnhoven, 2022). Single-loop learning is the creation of improvements within existing processes or an existing framework. Double-loop learning goes beyond single-loop learning by reflecting on the existing process or framework and is therefore the creation of innovation. Deutero learning is the creation and development of norms, rules, and conditions by which the knowledge-creation processes can be done. It can be seen as a process to discover elements necessary for learning, such as the infrastructure needed, policies, and the setting of norms and rules of a system. Symbiotic learning is the implementation of deutero learning outcomes.

Within organizational learning, a distinction is made between two types of knowledge convertible into each other: tacit knowledge and explicit knowledge (Nonaka, 1994; Wijnhoven, 2022). Tacit knowledge is knowledge that is based on an individual's experiences, skills, insights, and intuitions. It is often deeply ingrained in individuals, difficult to express in words, and thus person dependent. Explicit knowledge is formalized, communicable knowledge through an explicit representation. Tacit and explicit knowledge are continuously transformed and expanded following the SECI (Socialization, Externalization, Combination, and Internalization) model given in Figure 3. Tacit knowledge can be shared among individuals through direct interaction and experience via socialization. Tacit knowledge can also be made person independent by externalizing it via coding and writing, resulting in explicit knowledge. Explicit knowledge can be combined, which involves the integration of explicit knowledge from various sources. Explicit knowledge can also be transformed and incorporated into an individual's tacit knowledge and integrated with an individual's own experiences, values, and skills, which is named internalization.

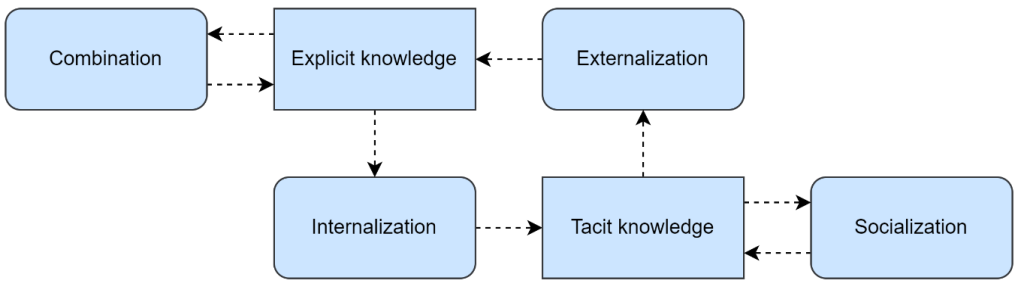

Figure 3: SECI (Socialization, Externalization, Combination, and Internalization) model, based on (Nonaka, 1994). The boxes with curved angles represent knowledge creation processes, and the boxes with sharp angles represent knowledge stocks.

The implementation of DDDM requires organizations to recognize a novel approach to decision-making, combining the creation of decisional knowledge through single- or double-loop processes with the integration of machine learning (Wijnhoven, 2022). Single- or double-loop learning realized by an interaction of human and machine learning is called triple-loop learning (Seidel et al., 2018). In triple-loop learning, single-loop learning occurs when people and machine collaborate to generate design outcomes (Wijnhoven, 2022). Double-loop learning within triple-loop learning involves people assessing design alternatives and adjusting input parameters and machine settings based on feedback and the machine learning from human feedback to enhance its model and to produce improved alternatives. So triple-loop learning assumes that human and machine single-loop and double-loop learning may mutually influence each other. Triple-loop learning is driven by the mental models present in both machine (the DDDM tool) and human (Seidel et al., 2018). These mental models encompass goals, cognitive rules, and reasoning. In triple-loop learning, there is an assumption that the mental models of human and machine mutually influence each other, leading to the enhancement of their mental models through error correction (single-loop) and the revision of norms (double-loop). Deutero learning outcomes can enable triple-loop learning in DDDM by creating elements such as motivations, conditions, facilities, infrastructure, and policies to effectively use data science techniques in decision-making.

A DDDM tool can provide valuable knowledge that generates fresh insights for people to internalize, resulting in the acquisition of new tacit knowledge and fostering intelligence amplification (Metcalf, Askay, & Rosenberg, 2019; Wijnhoven, 2022). Moreover, symbiotic learning can be enabled, meaning that machine and human work together and leverage their strengths and capabilities to enhance learning outcomes (Jarrahi, 2018). Symbiotic learning overlaps triple-loop and deutero learning, as it enables the realization of triple-loop learning by adopting and implementing the deutero learning outcomes.

2.2 PREDICTIVE MODELING

Predictive modeling is the process of using historical data, statistical algorithms, and (machine learning) techniques to make predictions about (future) events or outcomes (Waljee, Higgins, & Singal, 2014). Predictive modeling has a wide range of applications across industries and attempts to assist organizations in gaining insights, making informed decisions, and optimizing their operations for better outcomes. These models can be used as a tool to help decision-making, planning, risk assessment, resource allocation, and strategic forecasting in various industries (Waljee et al., 2014). Predictive models can be divided into prediction models and forecasting models. Prediction models make predictions based on historical data, assuming time invariance. They capture patterns and relationships between variables. Forecasting models explicitly consider the temporal aspect. They analyze time series data, capturing time-dependent patterns and trends to forecast future values. Prediction models focus on relationships, while forecasting models incorporate the time dimension for accurate predictions. In the remainder of this report, a predictive model refers to a prediction model because it is expected that the historical data available does not exhibit significant time-dependent patterns or trends. Predictive models can be built using multiple algorithms and techniques. The predictive models used in this research are multiple regression, random forests, and neural networks. To obtain a theoretical basis for this, a theoretical reflection has been added in Appendix C. This section dives into the concept of explainable AI and describes the phases of developing a predictive model.

For DDDM interpretability and organizational learning, predictive models need to be well explainable. Explainability stands out as a primary barrier in the practical implementation of AI (Arrieta et al., 2019). Explainability is also significant in the context of organizational learning with DDDM as the lack of

explainability may hinder the human internalization process (Wijnhoven, 2022). This barrier arises from the inability to understand the reasons by which predictive modeling algorithms perform, which is a problem that finds its roots in two different causes (Arrieta et al., 2019). Firstly, a gap between the research community and business sectors hinders the seamless integration of machine learning models into sectors. Secondly, research predominantly centers on results and performance metrics, a focus that may benefit certain disciplines in assisting AI to infer relations beyond human cognitive reach. Nevertheless, the focus on accuracy alone is increasingly coming under criticism as it hinders users from assessing, understanding, and correcting the system (Nauta et al., 2023). Hence, the necessity for explainable machine learning algorithms emerges, aiming to enhance the transparency of AI systems and make their outcomes more understandable to people (Nauta et al., 2023).

Explainable AI (XAI) encompasses a diverse range of explanations, presenting aspects of the reasoning, functioning, and behavior of a machine learning model in terms understandable to people (Nauta et al., 2023). Decision trees, represented as rooted graphs with conditional statements at each node, offer intuitive visualizations of decision pathways. Natural language processing techniques provide a textual explanation. Inherently interpretable white-box models, such as a scoring sheet or linear regression, offer explanations through the model's structure and parameters. Representation synthesis or representation visualization explains the form of visualizations, such as feature visualization, scatter plots, or cluster analysis, to explain a predictive model's representations. Feature importance is a set of non-binary scores to indicate feature relevance. These various methods of XAI enable people to gain a deeper understanding of machines, fostering them to learn from a machine.

Predictive modeling consists of explicit learning steps, which are defining the project objective, collecting and preparing data and selecting features, selecting and specifying a model, training and validating the model, and model presentation and implementation (Jakeman, Letcher, & Norton, 2006; Steyerberg & Vergouwe, 2014; Waljee et al., 2014).

Before starting with the development of a model, it is key to clearly define the problem that has to be solved and the objective that has to be achieved. The problem definition step consists of identifying the problem, defining the purposes and the scope, and building domain knowledge that is necessary to understand and solve the problem (Jakeman et al., 2006; Steyerberg & Vergouwe, 2014). The problem identification results in a problem statement that the predictive model should solve. The objective includes specifying what the model will predict. Scope definition is about defining the scope and boundaries of the predictive modeling project, taking into account factors such as resource availability, time constraints, and technical challenges. Despite challenges that may arise from various stakeholder interests, achieving a comprehensive understanding is beneficial for all parties involved, aiding in the definition of the problem and exploration of potential solutions (Jakeman et al., 2006). The problem and objective definition in the context of organizational learning can be seen as the externalization and combination of some knowledge.

The second step is to conceptualize the functioning of the system that has to be modeled. In this step, data, prior knowledge, and assumptions about the relevant processes are defined. Prior knowledge includes observational data, structural information, process characteristics, and parameter values with uncertainties. The step begins qualitatively by exploring knowledge about processes, available records, and instrumentation compatibility, and passes to a quantitative phase where decisions are made regarding inclusion, simplification, and neglect of variables, while also assessing their sensitivity (Jakeman et al., 2006). This can be seen as externalization and combination to arrive at combined explicit knowledge. Conceptual models that can be useful to set up during this phase are for example a business process model about the as-is and to-be situation, a causal model to visualize the correlation of variables, and a data structure model. Thoroughly examining and understanding business objectives

and project goals is essential for effective predictive modeling. This critical phase combines business perspective, understanding, and data comprehension, which is needed for alignment between modeling efforts and desired outcomes (Lukyanenko, Castellanos, Parsons, Chiarini Tremblay, & Storey, 2019).

A domain analysis increases the chance of successful development of a predictive model (van der Spoel, 2016). The type of domain analysis needed depends on the complexity of a domain. A complex domain has coordinated human actions, politics, myths, meanings, unstructured nature, partial observability, individual goals, probabilistic elements, environmental interactions, and behavioral influences (Jackson & Keys, 1984). The domain analysis can be made specific by collecting domain knowledge using brainstorming and a field study, which provides hypotheses for making predictions and identifies constraints that determine the usability and relevance of predictive models (van der Spoel, 2016).

The next step is to collect and prepare relevant data and select features from this data. This step consists of data preparation and feature selection, which are intertwined. The step involves identifying and gathering relevant data from various sources, ensuring data quality, and transforming the data into a suitable format for model training. Feature selection is crucial in preparing data for machine-learning problems. The goal of feature selection is to reduce dimensionality and prepare clean, understandable data (J. Li et al., 2018). Finally, the categorical and continuous variables of the model should be coded in this step.

Choosing the appropriate predictive model algorithm involves evaluating different models and selecting the one that performs best based on a specific evaluation metric, which depends on the problem, the available data, and the desired outcome (Waljee et al., 2014). The choice of a model structure can be made with prior scientific knowledge, which is not always sufficient, or by trial and error among a modest number of possibilities based on the credibility of model behavior (Jakeman et al., 2006).

For the training and validation of the model, the dataset is split into two sets: a training set and a validation set. During model training, the model learns from the training data by adjusting its parameters to capture patterns and relationships. This process involves optimization algorithms and aims to minimize the discrepancy between predicted and actual values. Once the model is trained, it is necessary to validate its performance on unseen data. Model validation involves measuring relevant metrics to assess how well the model generalizes. Through proper training and validation, models can be optimized and assessed for their effectiveness in making accurate predictions.

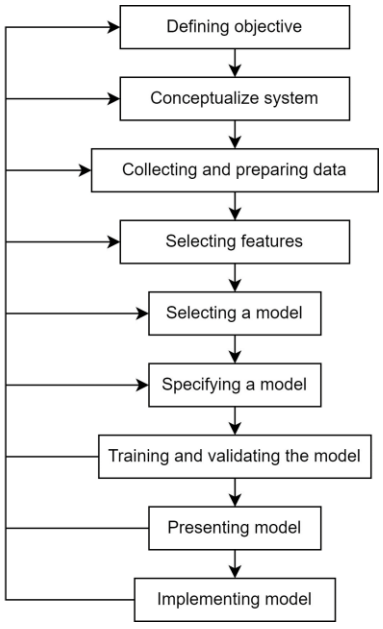

Figure 4: The iterative process of predictive model development, inspired by (Jakeman et al., 2006)*, but with modifications based on other literature input as described in 2.2.*

The final step is to present, introduce, and implement the predictive model (Steyerberg & Vergouwe, 2014). Model presentation is important to transform the model from a theoretical construct to a practical solution by effectively communicating the insights and capabilities. Model presentation facilitates understanding among stakeholders, enabling informed decision-making and building trust in the model's predictions. Implementation, which often regards the integration of the model into existing applications, is also necessary to make the model effective. The model presentation can be seen as

newly created explicit knowledge that is combined with existing explicit knowledge, and the presentation and implementation together can be seen as enablers for internalization of the newly created explicit knowledge.

The steps above are an iterative learning process (Jakeman et al., 2006). A model is developed based on assumptions and available knowledge, and stakeholders may not understand their own needs fully. A model is rarely perfect or optimal right from the start. Through iterations, the model can be refined, improved, and adapted to better fit the problem domain. In addition, iterations enable developers to adapt the model to changing data distributions, refine it based on observed errors or biases, and ensure its ongoing relevance and reliability in dynamic real-world scenarios. The described steps together can be found in Figure 4. Depending on the cause of the new iteration, the iteration might only contain respecifying the model, but it can also mean adjusting the goals and repeating the entire process.

2.3 PREDICTIVE MODEL DEVELOPMENT AS AN ORGANIZATIONAL LEARNING PROCESS

It has been described above how learning is crucial for predictive modeling. This section describes how this learning is organizational by nature in DDDM, and how triple-loop learning can be reached in the predictive model development method. Figure 5 shows how predictive modeling learning is added to human learning. Figure 5 is a combination of the SECI model given in Figure 3 and the predictive model development methodology given in Figure 4. In Figure 5, the solid arrows represent a machine development process flow, the broken arrows represent a learning process flow, the blue boxes with curved angles represent knowledge creation processes, the blue boxes with sharp angles represent knowledge constructs, and the white boxes with sharp angles represent machine learning steps. The proposed methodology is an iterative process that can be executed as single- or double-loop learning. The proposed methodology contains the following stages of learning, where the numbers of the phases correspond to the numbers placed in circles in Figure 5:

1. *Externalization* The process in Figure 5 starts as experts in a field provide input on what is important to them in the functioning of a predictive model. They externalize their tacit knowledge to explicit knowledge and help the predictive model developers with defining the objective and conceptualizing the predictive model with their domain knowledge. The developers also externalize their tacit knowledge to explicit knowledge by coding.
2. *Combination of human knowledge* The explicit knowledge is combined with existing explicit knowledge. An example of a combination process is that learning outcomes are combined with explicit data knowledge, and based on this combination, knowledge is created about how the learning outcomes can be implemented in the machine. A part of the combination of human knowledge is the problem and objective definition and the conceptualization of the system as described in section 2.2. The problem, objective, and conceptualized system therefore belongs to the human mental model.
3. *Machine learning* The combined explicit knowledge enables the model to learn. The predictive model passes a development iteration as described in section 2.2, and the machine learning process starts. Machine learning in this context is not a machine learning algorithm, but the process of which the predictive model is learning.
4. *Combination of machine knowledge* At the end of the cycle, the predictive model is presented and possibly even implemented, which leads to new explicit knowledge. A part of the combination of machine knowledge is the presentation and implementation of the predictive model as described in section 2.2. This explicit knowledge is combined with other explicit knowledge where necessary. This creation of new explicit knowledge and combination with

existing explicit knowledge can be seen as a change in the mental model of the machine triggered by the human mental model. The explicit knowledge is then internalized, which is the starting point of a human learning process.

5. *Internalisation* The involved individuals, such as the model developer and the domain experts, internalize explicit knowledge into tacit knowledge.
6. *Socialization* By interacting with each other and sharing experiences, this tacit knowledge undergoes socialization. As a result of internalization and socialization, experiences are gained and norms and values are created. Learning outcomes will rise during this process. After socialization, the learning outcomes can be externalized. The learning outcomes trigger a change in the mental models of human influenced by the machine mental model, and a new learning cycle will be initiated.

Both humans and machines are capable of generating new explicit knowledge throughout this process. Within the proposed methodology, they form a constant cycle in which the machine learns from the tacit knowledge generated by people, and people learn from the explicit knowledge generated by machines. So the proposed methodology represents a continuous learning process. By enabling symbiotic learning, or implementing deutero learning outcomes that follow from single- and double-loop learning processes, a predictive model should be developed that is based on human feedback, and thus triple-loop learning should emerge.

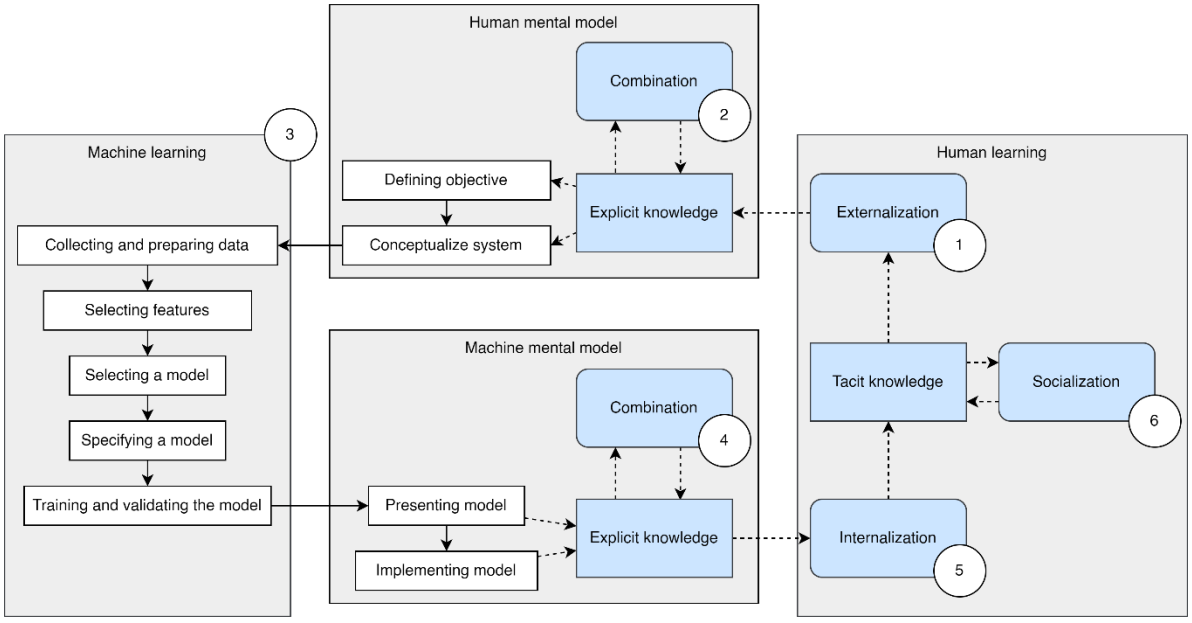

Figure 5: Framework to develop predictive model as DDDM tool by triple-loop learning. The solid arrows represent a machine development process, the broken arrows represent a learning process, the blue boxes with curved angles represent knowledge creation processes, the blue boxes with sharp angles represent knowledge stocks, and the white boxes with sharp angles represent machine learning steps. The proposed methodology is based on the theory described in 2.1.2 and 2.2.

2.4 CONCLUSION THEORETICAL FRAMEWORK

In conclusion, the theoretical framework reflects contemporary theoretical knowledge about organizational learning and predictive modeling. Integrating knowledge about these subjects has led to the development of a proposed methodology for creating predictive models as DDDM tool through triple-loop learning. This research is built on the proposed methodology.

3 RESEARCH DESIGN

The research design is based on the proposed methodology to develop predictive models as an organizational learning process illustrated in Figure 5. Section 3.1 introduces the research questions used within this research and provides an overview of the knowledge collected. Section 3.2 delves into how human involvement is realized, representing externalization (1) and combination (2) of human knowledge before the machine learning phase, and internalization (5) and socialization (6) after the machine learning phase of Figure 5. Section 3.3 gives the research design for machine learning (3) and the combination of machine-created knowledge with existing knowledge (4) in Figure 5.

3.1 RESEARCH QUESTIONS

The main research question is:

How can a model for predicting the usefulness of a re-inspection be developed by triple-loop learning?

The main research question consists of two components: the development of a re-inspection usefulness predictive model and the implementation of it within the re-inspection prioritization method. The main research question will be answered using the proposed methodology in section 2.3. When placing the research question in the context of Figure 5, usefulness is a created norm that changed the human mental model and thus is the starting point for this research.

As part of the development, it must be investigated which features can be selected from the data from the available data sources and which algorithms are suitable for the requirements and the available data. A usefulness predictive model must emerge that is most suitable and can be further developed. Following Figure 5, the development is the single-loop and double-loop learning process. This leads to the following research question and sub-questions:

RQ1: What model predicts the usefulness of a re-inspection?

- *SQ1.1: What is the usefulness of a re-inspection?*
- *SQ1.2: Which features can be selected out of the available data sources?*
- *SQ1.3: What is the optimal algorithm for building the predictive model based on the available data?*

Parallel to the development of the model, it must also be implemented. It must be integrated within the re-inspection prioritization mode currently in use and it must be implemented within the organization in a manner appropriate to the organizational culture. To ensure successful integration of the prioritization method within the organization, triple-loop learning must be achieved. Following Figure 5, the implementation of the model is about deutero and symbiotic learning. Based on this, the following research question and sub-questions are defined:

RQ2: What is needed to implement and activate the predictive model for re-inspection usefulness within the organization?

- *SQ2.1: What is an appropriate method for integrating the predictive model for re-inspection usefulness into the current re-inspection prioritization method?*
- *SQ2.2: What is needed to enable stakeholders to utilize the results generated by the predictive model?*

Within the organizational learning context, RQ1 pertains to single-loop and double-loop learning, while RQ2 focuses on deutero and symbiotic learning. The combined insights from both questions are aimed

to enable triple-loop learning of the usefulness predictive model and the prioritization method and should answer the main research question.

To develop a predictive model for the usefulness of re-inspections, the right knowledge has to be collected. The knowledge needed to develop a re-inspection usefulness predictive model can be divided into four types of knowledge: domain knowledge, data knowledge, procedural knowledge, and methodological knowledge. Domain knowledge is about knowledge of the domain and includes the definition of the usefulness of a re-inspection and the purpose of the re-inspection usefulness predictive model. Data knowledge is about the data that is needed to predict the usefulness. Procedural knowledge is about the decision-making process with which re-inspections are prioritized. Methodological knowledge involves understanding the methodology used to develop a usefulness predictive model from an organizational learning approach. Figure 6 shows a diagram of the knowledge that is required, classified by the four types of knowledge. The figure distinguishes between knowledge that is gathered already via literature review, knowledge that can be gathered via desk research (among others data exploration and the exploration of the technical functioning of current models), and knowledge gaps that require human input to externalize and combine tacit knowledge to achieve explicit human knowledge. The specific knowledge that is needed is as follows:

- *Domain knowledge* To determine the definition of the usefulness of a re-inspection, it must be clear what the purpose of re-inspections is and which factors influence this usefulness. Clarifying the definition of the usefulness of re-inspections and the advantages and disadvantages of the current prioritization method should contribute to settle the purpose of the usefulness predictive model. Furthermore, to understand the decision-making process and the decision-making norms that the prioritization method should have., knowledge is needed about the current process of prioritizing re-inspections and about the advantages and disadvantages of the re-inspection prioritization method.
- *Data knowledge* As part of feature selection for the re-inspection usefulness predictive model, it should be clear what factors influence the usefulness, which data fields needed to predict the re-inspection usefulness are available, and how the data is structured.
- *Methodological knowledge* Based on methodological knowledge that is collected by combining literature findings about predictive modeling and organizational learning, a methodology is proposed to develop a predictive model while enabling triple-loop learning. This study in turn contributes to new knowledge about how to achieve triple-loop learning in the development of predictive models.

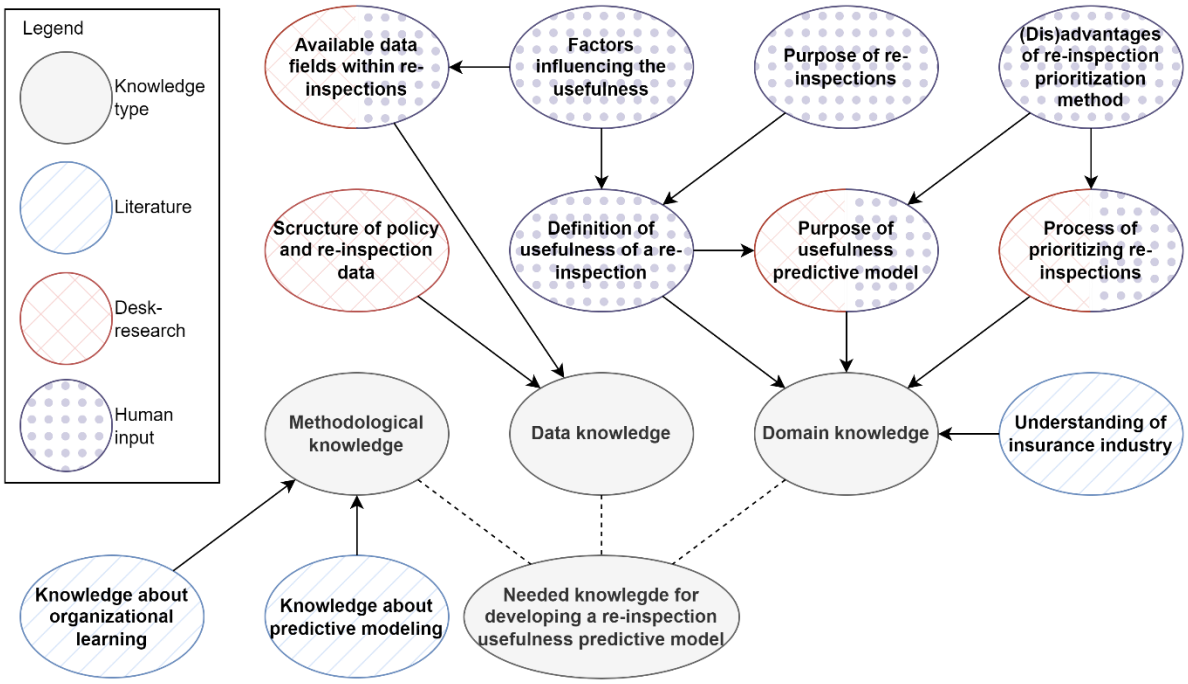

Figure 6: Diagram of the knowledge needed to develop a re-inspection usefulness predictive model, distinguishing between the knowledge already available via literature or desk research, and the knowledge to be gathered by human input. The gray ovals represent a type of knowledge, and the ovals filled with a pattern represent a knowledge subject. The solid arrows represent dependencies between knowledge subjects, the dashed lines are used to classify the required knowledge into three different types.

3.2 HUMAN INVOLVEMENT

Human involvement is crucial in this research, as the organizational learning approach emphasizes mutual learning between humans and machines. Therefore, it is key to engage individuals and gather human knowledge. This knowledge is collected and integrated at two key stages in the research process: interviews with people are conducted before the machine learning phase, and evaluations with people are conducted after the machine learning phase.

Interviews are conducted to externalize human tacit knowledge and analyzed to combine the externalized knowledge into explicit knowledge. The interviews are semi-structured, with a heterogeneous group of participants. The function title of the participants can be found in Table 1. The questions that are used for the semi-structured interview and the types of knowledge that are gathered per interview participant can be found in Appendix D and a summary of the subjects discussed with each interview participant can be found in Table 2. The interview subjects correspond with the human input knowledge subjects in Figure 6. The interviews are recorded and transcribed. Subsequently, these transcriptions are shared with the participants for their consent to be included in the research. Participants are also given the option to redact specific portions of the interview transcription if they wish to do so. Ethical permission from the Ethics Committee Computer and Information Science of the University of Twente has been granted to perform the interviews[1]. The interviews are analyzed by recording them, transcribing them, and then coding statements according to specific topics so that the statements could be grouped on a specific topic. An analysis was then made of the different participant perspectives on that topic. Based on the interview analysis, a conceptual model is made for the usefulness of re-inspections, and learning lessons are identified that

[1] https://www.utwente.nl/en/eemcs/research/ethics/ request number 230519

provide guidelines for the machine learning stage. The conceptual model should clarify the concept of the usefulness of re-inspections and which factors influence the usefulness of re-inspections. Based on the conceptual model, it should be able to select features in the machine learning phase.

Table 1: Participants and their function title within Achmea

Participant number	Function title
P1	Risk expert
P2	Manager risk expertise department
P3	Underwriter of business fire insurance
P4	Manager underwriting department of business fire insurance
P5	Product manager of business fire insurance
P6	Data scientist 1
P7	Data scientist 2

Table 2: Subjects discussed per interview participant

	Participant number						
Subject	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**
Purpose of re-inspections	X	X	X	X	X	X	X
Definition of the usefulness of a re-inspection	X	X	X	X	X	X	X
Purpose of usefulness predictive model	X	X			X	X	X
Factors influencing the usefulness	X	X	X	X	X	X	X
Available data fields	X	X			X	X	X
Process of prioritizing re-inspections	X	X	X	X	X	X	X
(Dis)advantages of the prioritization method	X	X	X	X	X	X	X

After the machine learning and the combination of the explicit machine knowledge, the usefulness predictive model and the prioritization method have been evaluated by different stakeholders. The stakeholders involved in the evaluation are P1 to P5 of Table 1, who are the decision-makers and end-users of the prioritization method. P6 and P7 (data scientists) have not been involved in the evaluation because data scientists were closely involved in the development process. The evaluation, like the interviews, was analyzed by recording them, transcribing them, and then coding them. An analysis was then made based on different perspectives on topics. First of all, the evaluations are intended to discuss the methods and results of developing and implementing the usefulness predictive model. Stakeholders are involved in how features are selected and how the usefulness predictive models are assessed for their performance. The evaluation is also intended to address topics that emerged from the interviews as being important for the machine mental model. During the evaluation, these aspects are discussed in more detail. Using what-if statements and different scenarios to represent topics, topics that may be complicated to comprehend for stakeholders have been made more understandable so that the stakeholders are still able to internalize those topics. During the evaluation, participants are asked to reflect on topics that are discussed. With this reflection, it is examined how the human and the machine mental model change.

3.3 MACHINE LEARNING

The development of the re-inspection usefulness predictive model as part of the re-inspection prioritization method will be accomplished using the predictive modeling development approach following section 2.2 and using different predictive modeling algorithms. Attention is paid to the relevant enterprise systems and the data preparation as part of the machine learning, considering that the data is sourced from multiple repositories and has not been consolidated before. After the completion of the machine learning, the newly created machine knowledge is combined with existing

knowledge by performing a simulation of the usefulness predictive model. The simulation provides insight into how the usefulness prediction included in the re-inspection prioritization method influences the usefulness of prioritized re-inspections.

Initially, by varying algorithms, different predictive models are created that need to be compared based on their technical performance. To compare the predictive models, k-fold cross-validation is applied. K-fold cross-validation is a method in which the dataset is divided into k equal parts, or 'folds' (Berrar, 2019). The model is trained and evaluated k times, each time using a different fold as the test set and the remaining folds as the training set. The model performance is averaged over the k iterations, providing a reliable estimate of the model's performance that is less dependent on the randomness of the data split. In this case, with little data, cross-validation techniques are needed. Five folds have been used in this research. Cross-validation is used to configure the hyperparameters of the modeling algorithms—external settings that influence the model's architecture or learning process—and test these configurations on the model's outcomes. In that way, for each type of algorithm, the hyperparameters are chosen that perform best on average based on the performance metric.

The classifier predictive models are validated using the accuracy and the macro-averaged F1 (ma-F1) score. The accuracy, depicted in (1), quantifies the overall correctness of predictions and provides a general measure of the model's effectiveness. However, the accuracy is not sufficient for imbalanced datasets as accuracy might be high even if the model performs poorly on classes that are underrepresented in data. The ma-F1 score is based on the F1 score, which can be calculated per class according to (2). The F1 score considers both the precision's focus on accurate positive predictions and the recall's emphasis on capturing actual positive instances. A high F1 score indicates a model that achieves a balance between minimizing false positives and false negatives, making it a valuable metric for tasks where achieving precision and recall are equally important. The ma-F1 score uses the F1 score that is calculated for each class and then averages these scores according to (3). The ma-F1 score is sensitive to the performance of minority classes, providing a more balanced assessment of the model's ability to perform well across all classes. Therefore, the ma-F1 score is particularly useful when there is a significant class imbalance. For the technical evaluation of the classifiers, the accuracy and the ma-F1 score are both considered.

$$Accuracy = \frac{Correct\ predictions}{All\ predictions} \tag{1}$$

$$F1_i = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2}$$

$$\text{ma-F1} = \frac{1}{\mathrm{K}} \sum_{\mathrm{i}=1}^{\mathrm{K}} \mathrm{F1_i} \tag{3}$$

The regressor predictive models are validated using the Mean Squared Error (MSE), a measure of the average squared differences between the predicted values and the actual values. For this research, MSE is preferred over other metrics such as Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) because MSE penalizes large errors significantly more than MAE and MAPE by squaring the error terms. The formula for calculating the MSE is shown in (4) and involves taking the average of the squared differences between predicted and actual values for each data point. A lower MSE indicates a better fit of the model to the data, as it signifies smaller deviations between predicted and actual values. However, the MSE does not take the imbalance of datasets into account. Therefore, the MSE is combined with the ma-F1 score to assess the performance of the regressors.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \quad (4)$$

The simulation is performed using the prioritization method output of 2023, which contains two lists: one of prioritized SME and one of large enterprises. The usefulness of a re-inspection was predicted for the top 500 prioritized business objects from both lists. The training data encompasses all re-inspections not included in the priority list. Including the re-inspections from the prioritization list in the training set would grant the usefulness predictive model prior knowledge about the usefulness of these re-inspections, incorrectly boosting the results. Due to the difference in the train-test split used for the simulation and used for the already trained models, the models must be retrained.

The simulation can be used to calculate the change in the average usefulness score of re-inspections. The usefulness score is known because the re-inspections were carried out and assessed on its usefulness in 2023. Only the top 100 re-inspections from the priority list were conducted and assessed with a usefulness score. This leads to an issue when simulating changes in the top 100 in the priority list because re-inspections that were originally outside the top 100 and move to the top 100 do not have a known usefulness score. The change in the average usefulness can then not be measured. To solve this issue, the change in average usefulness in the top 50 is simulated, leading to newcomers in the top 50 that almost all have a usefulness score.

4 EXTERNALIZATION AND COMBINATION OF TACIT KNOWLEDGE

In this chapter, the prioritization method currently in use is examined. The advantages and disadvantages of the prioritization method are analyzed based on stakeholder perspectives, and the futural requirements of the prioritization method according to stakeholders are outlined. Additionally, a definition of the usefulness of re-inspections is formulated by combining perceptions from different stakeholders. The overarching objective of this chapter is to externalize human tacit knowledge into explicit knowledge and to combine this newly generated explicit knowledge, as illustrated in Figure 5.

4.1 STAKEHOLDER PERSPECTIVE ON THE PRIORITIZATION METHOD

Initially, 100 companies were selected by the prioritization method to undergo re-inspections as a pilot. Of this selection, 30 were excluded because the customer had canceled the policy or because the companies had recently been inspected, and 70 were carried out. The data scientists wanted to compare the re-inspections selected with the prioritization method with the re-inspections selected by expert judgement on their average usefulness and on the damage occuring. However, of the 70 re-inspections carried out, some had an incomplete usefulness score or were not linkable to the policy data, which meant they were eliminated. Ultimately, the data scientists analyzed 35 re-inspections selected by the prioritization method. The analysis showed that concerning re-inspecting the companies where damages will occur, the re-inspections chosen with the prioritization method were slightly better, while concerning the usefulness assessed by the risk expert, the regular inspections scored slightly better. However, there was considerable variation in the data and the difference was not significant. This was considered good news by the data scientists. P7 (data scientist): *"We were able to make a prioritization based on data that was no better, but also no worse, than how things had gone until then. Moreover, we were able to explain why we went to certain locations, while previously this was often based on sentiment only."*

The fact that the re-inspections are linkable to policy data to a limited extent or that the usefulness score is not always filled in is a result of a lack of data standards and leads to limited, not-significant results in the evaluation of the DDDM tool, and thus limited learning outcomes. Learning outcomes are expected to be greater when data standards are introduced that improve the data quality. Therefore, deutero learning found place: data standards have been developed in the meanwhile. Symbiotic learning based on this deutero learning outcome occurred to some extent, resulting in the usefulness score being filled in for most inspections but the inspection data still hardly being linkable to policy data.

Despite the limited analysis possibilities of the re-inspection prioritization method, stakeholders have been able to internalize and socialize their experiences with the DDDM tool, and based on the experiences from the pilots, stakeholders have been able to form an idea of the advantages and disadvantages of the prioritization method, which are learning outcomes that might have changed the human or machine mental model. An advantage stakeholders have observed is that the prioritization method is based on data and, as a result, can prioritize based on the entire portfolio. Where expert judgment may tend to be influenced by incidents at companies, personal opinions of experts, and limited insight into the entire portfolio, the prioritization method can predict where damages can be expected based on damages that have occurred in history and help decision-makers make decisions about the entire portfolio based on recurring patterns. As a result, the prioritization method contributes to the efficient deployment of risk experts by reducing the costs of damage with minimal staffing expenses.

In addition, the prioritization method provides new insights into re-inspections that would not be available without the prioritization method. P2 (the manager of the risk expertise department) emphasizes that although risk experts themselves may have an intuitive sense of the severity of damages based on expert judgment, they do not have insights into the actual damage at companies. P3 (underwriter) notes that new insights can be obtained and that these insights may not emerge based on expert judgment alone. Subsequently, these new insights lead to the identification of companies with specific risks in the prioritization method, whereas they might go unnoticed through expert judgment because it was unknown that these risks are relevant, as mentioned by P5 (product manager) and P7 (data scientist).

The advantages mentioned by stakeholders are rather hypothetical than confirmed. Stakeholders are informed about the advantages and expect the advantages to be realized in practice, yet not all advantages have been experienced and confirmed in practice by the stakeholders. However, the advantages can be seen as deutero learning outcomes from the stakeholders about how the prioritization method should work and thus as norms for the DDDM tool to function in practice.

The main disadvantage that emerges is that the prioritization method is unable to provide context to companies and buildings as experts do. P5 (product manager) describes the assessment of whether a company or building should be re-inspected as a human activity that cannot be easily translated into data: *"It is the knowledge of experts, things they have picked up on the phone with a specific client, developments occurring in a particular area, information that is not reflected in the data. Without expert knowledge, you will miss things one way or another."* P4 (manager of the underwriting department) adds that experts can weigh in a societal context, something that, according to him, does not directly come across in the prioritization method: *"For example, if we see that we have arson and we notice that we are in a period of economic downturn and high unemployment, experts can anticipate that this will affect the damage burden. The question is whether this can be extracted from data."* P6 (data scientist) also mentions that the prioritization method does assess risks but does not distinguish between acceptable and non-acceptable risks. This means that risk experts are sent to companies that have a significant but acceptable risk, and as a result, risk experts can do little with such re-inspections. It can be inferred that, although the DDDM tool aligns with predetermined norms and rules —specifically, re-inspection capacity should be spent on buildings that are most likely to have high damage burden — its practical performance is suboptimal, leading to new learning norms and an adjustment in the human mental model: the prioritization method should be able to distinguish between risks. The realization of this changed norm in the prioritization method is something that the usefulness predictive model should do. In that way, the addition of the usefulness predictive model leads to a change in the norms in the machine mental model triggered by a change in the norms of the human mental model.

In addition, the prioritization method processes any developments that can be extracted from data with a delay, resulting in the prioritization of companies that may no longer require attention or unprioritized companies that need high prioritization. Training a predictive model on developments requires a significant history of these developments to be included in the data, and that takes time. The prioritization method experiences a delay before adapting to new trends. P7 (data scientist) provides an example: *"Suppose many bakers have suffered damage in recent years, but there is an innovation that makes bread ovens safer. Then bakers will still be at the top of the priority list next year because historically they have a high risk of damage. [...] On the other hand, the prioritization method misses business activities where damages begin to occur, for example, because they are new. In ten years, there has been a significant increase, so to speak, in the number of charging stations and therefore also an increase in the burden of damage to charging stations. It will take some time before*

this kind of new damage surfaces. You first need to build a history in the data. But that is exactly what you don't want because you want to be able to predict it in the future and on forehand."

Several stakeholders have highlighted that there is a lack of transparency: risk experts are tasked with re-inspecting companies based on a prioritization method without a clear understanding of why a specific company is chosen. This harms the internalization of the DDDM tool. When risk experts are uncertain about the reasons to re-inspect, it hinders their ability to form a well-informed opinion on the DDDM tool's functioning – what aspects work well and what aspects need improvement. The absence of a well-informed opinion results in limited symbiotic learning outcomes, hindering the potential for triple-loop learning from the DDDM tool among risk experts. To address this, it is crucial to enhance transparency in the prioritization method, necessitating the use of XAI. Although transparency is good for the internalization and socialization of the DDDM tool, transparency must be handled with care as too much transparency can also lead to an unfair decline in confidence in the prioritization method in the event of incidental setbacks.

Another aspect that comes to light is that the quality of data is limited in some cases. Several examples were given, such as P7 (data scientist) sharing, *"The first time we ran the prioritization method, a customer with hundreds of claims in one year was at the top. I thought this must be a large building, we should go there. Until I looked at the accompanying policy. It turned out to be a customer who had insured many homes that were all listed under one policy, causing that policy to be at the top of our prioritization method. However, the risk expert cannot re-inspect all homes during one re-inspection, and re-inspecting a single home is far from useful"* The lack of data standards for policy data led to differences in how data is registered within a policy, influencing the quality of the DDDM tool. The underwriter adds that in the event of a damage report, the cause of the damage is not always accurately and specifically indicated, which also affects the quality of the prioritization method. Different stakeholders mention that the quality of data will increase in the future due to the so-called project SKB+. With that project, the policy ontology, currently uneven due to policies originating from different brands, is being aligned, and the semantics of data fields are being revised and enhanced. More details about this project can be found in Appendix E.

By using the prioritization method, stakeholders have gained an idea of how it should function in the future. Several stakeholders indicate that people should always be able to make adjustments to the prioritization. Once the prioritization method has produced a list of buildings to re-inspect, the risk expertise department and the underwriting department must be able to determine together whether a re-inspection makes sense for a building. This corresponds to the concept of human-in-the-loop. According to the stakeholders, insight into the operation of the prioritization method and therefore transparency of the prioritization method and the results is crucial to make well-informed choices about whether adjustments should be made to the re-inspection list.

According to stakeholders, a re-inspection usefulness predictive model should integrate soft factors and context into the prioritization method to increase the effectiveness of re-inspections. P5 (product manager) emphasizes the value of including data entered by risk experts and distinguishing between acceptable and unacceptable risks, to enable the model to assess the impact of risk experts. P7 (data scientist) provides an example of this: *"Suppose bakers always have a lot of damage. At the moment, they come to the top of the list every year. But if a risk expert goes there and consistently indicates that it is pointless and that the high priority for bakers does not make sense. The usefulness score can indicate that the impact of risk experts on the risk is not significant, even though the bakers are at the top based on damage risk. This way, a prediction of usefulness can correct for the impact a risk expert can make."* P2 (manager of the risk expertise department) adds that customers who have been re-inspected and approved previously do not need to be reselected in the prioritization. Those learning

outcomes emphasize the necessity of incorporating a prediction of re-inspection usefulness to introduce soft factors and a feedback loop into the prioritization method. This inclusion facilitates a human-machine interaction, where risk experts influence the model by assigning a usefulness score as an evaluation of a re-inspection. Moreover, an automatic learning mechanism for the DDDM tool is created, enhancing its ability to generate improved alternatives by learning from user input.

Conclusion
The stakeholders that are interviewed have gained experiences with the DDDM tool by internalizing and socializing. Although internalization was limited due to insufficient transparency, stakeholders did achieve new learning outcomes through the process of internalization and socialization. An important learning outcome is that the current prioritization method is not sufficiently capable of distinguishing between acceptable and non-acceptable risks and of the possible reduction of a risk by a re-inspection. Adding a prediction of the usefulness of a re-inspection to the prioritization method can enhance this context. This learning outcome serves as a basis for further development of the prioritization method and as the starting point for the development of a usefulness predictive model.

4.2 USEFULNESS OF A RE-INSPECTION

Usefulness, commonly perceived as the extent to which something improves job performance or facilitates task completion (Ma & Liu, 1986), encompasses the functionality needed to fulfill work domain objectives, rather than prioritizing ease of use alone (Burns, Vicente, Christoffersen, & Pawlak, 1997). Usefulness is also defined as the extent to which the content and services offered meet user requirements (Buchanan & Salako, 2009). In summary, usefulness denotes the degree to which something is helpful, valuable, or advantageous for achieving a specific purpose. Defining the usefulness of a re-inspection requires clarity on its purpose and the contributing factors to it before developing a predictive model.

Stakeholders unanimously define the purpose of re-inspections as preventing damage burden. Risk experts achieve this by advising customers on damage reduction or by setting prevention requirements recorded as claims in clauses. The overarching goal is to minimize damage across the entire insurance portfolio and not necessarily at company level. However, stakeholders identify additional benefits that can be seen as a goal. Re-inspections offer personalized support, enhancing customer satisfaction and leading to a stronger customer relationship. Re-inspections also contribute to the acquisition of new tacit knowledge for risk experts by providing risk experts with valuable insights into current risk trends, contributing to the improvement of their expertise.

A re-inspection is initially considered useful if it can reduce the damage burden, however, it is uncertain if the damage can be prevented, as damage depends on chance. Nevertheless, risk reduction helps in preventing damage. Risk experts preferably focus on the risks over which they can exert the most positive influence. They primarily exert influence by mitigating the impact of risks because the chance can often be reduced to a limited extent. This impact is measured by the estimated maximum loss (EML), representing the estimated extent of possible damage under normal circumstances. Companies with a higher EML are more likely to see reductions. The metric is only known for previously inspected buildings. In addition, insured interest indirectly indicates the potential extent of damage, with lower interest linked to less potential reduction. However, it does not directly reveal a potential reduction in the risk of damage, as it also depends on building type; an equal insured interest may mean a different risk of damage for an office than for a bakery. Stakeholders therefore emphasize considering insured interest in combination with the sector. Both insured interest and sector are recorded in policies and are available for all insured buildings. A factor that also influences the risk of damage are ABC risk scores, which are categorical scores used by risk experts to indicate the risk severity. According to

stakeholders, the risk scores with the highest influence are the risk scores for fire and burglary. Risk scores are only known for buildings that have been inspected previously. Furthermore, historical damage data is mentioned as a factor for the risk of damage when taken into account in combination with the business sector. Stakeholders suggest that if a lot of damage occurs within certain sectors, there is a good chance that significant damage reduction is possible within those sectors. However, using historical damage data at the building level may be questionable, as recent building damages may lead to building replacements meeting high safety standards.

What, according to the stakeholders, influences the reduction of risk of damage and thus contributes to the usefulness of a re-inspection is creating awareness among the customer. Risk experts attempt to create awareness with customers through advising about potential damages and the impact of those damages on the customer's business operations to motivate the customer to implement advice. When awareness is present, the chances increase that the customer will implement advice or preventive measures. Therefore, awareness plays a crucial role in the usefulness of a re-inspection. Risk experts use expert judgment to assess a customer's awareness of specific risks and record this as an ABC risk score for each inspection as prevention awareness. Prevention awareness is only recorded for previous inspections. Prevention awareness is also represented in the management score because a good management score indicates strong prevention awareness within the management.

Noticing developments at companies also contributes to the usefulness of re-inspections. Developments such as unexpected expansions in production activities or structural changes in business processes can pose risks that were not yet known and thus increase the damage burden. Noticing these developments enables risk experts to identify and anticipate these new risks to manage the risk of damage. Stakeholders indicate that factors affecting the relevance of re-inspections are mainly determined by changes in business activities or risk scores. The changes are recorded as a Boolean value under the data field "change in policy". The problem, however, is that these changes in business activity or risk scores are not known in advance. The developments can only be observed after a re-inspection has been carried out. In practice, developments are closely monitored by, among others, the underwriting department and the risk expertise department, and changes in business activity or risks are somewhat known before a re-inspection is carried out. Experts are aware of developments but do not record these developments yet in the form of data.

Additionally, it is useful if a re-inspection is used to check whether previously proposed preventive measures have been implemented. This is necessary to monitor whether there are improvements in risks. If it turns out that risk improvements have not been implemented, and as a result, the risk is no longer considered acceptable, it may be necessary to establish preventive conditions with clauses, increase the premium, or even terminate the policy. However, stakeholders indicate that this approach feels less useful than advising, encouraging, and motivating the customer. If preventive measures are required, these will be included as a clause in the policy. According to stakeholders, the presence of these prevention requirement clauses is therefore an attribute that can contribute to the usefulness of re-inspections.

Conclusion

Using the described definition of and contributing factors to the usefulness, a representation of the human mental model is made about the usefulness of a re-inspection using a diagram that indicates the factors that have a relation to the usefulness of a re-inspection, which can be found in Figure 7. The figure illustrates the factors, indicating the available data that represent each factor and specifying whether the data is accessible for all policies or exclusively for those that have undergone re-inspection. This overview is the basis for the selection of features. A distinction is made between factors that contribute to the usefulness of a re-inspection, which are formative indicators, and factors

that reflect the usefulness of a re-inspection, which are reflective indicators (Appelman & Sundar, 2016; Kline, 2016). Based on this mental model and the corresponding data representations, features can be selected for the re-inspection usefulness predictive model. Now that the human mental model has been made explicit, the machine learning phase following Figure 5 can start.

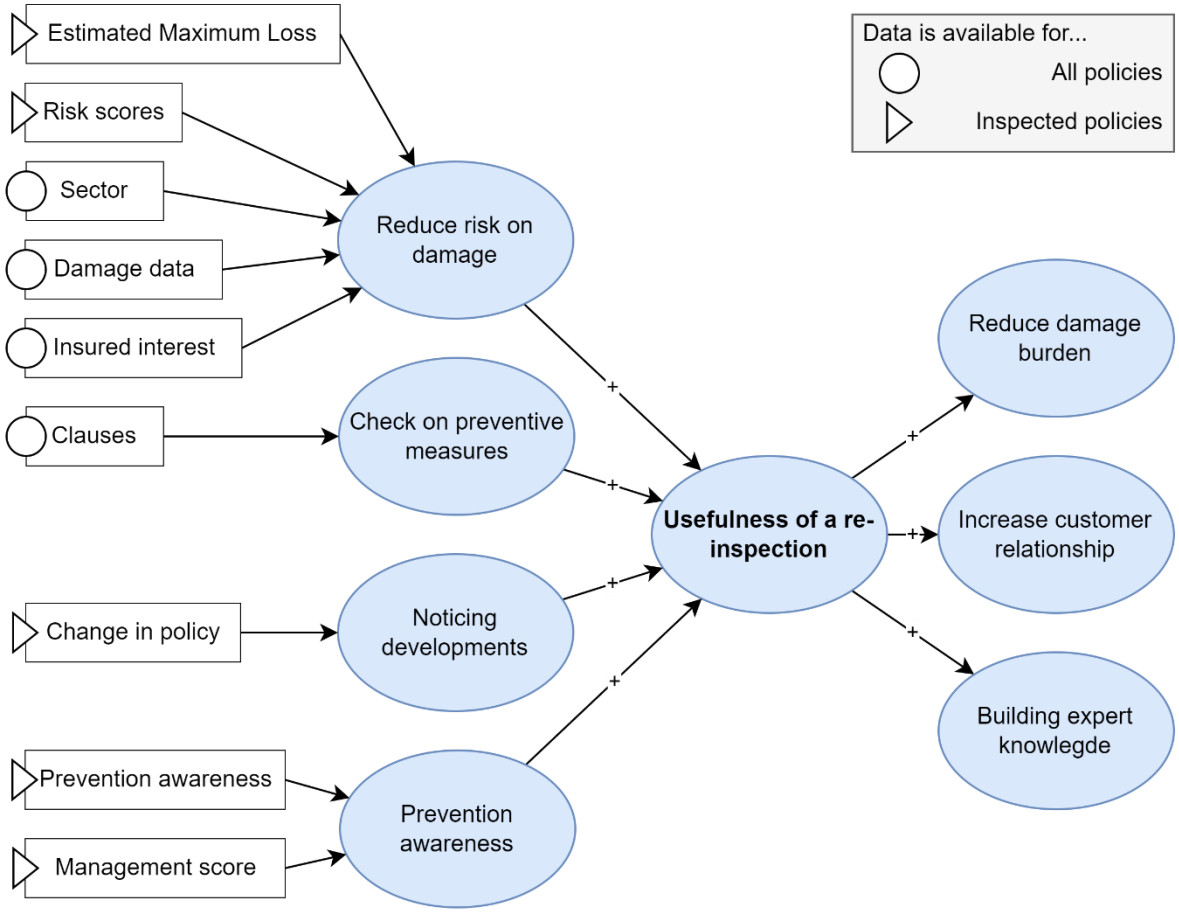

Figure 7: Representation of human mental model about factors that have a relation to the re-inspection usefulness. The blue oval boxes represent factors that have a relation to the usefulness of a re-inspection, and the white rectangular boxes are data representations of specific factors.

5 MACHINE LEARNING

In this chapter, the data preprocessing of the dataset is addressed and the results of the development of the usefulness predictive model and the simulation of it on the usefulness of re-inspections are presented. The overarching goal of this chapter is to execute machine learning, as depicted in Figure 5. Part of machine learning is data exploration, collection, and preparation, necessary to arrive at a dataset to develop a usefulness predictive model. The data exploration and preparation can be found in Appendix E. The data exploration and preparation shows that a connection must be made between policy and inspection data based on the zip code to create a usable data set. This reduces the reliability of the dataset and the number of observations.

5.1 DATA PREPROCESSING

A predictive model for re-inspection usefulness can be applied at various scales over the data, and multiple datasets can be used to train. First, the scale on which a prediction is performed varies. Predictions can be conducted on a dataset limited to data points where a prior inspection has taken place and inspection data is available (constituting a segment of the complete portfolio), and on a dataset covering the entire portfolio, including data points where no inspection data is available. Second, the datasets differ in the type of data used for feature selection, involving policy data, inspection data, or a combination of both. This leads to the creation of four datasets with which a predictive model can be developed:

1. The first dataset includes only the features originating from inspections, and a usefulness prediction can be made exclusively for companies that have undergone inspections. The sample size of this dataset is 914.
2. The second dataset includes features from both policy and inspection data, and a usefulness prediction can be made exclusively for companies that have undergone inspections. The sample size of this dataset is 484. The reduced sample size in the second dataset compared to the first dataset is due to the barrier in merging policy- and inspection data, primarily caused by the absence of policy numbers in the inspection dataset.
3. In the third dataset, only features from the policy are considered, and a usefulness prediction can be made for all companies, including companies with buildings that have never undergone inspections. The sample size of this dataset is 1049.
4. The fourth dataset involves features from both policy and inspection data, and a usefulness prediction can be made for all companies, including companies with buildings that have never undergone inspections. The sample size of this dataset is 1049. However, it is important to note that missing values in the inspection features may be encountered because not every policy is linked to inspection data.

The selection of features is guided by factors influencing the effectiveness of re-inspections, identified through interviews and depicted in Figure 7. Based on these factors, a data exploration for features has taken place, with details provided in Appendix E. The features found and their relationship to the factors in Figure 7 are presented in Table 3. The table makes clear that one or more features have been identified for all formative factors for the usefulness of re-inspections in Figure 7. However, two features, the previously judged usefulness score and time between inspections, were not mentioned by stakeholders as factors contributing to usefulness. During the data exploration, these factors were found and considered to influence the usefulness score, leading to their inclusion in the selection of features. The table therefore shows the relation between the machine mental model and the human mental model.

Table 3: Features and their relation to the data factors influencing the usefulness in Figure 7

Feature (machine mental model)	**Relation with factor in Figure 7 (human mental model)**
Previous judged usefulness score	-
Previous EML (estimated maximum loss)	Estimated Maximum Loss
Time between inspections	-
Previous prevention awareness	Prevention awareness
Insured amount	Insured interest
Previous judged management score	Management score
Damage burden last year	Damage data
Sector	Sector
Presence of prevention clause in past year	Clauses
Previous calculated fire risk	Risk scores
Previous calculated management score	Management score
Previous judged burglary risk	Risk scores
Presence of prevention clause in past five years	Clauses
Damage burden last five years	Damage data
Previous change in policy activity	Change in policy
Number of damages last five years	Damage data
Number of damages last year	Damage data

Details about the chosen features, the extent to which they have a value entered in the dataset, their respective Pearson correlation coefficient, and their presence in various datasets as elaborated in this section are presented in Table 4. Calculations on the correlation and completeness are based on the second dataset described in this part as this dataset includes features from both policy and inspection data and is therefore the most suitable dataset to calculate the correlation to the usefulness of all features. The features that start with *Previous* are features from the inspection carried out before the re-inspection. Further distinctions are made between a judged risk score, influenced by expert judgment, and a calculated risk score, derived from technical information and expert evaluation of corresponding risks.

Table 4: Features with corresponding correlation to the usefulness and their presence in the different datasets

Feature (machine mental model)				**In dataset**			
Name	**Type**	**Completeness**	**Correlation**	**1**	**2**	**3**	**4**
Previous judged usefulness score	Numeric	15%	0.4321	X	X		X
Previous EML (estimated maximum loss)	Numeric	97%	0.1917	X	X		X
Time between inspections	Numeric	100%	0.1103	X	X		X
Previous prevention awareness	Categorical	98%	0.0924	X	X		X
Insured amount	Numeric	100%	0.0890		X	X	X
Previous judged management score	Categorical	97%	0.0794	X	X		X
Damage burden last year	Numeric	100%	0.0758		X	X	X
Sector	Categorical	98%	0.0446		X	X	X
Presence of prevention clause in past year	Categorical	100%	0.0384		X	X	X
Previous calculated fire risk	Categorical	95%	0.0379	X	X		X
Previous calculated management score	Categorical	84%	0.0278	X	X		X
Previous judged burglary risk	Categorical	97%	0.0236	X	X		X
Presence of prevention clause in past five years	Categorical	100%	0.0200		X	X	X
Damage burden last five years	Numeric	100%	0.0197		X	X	X
Previous change in policy activity	Categorical	97%	0.0152	X	X		X
Number of damages last five years	Numeric	100%	0.0077		X	X	X
Number of damages last year	Numeric	100%	0.0016		X	X	X

Correlation assumes linear relationships, which have not been proven for the variables in question. Consequently, caution should be given when drawing conclusions based on correlation, particularly for categorical features. The categorical features have been converted into numerical values temporarily to facilitate correlation testing. This makes more sense for some categorical variables than for others. For risk and management scores, this conversion is logical because each category represents the extent to which a company meets the score. The earlier change in policy activity and the presence of prevention clauses can also be logically converted, as they represent a binary (True-False) value. However, for the sector variable, it is less logical to convert categorical values into numerical values. The categories have been transformed into numerical values to facilitate correlation testing.

The preprocessing of data for the predictive model also involves the normalization of numeric input features and the encoding of categorical features. Numeric features undergo normalization through z-score normalization, resulting in these features being scaled to have a mean of 0 and a standard deviation of 1 (Kappal, 2019). Normalization prevents larger-scaled features from disproportionately influencing the learning process. Simultaneously, categorical input features undergo one-hot encoding, a method essential for representing categorical variables in a format suitable for machine learning algorithms. One-hot encoding transforms each unique category within a categorical feature into a binary column, with '1' indicating the presence of a specific category and '0' indicating its absence (Yu, Zhou, Chen, & Lai, 2022). This process allows a predictive model to interpret categorical information and handle missing data. These preprocessing steps collectively contribute to a prepared input dataset for the training of a predictive model.

Risk experts input the usefulness score via a slider interface, although they see five categories with equal intervals instead of a numerical score. Nonetheless, they can specify the extent to which the usefulness falls within one of these categories, which makes the usefulness score a numeric value between 0 and 100 that is transferable to a categorical value. A score from 0 to 20 is "Crucial", a score from 21 to 40 is "Very useful", a score from 41 to 60 is "Useful", a score from 61 to 80 is "Unnecessary", and a score from 81 to 100 is "Useless". By transferring the numeric usefulness score to categories, it is possible to use both regression and classification algorithms.

The distribution of the numerical usefulness score in the dataset can be seen in Figure 8 and the distribution of the categorical usefulness score in the dataset can be seen in Figure 9. The fourth dataset, as outlined, serves as the basis for visualizing this distribution. However, it is noteworthy that all datasets have a comparable distribution of usefulness. The distribution shows that the dataset that will be used for the usefulness predictive model is imbalanced: a moderately useful re-inspection occurs significantly more frequently than a crucial or useless re-inspection. This imbalance poses a challenge for training a predictive model that accurately distinguishes between various levels of usefulness. To allow a predictive model to distinguish between different categories or scores of usefulness and not predict everything as averagely useful, a balanced dataset is needed. To address the imbalanced nature of the training dataset, the Synthetic Minority Over-sampling Technique (SMOTE) can be employed. SMOTE works by generating synthetic instances of the minority class to augment the training dataset, thereby balancing the class distribution (Fernandez, Garcia, Herrera, & Chawla, 2018). SMOTE enhances the predictive model's ability to generalize to minority class patterns, ultimately improving the overall performance of the machine learning model in handling imbalanced datasets. By applying SMOTE, the class containing the majority of samples, which is the 'useful' class, remains constant, while classes with fewer samples are filled with synthetic data until they reach the same sample size. The impact of applying SMOTE on the usefulness score is shown in Figure 10.

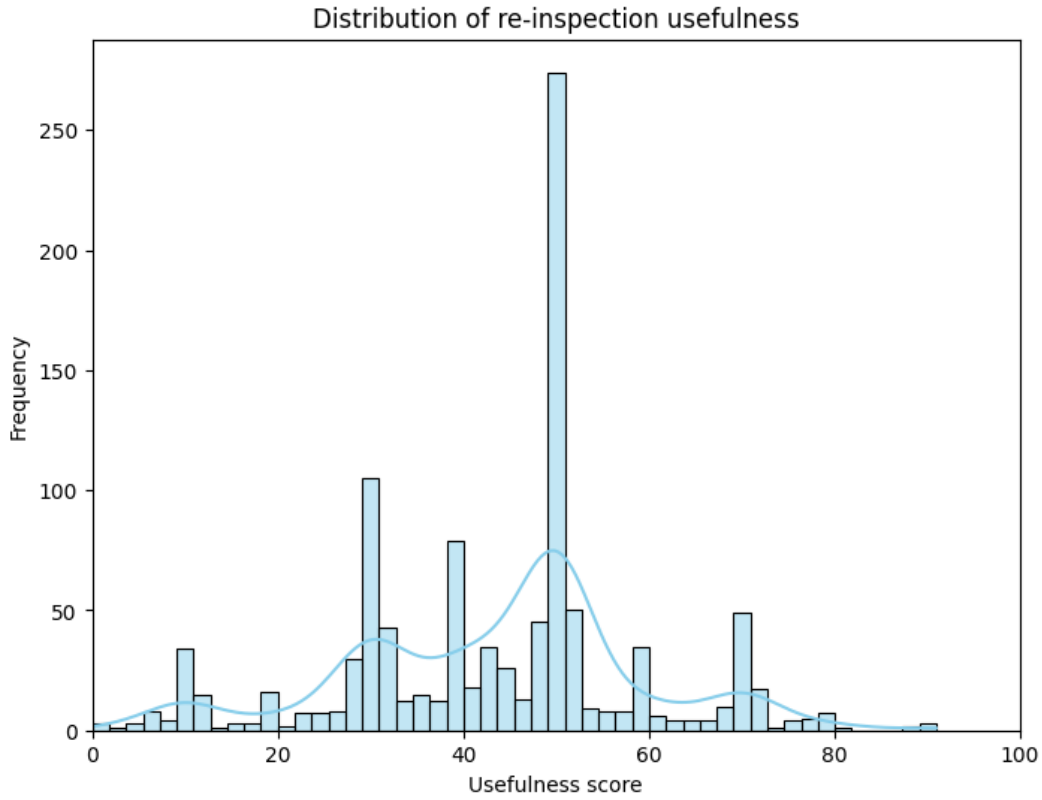

Figure 8: Distribution of the numerical re-inspection usefulness score

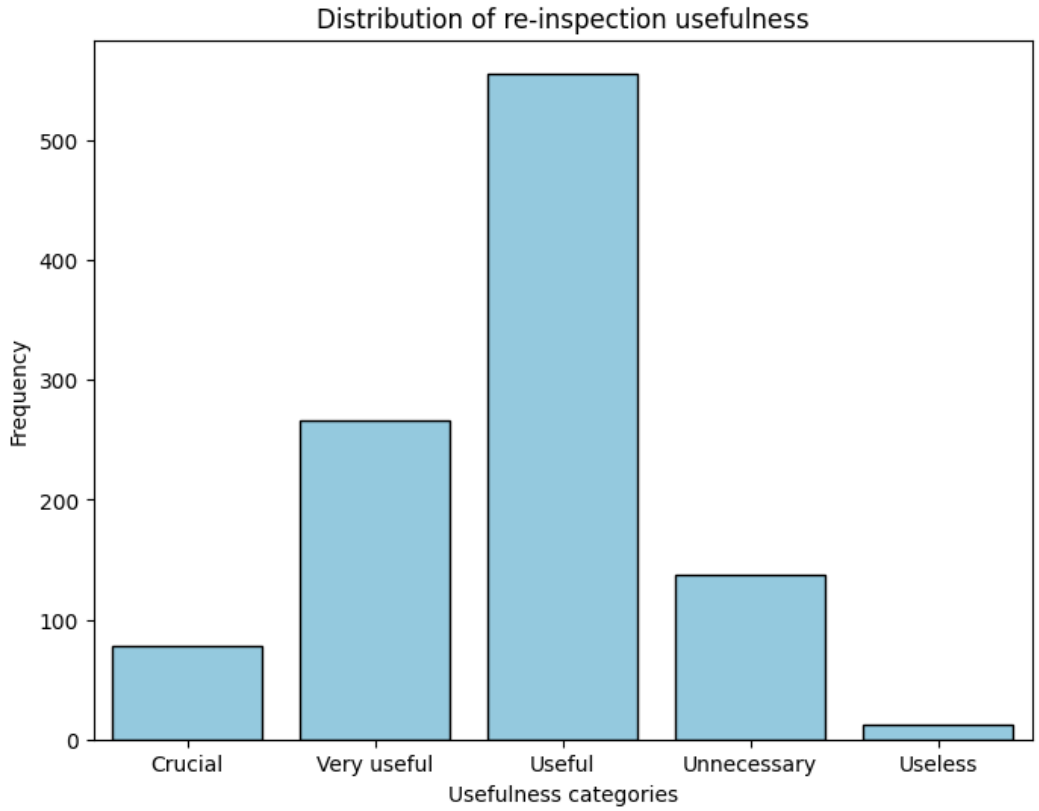

Figure 9: Distribution of the categorical re-inspection usefulness score

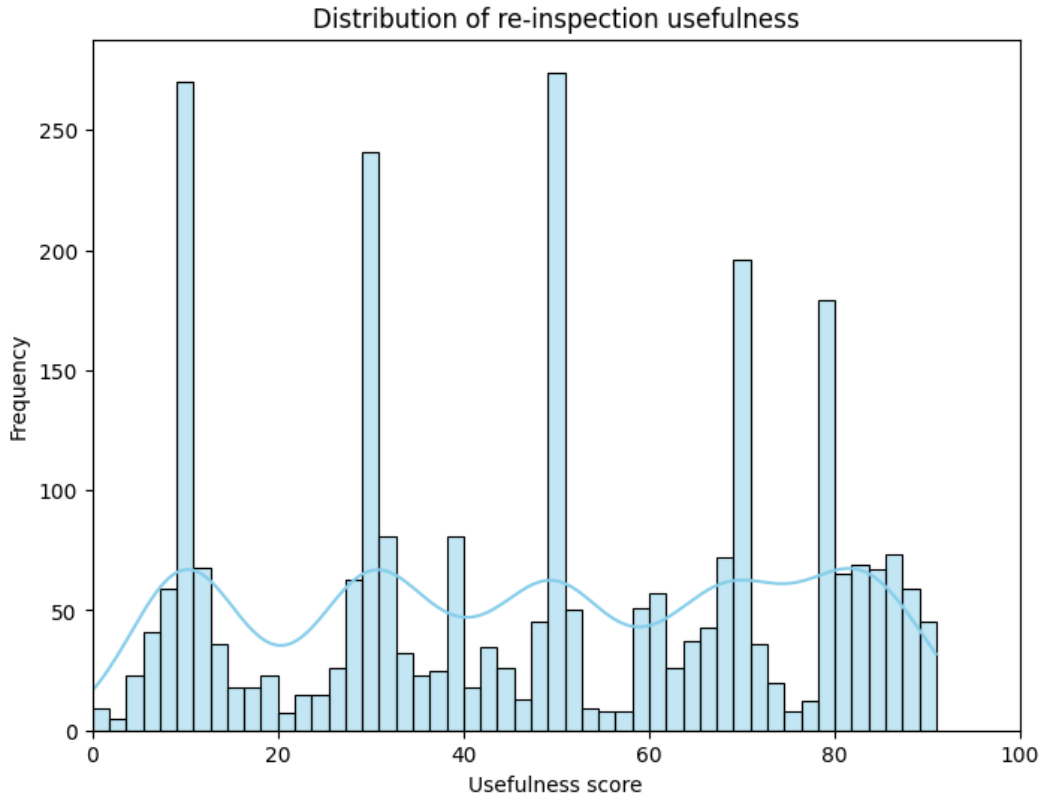

Figure 10: Distribution of the numerical re-inspection usefulness score after using SMOTE to balance the dataset

5.2 Model development results

Re-inspection usefulness predictive models have been trained using various algorithms and the four datasets as described in section 5.1. The utilized algorithms include multiple regression, random forest regressor, random forest classifier, neural network regressor, and neural network classifier. The model has been trained on the five categories as described in section 5.1 and on three categories, where the categories “Crucial” and “Very useful” together form the category “Very useful”, the categories “Unnecessary” and “Useless” together form the category “Unnecessary”, and the category “useful” remains the same. All features described in section 5.1 were incorporated into the training of the models.

The performance of the trained predictive models is visualized using a heatmap, which can be found in Table 5. The heatmap shows the relevant performance metrics for each unique combination of algorithm, dataset, and number of categories. The heatmap also indicates the algorithm ‘No algorithm’, meaning that an average has been taken to predict usefulness, allowing for a comparison of how well the algorithms perform about a baseline performance. A complete overview detailing the mean and the standard error of each performance metric for every trained model, derived from the cross-validation process, is provided in Appendix F.

In terms of accuracy for classification algorithms and MSE for regression algorithms, none of the algorithms surpasses the baseline of taking the average. However, when it comes to the ma-F1 score, the algorithms demonstrate considerable improvement. Algorithms using three categories for making a prediction outperform those using five categories in terms of the performance metrics. This outcome is somewhat expected, as the five-category classification lacks sufficient data for the two extreme categories, and it may be harder to distinguish between five categories than to distinguish between three categories. The expectation is that by including more data in the extreme categories, the

performance difference between the three and five categories will diminish. In such a scenario, it may be preferable to opt for a more precise distinction in usefulness based on five categories. It makes sense to select algorithms for further development that make predictions based on three categories.

The differences in performance are less evident when evaluating the combinations of datasets and algorithms within three categories. In certain instances, the standard error in model performance that originates from cross-validation is even greater than the performance differences between combinations of datasets and algorithms, as indicated in Appendix F. The results do not offer a compelling rationale for the selection of a specific combination of a dataset and an algorithm because of the small differences in performance metrics and because the performance metrics of regression algorithms cannot be directly compared with those of classification algorithms. To have diverse types of usefulness predictive models to test in the simulation and to allow the models to be internalized by stakeholders, it is decided to further develop three usefulness predictive models based on three different algorithms and three different datasets to have usefulness predictive model alternatives that exhibit notable differences.

Table 5: The performance of the trained usefulness predictive models using the different algorithms, different category configurations, and different datasets. The colors of the heatmap show relative differences between the results within a given performance metric.

	Algorithm	Performance metric	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Three categories	No algorithm	MSE	259.8	248.3	250.4	250.4
		Accuracy	0.54	0.54	0.53	0.53
		ma-F1	0.23	0.23	0.23	0.23
	Multiple regression	MSE	278.2	316.5	268.0	276.9
		ma-F1	0.37	0.37	0.37	0.37
	Random forest regressor	MSE	267.2	256.5	284.7	273.1
		ma-F1	0.36	0.31	0.37	0.35
	Random forest classifier	Accuracy	0.47	0.48	0.46	0.47
		ma-F1	0.36	0.35	0.38	0.36
	Neural network regressor	MSE	273.3	464.9	379.8	386.3
		ma-F1	0.34	0.36	0.38	0.36
	Neural network classifier	Accuracy	0.37	0.42	0.40	0.43
		ma-F1	0.32	0.34	0.37	0.39
Five categories	No algorithm	MSE	259.8	248.3	250.4	250.4
		Accuracy	0.54	0.54	0.53	0.53
		ma-F1	0.14	0.14	0.14	0.14
	Multiple regression	MSE	385.0	327.2	432.9	562.4
		ma-F1	0.23	0.22	0.21	0.23
	Random forest regressor	MSE	276.3	275.6	284.0	273.1
		ma-F1	0.20	0.19	0.20	0.21
	Random forest classifier	Accuracy	0.46	0.44	0.42	0.46
		ma-F1	0.23	0.20	0.21	0.21
	Neural network regressor	MSE	296.9	484.5	335.8	416.9
		ma-F1	0.19	0.22	0.23	0.22
	Neural network classifier	Accuracy	0.30	0.44	0.33	0.38
		ma-F1	0.20	0.27	0.22	0.22

This choice was made by looking at algorithms that have a relatively high accuracy or MSE in combination with a relatively high ma-F1 score. The ma-F1 score is emphasized because only evaluating based on accuracy or MSE increases the likelihood of the predictive model being unable to differentiate between classes, especially since the moderately useful class is much more frequent in the data. Because the performance of classifiers cannot be compared with the performance of regressors at this stage, it was decided that at least one model of both types of algorithms should be selected. The three chosen alternatives are multiple regression for dataset 3, random forest regressor for dataset 1, and random forest classifier for dataset 2, all for three categories.

The models chosen are optimized using cross-validation within the training set. Within the cross-validation, the model that produces the best predictions according to the validation set is selected as the model, assuming that this combination of training and validation data is best for recognizing patterns. The performance of the selected model is then tested with a test dataset, which is held apart and thus unseen, providing an unbiased view of the performance. A confusion matrix is presented for the three optimized models, illustrating the distribution of predicted usefulness classes relative to the actual usefulness classes. This matrix includes detailed calculations for recall, precision, and F1 score per usefulness class, as well as accuracy and ma-F1 score for the entire dataset. Additionally, for regression algorithms, a scatter plot is included to visually assess the alignment of predicted values with actual values.

The performance results of the usefulness predictive model using multiple regression for dataset 3 can be found in Table 6 and Figure 11. The MSE of the multiple regression model is calculated as 246.63, indicating a marginal improvement compared to an MSE of 250.37 obtained when taking the average of this dataset as a prediction according to Table 5. Despite this modest improvement, the model demonstrates the ability to differentiate between various usefulness scores. As illustrated in Figure 11, data points are distributed along the diagonal line with a notable spread, signifying the model's ability to distinguish between usefulness values. This observation is consistent with the findings in Table 6, where multiple regression achieves a ma-F1 score of 0.40 and an accuracy of 0.50. Multiple regression distinguishes between useful and very useful categories with a considerable margin of error but faces challenges in differentiating the unnecessary re-inspections, frequently predicting instances from this class as useful. This is also reflected in the F1 score of the corresponding classes. The disability to differentiate unnecessary re-inspections may result from insufficient data in that category to learn the associated patterns.

Table 7 illustrates the confusion matrix for the optimized random forest classifier applied to dataset 2. With an accuracy of 0.49 and a ma-F1 score of 0.40, the performance of the random forest classifier on dataset 2 is comparable to multiple regression on dataset 3. In contrast to multiple regression, the random forest classifier exhibits lower capability in predicting very useful re-inspections, thereby reducing the F1 score for this category. However, it demonstrates moderate improvement in distinguishing unnecessary re-inspections, resulting in an increased F1 score for this class, making the ma-F1 score comparable to the optimized multiple regression model.

Table 8 presents the confusion matrix and Figure 12 displays the scatter plot of the optimized random forest regressor on dataset 1. The MSE for the optimized random forest regressor on dataset 1 is 242.22, indicating a modest improvement compared to an MSE of 259.75 obtained using the average usefulness of dataset 1 as a prediction. In Figure 12, the model demonstrates some capability in aligning predicted values with actual values, although it has a notable margin of error. Notably, the test dataset lacks observations with a usefulness score above 80, potentially influencing the MSE positively. The confusion matrix in Table 8 highlights the model's ability to distinguish the classes very

useful and useful, but it faces challenges in distinguishing the class unnecessary. As a result, the model achieves an average ma-F1 score of 0.42 and an average accuracy of 0.54.

Table 6: Confusion Matrix for the optimized multiple regression model in combination with dataset 3. Rows correspond to actual classes, columns represent predicted classes. The matrix displays precision, recall, and F1-score for each class, along with overall accuracy and the ma-F1-score.

		Predicted				
		Very useful	Useful	Unneccessary	Recall	F1-score
Actual	Very useful	23	29	4	41%	0.45
	Useful	21	57	6	68%	0.60
	Unnecessary	3	20	3	12%	0.15
	Precision	49%	54%	23%	Accuracy = 0.50	ma-F1 score = 0.40

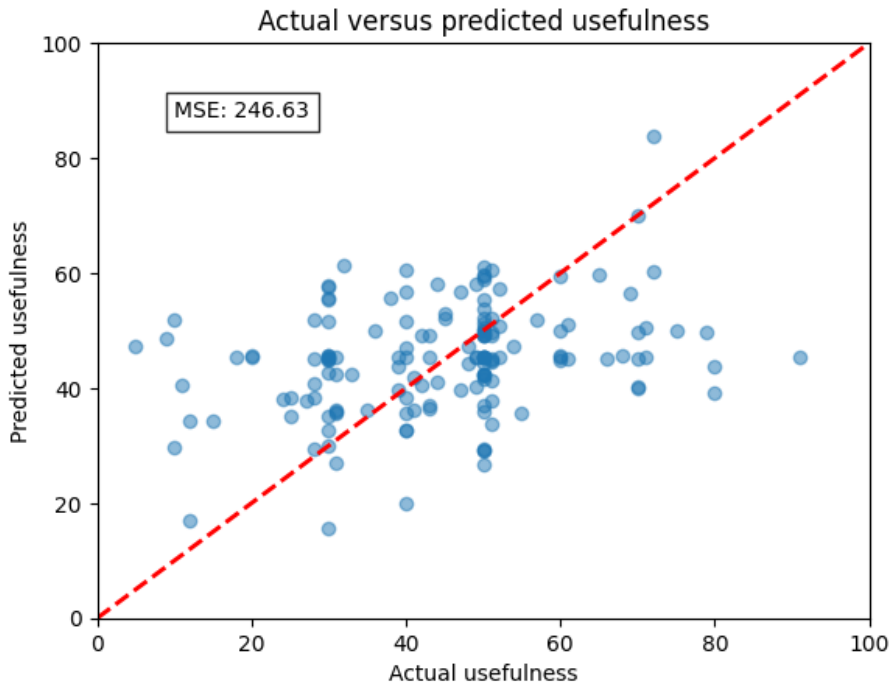

Figure 11: Scatter plot illustrating the comparison between predicted and actual values of the test set using the optimized multiple regression model in combination with dataset 3. Each point represents an observation, with the x-axis indicating the actual values and the y-axis representing the corresponding predicted values. The proximity of points to the red dotted line suggests the accuracy of the model predictions.

Table 7: Confusion Matrix for the optimized random forest classifier model in combination with dataset 2. Rows correspond to actual classes, columns represent predicted classes. The matrix displays precision, recall, and F1-score for each class, along with overall accuracy and the ma-F1-score.

		Predicted				
		Very useful	Useful	Unneccessary	Recall	F1-score
Actual	Very useful	6	22	0	0.21	0.29
	Useful	6	28	5	0.72	0.59
	Unnecessary	1	6	3	0.30	0.33
	Precision	0.46	0.50	0.38	Accuracy = 0.48	ma-F1 score = 0.41

Table 8: Confusion Matrix for the optimized random forest regression model in combination with dataset 1. Rows correspond to actual classes, columns represent predicted classes. The matrix displays precision, recall, and F1-score for each class, along with overall accuracy and the ma-F1-score.

		Predicted				
		Very useful	Useful	Unneccessary	Recall	F1-score
Actual	Very useful	21	32	0	0.40	0.46
	Useful	16	54	0	0.77	0.62
	Unnecessary	1	17	2	0.10	0.18
	Precision	0.55	0.52	1.00	Accuracy = 0.54	ma-F1 score = 0.42

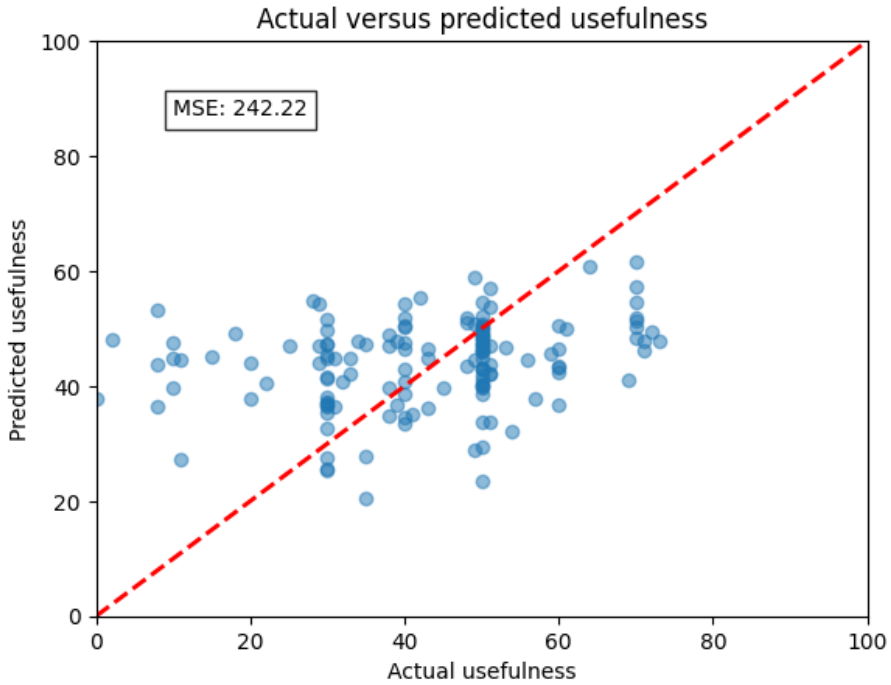

Figure 12: Scatter plot illustrating the comparison between predicted and actual values of the test set using the optimized random forest regression model in combination with dataset 1. Each point represents an observation, with the x-axis indicating the actual values and the y-axis representing the corresponding predicted values. The proximity of points to the red dotted line suggests the accuracy of the model predictions.

5.3 SIMULATION OF USEFULNESS PREDICTION ON PRIORITIZATION METHOD

Now that there are three optimized usefulness predictive models, a simulation can be made of the influence of their incorporation into the prioritization method on the usefulness. The simulation is carried out considering usefulness as a factor that should be integrated into the prioritization calculation and considering the usefulness as a correction factor that should be applied after the companies have been ranked based on damage probability and burden:

- *New prioritization calculation* The usefulness prediction is incorporated into the prioritization method given in Figure 2. For this purpose, formula (6) is used for the prioritization lists of both large enterprises and SMEs. In formula (6), $Ranking_{damage\ probability\ and\ burden}$ is the final ranking for companies according to Figure 2 and $Ranking_{usefulness}$ is the ranking number for companies based on their predicted usefulness. $Ranking_{usefulness}$ for regression usefulness predictions are created by ranking the numerical usefulness predictions. Since the classified usefulness predictions can adopt one of three fixed classes, these classes cannot be ranked. Therefore, $Ranking_{usefulness}$ for classified usefulness predictions is determined by a default ranking value assigned to each possible usefulness class, which are V_{vu} for the class very useful, V_u for the class useful, and V_{un} for the class unnecessary. After applying formula (6) for all companies, a new prioritization has been made for the large enterprises and SMEs separately by ranking the Priority score from high to low.
- *Correction on existing prioritization* The correction on the existing prioritization method is performed by first prioritizing all companies following Figure 2, then applying a correction factor for the usefulness of all prioritized companies by adding the correction factor to the ranking number, and then re-sort the prioritization lists for large enterprises and SME's based on the corrected ranking number. The correction for regression usefulness predictions is the

deviation of the numerically predicted usefulness to the average predicted usefulness multiplied by a usefulness coefficient r. With r, the influence of the usefulness correction on the prioritization can be set. The correction for classified usefulness predictions is made based on certain weights belonging to the given classes, W_{vu} for the class very useful, W_u for the class useful, and W_{un} for the class unnecessary.

$$Priority\ score = \frac{1}{Ranking_{usefulness}} + \frac{1}{Ranking_{damage\ probability\ and\ burden}} \quad (5)$$

Table 9: results of simulation on the change in mean usefulness in the top 50 prioritized buildings after applying the predicted usefulness in the prioritization. The simulation is performed for the different optimized usefulness predictive models and using different methods to integrate the usefulness prediction.

	Method to implement the predicted usefulness		SMEs		Large Enterprises	
			Change in mean usefulness	t-test (p) usefulness	Change in mean usefulness	t-test (p) usefulness
Multiple regression	Calculation, classifier	V_{vu}= 10, V_u= 80, V_{un}= 200	0.88	0.21	1.11	0.54
		V_{vu}= 20, V_u= 100, V_{un}= 200	0.92	0.35	1.04	0.82
		V_{vu}= 50, V_u= 125, V_{un}= 250	0.97	0.74	1.00	0.98
		V_{vu}= 100, V_u= 300, V_{un}= 500	0.99	0.89	1.00	0.98
	Calculation, regressor		0.94	0.53	1.11	0.56
	Correction, classifier	W_{vu}= -50, W_u= 0, W_{un}= 50	0.97	0.74	1.00	0.98
		W_{vu}= -100, W_u= 0, W_{un}= 100	0.97	0.69	1.04	0.82
		W_{vu}= -200, W_u= 0, W_{un}= 200	0.98	0.85	1.08	0.71
		W_{vu}= -300, W_u= 0, W_{un}= 300	0.89	0.24	1.04	0.85
	Correction, regressor	r= 5	0.88	0.16	1.12	0.44
		r= 10	0.74	0.02	1.07	0.67
		r= 20	0.79	0.03	0.98	0.92
Random forest classifier	Calculation, classifier	V_{vu}= 10, V_u= 80, V_{un}= 200	0.99	0.89	1.21	0.28
		V_{vu}= 20, V_u= 100, V_{un}= 200	0.99	0.89	1.18	0.35
		V_{vu}= 50, V_u= 125, V_{un}= 250	0.99	0.91	1.02	0.89
		V_{vu}= 100, V_u= 300, V_{un}= 500	1.00	1.00	1.02	0.91
	Correction, classifier	W_{vu} = -50, W_u= 0, W_{un}= 50	0.99	0.91	1.02	0.89
		W_{vu}= -100, W_u= 0, W_{un}= 100	0.99	0.91	1.18	0.35
		W_{vu}= -200, W_u= 0, W_{un}= 200	0.99	0.91	1.18	0.35
		W_{vu}= -300, W_u= 0, W_{un}= 300	0.99	0.93	1.21	0.28
Random forest regressor	Calculation, classifier	V_{vu}= 10, V_u= 80, V_{un}= 200	1.01	0.91	1.01	0.96
		V_{vu}= 20, V_u= 100, V_{un}= 200	1.01	0.91	1.06	0.76
		V_{vu}= 50, V_u= 125, V_{un}= 250	0.99	0.91	0.98	0.93
		V_{vu}= 100, V_u= 300, V_{un}= 500	1.00	1.00	1.00	1.00
	Calculation, regressor		1.06	0.50	1.02	0.92
	Correction, classifier	W_{vu}= -50, W_u= 0, W_{un}= 50	0.99	0.91	0.98	0.93
		W_{vu}= -100, W_u= 0, W_{un}= 100	0.99	0.91	1.06	0.76
		W_{vu}= -200, W_u= 0, W_{un}= 200	0.99	0.89	1.06	0.76
		W_{vu}= -300, W_u= 0, W_{un}= 300	0.98	0.84	1.03	0.85
	Correction, regressor	r= 5	0.98	0.77	1.11	0.54
		r= 10	0.97	0.70	1.17	0.39
		r= 20	0.98	0.76	0.99	0.97

The results of each simulation can be found in Table 9. The change in average usefulness within the top 50 prioritized re-inspections for both SMEs and large enterprises is calculated for each combination of usefulness predictive model and method used to integrate the predicted usefulness. This calculation has been performed to compare the integration of the usefulness score with the original prioritization. Additionally, a two-sided t-test has been conducted to determine the significance level and assess if the observed changes were statistically significant. In general, the average usefulness score for SMEs decreases with the tested prioritization methods in which the usefulness prediction is integrated. This implies an increase in the usefulness of re-inspections for SMEs, as a higher re-inspection usefulness score corresponds to a lower overall usefulness of a re-inspection. Conversely, the mean usefulness score for large enterprises increases, indicating a decrease in the usefulness of re-inspections for large enterprises. Considering a significance level of 95% for the two-sided t-test, there is a significant difference in the average usefulness score between the simulated top 50 and the top 50 of the current prioritization method for the multiple regression algorithm applied as a correction to the existing prioritization method based on its numerical prediction of usefulness. At $r = 10$, the usefulness changes by a factor of 0.74, and at $r = 20$, the usefulness changes by a factor of 0.79.

5.4 Conclusion machine learning

Based on the human mental model, features are selected to initiate the machine learning phase of Figure 5. Predictive models are trained and evaluated for their performance across various algorithm types and scales of performing the usefulness prediction. Three models are chosen for optimization and simulation to assess how they could enhance the usefulness of the inspections prioritized by the method. The simulation also considers different manners to implement the usefulness score. The simulation results indicate that the multiple regression algorithm, when applied as a correction to the existing prioritization method, significantly improves the usefulness of prioritized re-inspections for SMEs. Thus, with the usefulness of inspections as the norm, this model emerges as the best-performing one.

6 COMBINATION AND INTERNALIZATION OF EXPLICIT KNOWLEDGE

In this chapter, the outcomes of the machine learning are presented to the stakeholders, initiating an evaluation process. The overarching objective of this chapter is to generate explicit knowledge from the machine, combine this knowledge with existing explicit knowledge, and facilitate the internalization of explicit knowledge by individuals, as depicted in Figure 5. The stakeholders have been asked to reflect on the results and findings of this study. This section describes the reflection on the results and findings and covers the factors employed in predicting usefulness, the distribution of usefulness, end-user acceptance, and the interpretability of the prioritization method.

Factors used for making the usefulness prediction
Based on the interviews with stakeholders, in section 4.1, factors were chosen as features to predict usefulness. During the data preprocessing, in section 5.1, it became clear that not every mentioned factor was strongly correlated with the usefulness. Therefore, during the evaluation, stakeholders were asked to assess whether and why the factors are predicting the usefulness. This section describes the most striking findings of the factors used.

The impact of the number of damages in the past year was found to have a low influence on usefulness. According to P5 (product manager) and P2 (manager of the risk expertise department), this makes sense on second thought. P5 explains that fire is not a high-frequency risk, while fire risk is a crucial aspect of a re-inspection. Additionally, he suggests that it may be useful to re-inspect companies where the amount of damages and damage burden have been low in the past year because the damage has not occurred there yet, and it could still happen in the future. This means a change in the human mental model.

The low correlation between the presence of a prevention clause and the usefulness of a re-inspection is caused by the limited impact a clause has on the actual implementation of preventive measures. P3 (underwriter): *"I have the impression that customers might poorly read their policy, let alone a clause stating that a customer must meet preventive measures. If you include preventive measures in a clause, they fade into the background and therefore have limited impact. (...) If you want to seriously work on prevention, you need direct contact with the customer to emphasize the relevance of prevention. If it's only on the policy, it has a limited effect."* The presence of a prevention clause was first seen by people as a possible feature, but upon further consideration, it appears to be an inadequate feature, which is a learning outcome that changes the human mental model.

P1 (risk expert) emphasizes that the usefulness predictive model should not overly rely on the previous usefulness: *"If a risk expert indicated the last time that it was very useful, I can imagine that in the next re-inspection, if everything is resolved, it may not be useful."* A learning outcome suggests that, although there is a high correlation between usefulness and previous usefulness, a usefulness prediction should primarily rely on other information. This changes the mental model of the machine.

P1 generally had higher expectations regarding the influence of risk data. The evaluation with the risk expert revealed that ABC risks are based on a numerical value ranging from 0 to 100, making the numerical value a more precise feature. However, in the development of the predictive model, ABC risks were used as features. Based on this, P1 recommends using the numerical scale for the risk scores instead of the categorical approach. A learning outcome indicates the need for a change in the data used to represent risk scores. This learning outcome influences the machine's mental model.

Compared to the usefulness factors according to the human mental model in Figure 7, changes have taken place in the human mental model after a learning cycle: prevention clauses and damage data

are, on closer inspection, not good predictors of usefulness. Compared to the factors according to the original machine mental model, which are visualized in Table 4, differences have also been found after a learning cycle: the previous usefulness score is not a desirable predictor and the categorical risk scores should be replaced by numerical risk scores. The changes described in the human mental model should also lead to changes in the machine mental model.

Distribution of usefulness

As indicated in the data preprocessing, section 5.1, the distribution of the usefulness score is unequal. The vast majority has been assessed as moderately useful, and as the score becomes more extreme, fewer re-inspections are found in the data. Therefore, stakeholders were asked why the score could be distributed in such a way.

The distribution of usefulness across the re-inspections aligns with the expectations of various stakeholders. Stakeholders express that they expect the distribution to correspond to reality and that they find it reassuring to see this distribution. P4 (manager of the underwriting department) states, *"If we conclude that we send a risk expert on a mission, and it turns out to have not been useful in many cases, then I would be more concerned."* However, P3 (underwriter) notes, *"It is unfortunate that there are so few non-useful re-inspections because these are inspections we would like to eliminate from the prioritization list."* A distribution of usefulness corresponding to reality would represent a learning outcome impacting the machine's mental model: the machine was anticipated to effectively differentiate between the extreme usefulness categories, particularly identifying non-useful re-inspections, which is not possible due to the few available non-useful observations. P3 (underwriter) suggests another possible explanation for the fact that usefulness is often rated as average: *"I think you have to be very confident to say: this is crucial or pointless. I don't think people are wired to assess something extremely on paper or in data, even if they think so."* If that is the case, then a change in the human mental model may be needed to promote the effectiveness of the usefulness predictive model, namely a more extreme assessment of re-inspections.

Different stakeholders point out that subjectivity and the lack of a standardized definition of usefulness impact the distribution of usefulness. P2 (manager of the risk expertise department) emphasizes, *"When you deal with 32 different individuals assessing this, one may find something useful while another does not at all."* P5 (product manager) adds that usefulness depends on the person assessing it, their experience, and the perception with which someone approaches an inspection. P4 (manager of the underwriting department) underscores the influence of an aligned definition of usefulness, stating, *"If we look at the criterion of usefulness, no definition has been provided. It is therefore a personal estimate of how useful a re-inspection is. (...) If you have a set of criteria, you can objectively assess whether it is useful."* A deutero learning outcome here is that standards are needed in the form of assessment criteria for the usefulness of re-inspections to improve the usefulness predictive model's distinctiveness.

P2 (manager of the risk expertise department) describes a difference between the concepts of usefulness and influenceability, explaining that influenceability may potentially be better to consider in prioritization: *"We can also talk about influenceability: to what extent can the risk expert influence the risk during a re-inspection? Imagine arriving at a client and identifying a very high risk, but you can't change anything about it. In that case, the influenceability is zero, but it is still useful. Because if you know you have a high-risk situation on record, underwriting can take that into account. Usefulness does not necessarily mean that influenceability is high, but when we can exert influence on the risk, usefulness will increase significantly. Those are two different concepts. (...) For re-inspections, you may need to look more at influenceability."* A learning outcome of this double-loop learning is that a refinement in the target that has to be predicted might be needed.

Acceptance among the end users

One of the goals of the usefulness prediction is to incorporate the knowledge and experience of a risk expert to enhance the acceptance of and confidence in the prioritization method. To investigate this, stakeholders were asked to explain whether they agree with the statement: *"Adding a usefulness prediction enables a risk expert to leverage their knowledge and experience in future prioritization, increasing their acceptance of the prioritization method."*

P2 (manager of the risk expertise department), P3 (underwriter), and P4 (manager of the underwriting department) expect that in the future, when the usefulness predictive model is accurate enough, trust in the model and acceptance of the model will increase. They explain that the actual usefulness of re-inspections rises, and risk experts will experience that there will be fewer pointless re-inspections suggested by the model. P1 (risk expert) and P5 (product manager) describe that, while a usefulness prediction can increase the acceptance of the prioritization method, there are more dependencies for acceptance. According to P5, explainability and transparency about why a re-inspection is prioritized by the model are important for model acceptance. The explainability and transparency should promote the internalization of the re-inspection method among risk experts. P1 (risk expert) adds that understanding the model is crucial, but communication about it is also essential. If risk experts are limited in their awareness of the usefulness prediction, this is likely not to increase acceptance, even if the average usefulness improves.

A point that emerged during the evaluations is the consideration of how far one wants to emphasize the acceptance of the prioritization method among risk experts. When management indicates that re-inspections must follow a certain policy, risk experts must accept that policy. While emphasizing usefulness might contribute to higher acceptance, the policy will always be decisive. Therefore, the focus in development should not be too strongly directed towards maximizing model acceptance among risk experts. A learning outcome is a weakening of the norm that the re-inspections given by the prioritization methods are not perceived as useful and should be considered as a problem.

Interpretability

From the interviews described in section 4.1, it became clear that the prioritization method provides limited transparency. Therefore, during the evaluation, attention was dedicated to exploring possibilities for interpretation. This was carried out through six visual scenarios within the context of the re-inspection prioritization method, which can be found in Appendix G. Stakeholders were asked about the type of interpretability desired and how they anticipate that providing different levels of transparency would affect the use of the prioritization method in their way of working. The stakeholder's reflections on interpretability are combined to create new explicit knowledge in the form of a shared vision of how interpretability should be provided in the future.

The evaluations have outlined a vision for interpretability in the short and long term. In the short term, stakeholders assert that the scenario in which the ranking is provided in combination with specific predictions will reasonably enhance the interpretability of the prioritization method. This is considered the minimum requirement for decision-makers and end users to gain some insight into the prioritization made. However, P1 (risk expert) wonders how a particular predicted damage burden or damage probability without any further context would be interpreted by end users. "You must explain to risk experts how to interpret and apply the predicted numbers. Or you should qualify the numbers, for example, by categorizing them from low to high." With this, the risk expert aims to prevent end users from taking a prediction too literally.

In the long term, stakeholders express a desire to see an addition to the short-term scenario in the form of predicted reasons for the predictions made and factors that most influence a specific

prediction. According to stakeholders, a predicted reason for why a prediction is made is quickly understandable for both decision-makers and end users. P4 (manager of the underwriting department) emphasizes that this increased interpretability can resolve the current issue of misunderstanding the prioritization list, where decision-makers do not understand why an inspection is on the list, and end users do not know what to focus on during a re-inspection. The stakeholders argue that providing the factors that have the greatest influence on a specific prediction provides a quantitative description to both decision-makers and end users. P5 (product manager) explains: *"I think this scenario illustrates why it is important for a company to be re-inspected. (…) By identifying the most important factors that determine the priority, you provide context for why a company is a high priority".* The predictive models within the prioritization method should therefore incorporate feature relevance techniques to provide the most important factors in a specific prediction.

Conclusion

In this chapter, the results of the machine learning were presented to stakeholders and internalized by them. As a result of internalization, learning lessons are observed, prompting adjustments in both human and machine mental models. Consequently, this section illustrates how the development of DDDM via triple-loop learning, as depicted in Figure 5, generates learning outcomes that influence the human and machine mental model.

7 DISCUSSION

7.1 MAIN FINDINGS

This research addresses the question of how a model for predicting the usefulness of re-inspections can be developed and implemented while enabling triple-loop learning. The implementation of the usefulness predictive model within the prioritization method should lead to the incorporation of risk experts feedback and soft information to acquire more impactful re-inspections, as the current re-inspection prioritization method lacks efficient prioritization and results in diminished confidence among experts. This gives rise to the research question: "*How can a model for predicting the usefulness of a re-inspection be developed by triple-loop learning?*"

To answer this question, the usefulness of re-inspections first had to be defined. A definition is created based on learning outcomes from previous single-loop and double-loop learning: the usefulness of a re-inspection is mainly determined by the possible reduction in the risk of damage. However, it appears to depend on multiple factors, does not have a standardized definition, and is fairly subjective. By using the factors that determine the usefulness of a re-inspection, features were selected to develop the model, and multiple modeling algorithms were developed using different compositions of the dataset. The next step was to implement the usefulness predictive models for re-inspection usefulness within the existing prioritization method to simulate which usefulness predictive model performs best within the organization. The simulation indicates that the multiple regression algorithm, when applied as a correction to the existing prioritization method, significantly improves the usefulness of prioritized re-inspections for SMEs. With the usefulness of inspections as norm, this model emerges as the best-performing one.

This research demonstrated how triple-loop learning via single- or double-loop learning take place and lead to changes in the machine or the human mental model. In this case study, the machine mental model has influenced the human mental model via single-loop learning by revealing that features, such as damage or clause, have limited impact on usefulness. Additionally, the machine's mental model has influenced the human mental model through a revision of norms via double-loop learning, namely that re-inspection influenceability should be considered to predict instead of re-inspection usefulness. The human mental model has led to changes in the machine mental model via double-loop learning by adding the usefulness prediction to make the re-inspections more useful and impactful. The human mental model also changed the machine mental model via single-loop learning, for instance by indicating that the risk features used within the usefulness predictive model have a numerical score that is more accurate than the ABC risk score, allowing for model improvement. The provided examples of single- and double-loop learning, through collaboration between human and machine, can be considered as triple-loop learning. The study also shows how triple-loop learning can be integrated into the machine functionality. The usefulness predictive model supports a human-machine interaction as people can give feedback on the machine outcomes to generate better outcomes in the future. This can therefore be seen as an automated form of triple-loop learning via single-loop learning because the learning loop is built into the functionality of the prioritization method.

To implement the usefulness predictive model within the prioritization method and to improve collaboration and workflow between individuals and the prioritization method, deutero and symbiotic learning are crucial. These learning processes were also observed in this study. For instance, there is a necessity for a standardized definition of usefulness score, which is deutero learning, and for effective communication to ensure that all risk experts possess this standardized definition as part of their tacit

knowledge, which is symbiotic learning. Additionally, there is a need for interpretability of the tool to enable internalization of results, which is also a result of deutero learning.

7.2 PRACTICAL IMPLICATIONS

The implementation of soft information and the creation of a feedback loop lead to more useful re-inspections. A feedback loop is created by implementing the usefulness predictive model: inspections currently selected and inaccurately predicted due to performance issues will be used as training data next year, enabling the usefulness predictive model to improve itself in a targeted manner where errors occur. The feedback loop promotes human-machine interaction and fosters the alignment of human and machine mental models. Increased acceptance among the end-users is expected to be achieved by enabling them to refine the model based on their tacit knowledge and expertise. However, strong communication between decision-makers and end-users is crucial to enhance acceptance. Risk experts need to be aware of their influence on the model outcomes, the positive consequences when they fill in the score correctly, and the potential negative consequences when the score is left incomplete. In other words, symbiotic learning about the feedback loop has yet to occur.

By incorporating re-inspection usefulness via the usefulness predictive model, the prioritization method can take into account feedback from risk experts, and a human-machine interaction is enabled. In that way, the prioritization method can indirectly include the norms, values, and experiences of the human mental model. While the usefulness predictive model currently exhibits limitations in performance due to data quality issues, which may prompt consideration of waiting for additional data before implementing it in the prioritization method, the feedback loop triggers a direct learning process from the usefulness predictive model's incorrect predictions. Therefore, the advice is to implement the usefulness predictive model instead of waiting to gather more data, despite the current limitations in its performance. The implementation of a re-inspection usefulness predictive model also leads to the practical integration of triple-loop learning, where human and machine enhance each other to improve the decision-making process about which companies have to be re-inspected.

7.3 THEORETICAL IMPLICATIONS

There is currently no established methodology for developing a DDDM tool, especially a predictive model, that integrates triple-loop learning. Consequently, DDDM tools often fail to align with human decision-making norms and values. This research combines prior studies on organizational learning in the context of DDDM with established methodologies for constructing predictive models as DDDM tools. It offers a proof-of-concept framework for developing predictive models for DDDM by triple-loop learning, contributing to the literature on the development of DDDM tools, particularly predictive models, through mutual human-machine learning. This research presents a nascent approach to develop DDDM tools by considering mutual learning between people and the DDDM tool, resulting in the tool adopting human decision-making norms and values. The approach used in this research distinguishes from other iterative predictive modeling development methods that place less emphasis on the human learning process and do not consider the dynamic nature of human norms and values. This research fills the gap in scientific knowledge regarding the methodology with which triple-loop learning can be enabled to effectively develop a predictive model as a DDDM tool. The type of contribution is therefore an improvement to the Design Science Research (Gregor & Hevner, 2013). This has theoretical implications that are listed here.

First, this research demonstrates how existing methods can be combined to develop a predictive model as a DDDM tool with triple-loop learning. Human tacit knowledge about the decision-making

topic can be externalized by conducting interviews and can be combined into explicit knowledge by analyzing the interview results and combining them with existing knowledge. Machine learning can then take place based on the explicit knowledge following existing predictive modeling development methodology. Machine learning leads to explicit knowledge, which can be presented to individuals and subsequently reflected upon by them, thereby realizing the human-machine learning loop following the framework in Figure 5.

Second, this research illustrates how DDDM development by triple loop learning results in insight into the conditions that are necessary for learning, or deutero learning outcomes. Moreover, this research shows specific conditions that are generally essential to enable DDDM development by triple-loop learning. A DDDM tool must generate interpretable results. Without interpretable outcomes of a DDDM tool, users cannot internalize the DDDM tool and form an opinion about how the DDDM tool fits with their norms and values. Also, data quality is crucial for triple-loop learning. Limited data quality hinders the development of a DDDM tool, thereby impacting machine learning and, consequently, the triple-loop learning cycle. Limited data quality is a problem that lies in the intersection of human and machine because people generate the data.

Third, DDDM development through triple loop learning proves to be an effective method for finding solutions to issues arising from the interaction between human and machines. In this case study, an example of a problem arising from the interaction between human and machine is the unequal distribution of usefulness. Triple loop learning led to the learning outcome that there is a need for refining, communicating, or even completely revising the definition of the usefulness. The problem cannot be resolved through pure data science knowledge. DDDM development via triple-loop learning does, however, yield solutions to such problems through the iterative learning process between human and machine. This concept extends to numerous domains where challenges that find its roots in the interaction between human and machine, for instance data quality issues, cannot be resolved through conventional data science but can be tackled through a triple loop learning approach.

Finally, this research makes clear how human and mental machines change and why and how the mental models of humans and machines must be aligned. The human and machine mental models undergo transformation as a result of triple loop learning, as evidenced by several examples in this research. A shift in the human mental model is a revision of human norms and values, influenced by insights generated by a machine or external factors. Conversely, adjustments in the machine's mental model primarily occur based on the human mental model, aiming to align with the human mental model. When aligned, the machine acts by the norms, values, rules, and policies it should have according to people. As the mental model of humans will always continue to evolve and change due to external factors (norms and values are dynamic), the organizational learning cycle will have no ending point. It may happen that at a certain point, the human and machine mental models are aligned, but at a later stage, there may be a disparity without any changes happening in the machine mental model. This method takes into account the changing norms and values of humans and should triggers the machine to relearn promptly to align with the human mental model.

7.4 LIMITATIONS AND FUTURE WORK

The current state of the case study is that the re-inspection usefulness predictive model has not been functional in practice. As a result, decision-makers and end-users have not been able to internalize the usefulness predictive model into their tacit knowledge, leading to insufficiencies in measuring the success of a usefulness predictive model developed by triple-loop learning. Within this case, the usefulness predictive model must be implemented for people to gain practical experience with it. Subsequently, future work in this case study is needed to identify new learning lessons representing

the practical experiences of decision-makers and end users with the usefulness predictive model. A further development point could be to improve the interpretability of the results of the prioritization method because until now, triple-loop learning has been limited as people are constrained in learning from the machine due to a lack of transparency and interpretability in the prioritization method. As a result, decision-makers find it challenging to adjust the prioritization, and for the risk experts, who are the end-users, it is unclear why they are directed to a specific company for a re-inspection. This has led to constraints for decision-makers and end-users in forming an opinion about the prioritization method and creating learning outcomes.

A prediction for three usefulness categories performs better than a prediction based on five categories. This is remarkable, considering that the actual usefulness of re-inspection is defined with five categories. Research into the distribution of usefulness revealed that there is very little data available in the extreme categories of usefulness to learn the pattern that forms the extreme usefulness categories as a result of the lack of an aligned definition among all stakeholders, combined with subjectivity, the human tendency not to rate something extremely on paper, and the fact that re-inspections are mostly considered as moderately useful. An aligned definition of the usefulness of re-inspections is needed. Also, intrinsic motivation needs to be created among stakeholders by communicating about how correctly filling in the usefulness score through the feedback loop contributes to more useful re-inspections in the future to improve the accuracy of usefulness predictions.

The prioritization method is built using data from buildings that have undergone re-inspections. If these re-inspections are deemed useful, they receive high scores in the next prioritization, leading to more re-inspections and thus the collection of more data from similar buildings. However, the prioritization method does not have insight into the usefulness of re-inspections for buildings from sectors that are rarely re-inspected, and as a result, these buildings are not included in the prioritization method, even if they could be more beneficial. This creates a biased feedback loop. Predictions based on potentially biased historical data can themselves be biased.

Although the proposed method fills a gap in knowledge about how to effectively develop a predictive model as DDDM tool by triple-loop learning, more evidence should be found to demonstrate the validity of the method. The method has now been validated as a proof of concept based on one case study. To develop a proven design science methodology, the method must be demonstrated in other (case) studies to make it more generalizable (Gregor & Hevner, 2013). Therefore, future work is needed to acquire further scientific evidence to prove the method. Moreover, the method might not only be relevant for the development of predictive models as DDDM tools but also applicable for other kinds of DDDM tools and in other domains of information systems design science. Future work can therefore also be conducted in the broader application of an organizational learning development method for information systems.

8 CONCLUSION

The research question "*How can a model for predicting the usefulness of a re-inspection be developed by triple-loop learning?"* can be answered by declaring that a predictive model can be developed by triple-loop learning using the proposed framework in Figure 5. Learning lessons from people using a DDDM tool can be utilized to develop a predictive model in such a way that the machine operates by the norms, values, rules, and policies as intended by people. In this case study, the multiple regression algorithm that is capable of making predictions across the entire portfolio proves to perform best to the human norms. The research found out that the usefulness predictive model within the existing re-inspection prioritization method itself also leads to triple-loop learning as human feedback on the prioritization method outcomes leads to an automated learning process to generate better outcomes. Ultimately, triple-loop learning leads to the development of a DDDM tool in such a way that the human and the machine mental model are aligned with each other.

REFERENCES

Altman, N., & Krzywinski, M. (2017). Ensemble methods: bagging and random forests. *Nature Methods*, *14*(10), 933–934. https://doi.org/10.1038/nmeth.4438

Andrews, R., Diederich, J., & Tickle, A. B. (1995). Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems*, *8*(6), 373–389. https://doi.org/10.1016/0950-7051(96)81920-4

Appelman, A., & Sundar, S. S. (2016). Measuring Message Credibility. *Journalism & Mass Communication Quarterly*, *93*(1), 59–79. https://doi.org/10.1177/1077699015606057

Argote, L., & Miron-Spektor, E. (2011). Organizational Learning: From Experience to Knowledge. *Organization Science*, *22*(5), 1123–1137. https://doi.org/10.1287/orsc.1100.0621

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., … Herrera, F. (2019). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. Retrieved from http://arxiv.org/abs/1910.10045

Aven, T., & Renn, O. (2009). On risk defined as an event where the outcome is uncertain. *Journal of Risk Research*, *12*(1), 1–11. https://doi.org/10.1080/13669870802488883

Berrar, D. (2019). Cross-Validation. In *Encyclopedia of Bioinformatics and Computational Biology* (pp. 542–545). Elsevier. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Brynjolfsson, E., Hitt, L. M., & Kim, H. H. (2011). Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1819486

Buchanan, S., & Salako, A. (2009). Evaluating the usability and usefulness of a digital library. *Library Review*, *58*(9), 638–651. https://doi.org/10.1108/00242530910997928

Burns, C. M., Vicente, K. J., Christoffersen, K., & Pawlak, W. S. (1997). Towards viable, useful and usable human factors design guidance. *Applied Ergonomics*, *28*(5–6), 311–322.

Cech, T. G., Spaulding, T. J., & Cazier, J. A. (2018). Data competence maturity: developing data-driven decision making. *Journal of Research in Innovative Teaching & Learning*, *11*(2), 139–158. https://doi.org/10.1108/JRIT-03-2018-0007

Chen, Chiang, & Storey. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, *36*(4), 1165. https://doi.org/10.2307/41703503

de Ville, B. (2013). Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*(6), 448–455. https://doi.org/10.1002/wics.1278

Eisenhardt, K. M., & Zbaracki, M. J. (1992). Strategic decision making. *Strategic Management Journal*, *13*(S2), 17–37. https://doi.org/10.1002/smj.4250130904

Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, *61*, 863–905. https://doi.org/10.1613/jair.1.11192

Fu, C., Xu, C., Xue, M., Liu, W., & Yang, S. (2021). Data-driven decision making based on evidential reasoning approach and machine learning algorithms. *Applied Soft Computing*, *110*, 107622. https://doi.org/10.1016/j.asoc.2021.107622

Gregor, S., & Hevner, A. R. (2013). Positioning and Presenting Design Science Research for Maximum Impact. *MIS Quarterly*, *37*(2), 337–355. https://doi.org/10.25300/MISQ/2013/37.2.01

Hutton, R. J. B., & Klein, G. (1999). Expert decision making. *Systems Engineering*, *2*(1), 32–45. https://doi.org/10.1002/(SICI)1520-6858(1999)2:1<32::AID-SYS3>3.0.CO;2-P

Jackson, M. C., & Keys, P. (1984). Towards a System of Systems Methodologies. *Journal of the Operational Research Society*, *35*(6), 473–486. https://doi.org/10.1057/jors.1984.101

Jakeman, A. J., Letcher, R. A., & Norton, J. P. (2006). Ten iterative steps in development and evaluation of environmental models. *Environmental Modelling & Software*, *21*(5), 602–614.

https://doi.org/10.1016/j.envsoft.2006.01.004
Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons*, *61*(4), 577–586. https://doi.org/10.1016/j.bushor.2018.03.007
Kappal, S. (2019). *Data Normalization Using Median & Median Absolute Deviation (MMAD) based Z-Score for Robust Predictions vs. Min-Max Normalization*. https://doi.org/10.13140/RG.2.2.32799.82088
Kline, R. B. (2016). *Principles and practice of structural equation modeling*. New York, NY: Guilford Press.
Kokina, J., & Davenport, T. H. (2017). The Emergence of Artificial Intelligence: How Automation is Changing Auditing. *Journal of Emerging Technologies in Accounting*, *14*(1), 115–122. https://doi.org/10.2308/jeta-51730
Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2018). Feature Selection. *ACM Computing Surveys*, *50*(6), 1–45. https://doi.org/10.1145/3136625
Li, L., Lin, J., Ouyang, Y., & Luo, X. (2022). Evaluating the impact of big data analytics usage on the decision-making quality of organizations. *Technological Forecasting and Social Change*, *175*, 121355. https://doi.org/10.1016/j.techfore.2021.121355
Lukyanenko, R., Castellanos, A., Parsons, J., Chiarini Tremblay, M., & Storey, V. C. (2019). Using Conceptual Modeling to Support Machine Learning (pp. 170–181). https://doi.org/10.1007/978-3-030-21297-1_15
Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Retrieved from http://arxiv.org/abs/1705.07874
Lunenburg, F. C. (2010). The Decision Making Process. *National Forum of Educational Administration and Supervision Journal*, *27*(4).
Ma, Q., & Liu, L. (1986). The Technology Acceptance Model. In *Advanced Topics in End User Computing, Volume 4*. IGI Global. https://doi.org/10.4018/9781591404743.ch006.ch000
Metcalf, L., Askay, D. A., & Rosenberg, L. B. (2019). Keeping Humans in the Loop: Pooling Knowledge through Artificial Swarm Intelligence to Improve Business Decision Making. *California Management Review*, *61*(4), 84–109. https://doi.org/10.1177/0008125619862256
Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, *56*(4), 3005–3054. https://doi.org/10.1007/s10462-022-10246-w
Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., … Seifert, C. (2023). From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Computing Surveys*, *55*(13s), 1–42. https://doi.org/10.1145/3583558
Nissim, D. (2010). Analysis and Valuation of Insurance Companies. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1739204
Nonaka, I. (1994). A Dynamic Theory of Organizational Knowledge Creation. *Organization Science*, *5*(1), 14–37. https://doi.org/10.1287/orsc.5.1.14
Prieto, A., Prieto, B., Ortigosa, E. M., Ros, E., Pelayo, F., Ortega, J., & Rojas, I. (2016). Neural networks: An overview of early research, current frameworks and new challenges. *Neurocomputing*, *214*, 242–268. https://doi.org/10.1016/j.neucom.2016.06.014
Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, *1*(1), 51–59. https://doi.org/10.1089/big.2013.1508
Raisch, S., & Krakowski, S. (2021). Artificial Intelligence and Management: The Automation–Augmentation Paradox. *Academy of Management Review*, *46*(1), 192–210. https://doi.org/10.5465/amr.2018.0072
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939778
Seidel, S., Berente, N., Lindberg, A., Lyytinen, K., & Nickerson, J. V. (2018). Autonomous tools and design. *Communications of the ACM*, *62*(1), 50–57. https://doi.org/10.1145/3210753

Shrestha, A., & Mahmood, A. (2019). Review of Deep Learning Algorithms and Architectures. *IEEE Access*, *7*, 53040–53065. https://doi.org/10.1109/ACCESS.2019.2912200

Steyerberg, E. W., & Vergouwe, Y. (2014). Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European Heart Journal*, *35*(29), 1925–1931. https://doi.org/10.1093/eurheartj/ehu207

Stulp, F., & Sigaud, O. (2015). Many regression algorithms, one unified model: A review. *Neural Networks*, *69*, 60–79. https://doi.org/10.1016/j.neunet.2015.05.005

Uyanık, G. K., & Güler, N. (2013). A Study on Multiple Linear Regression Analysis. *Procedia - Social and Behavioral Sciences*, *106*, 234–240. https://doi.org/10.1016/j.sbspro.2013.12.027

van der Spoel, S. (2016). *Prediction instrument development for complex domains*. University of Twente, Enschede, The Netherlands. https://doi.org/10.3990/1.9789036541749

Waljee, A. K., Higgins, P. D. R., & Singal, A. G. (2014). A Primer on Predictive Models. *Clinical and Translational Gastroenterology*, *5*(1), e44. https://doi.org/10.1038/ctg.2013.19

Wijnhoven, F. (2022). Organizational Learning for Intelligence Amplification Adoption: Lessons from a Clinical Decision Support System Adoption Project. *Information Systems Frontiers*, *24*(3), 731–744. https://doi.org/10.1007/s10796-021-10206-9

Wu, L., Hitt, L., & Lou, B. (2020). Data Analytics, Innovation, and Firm Productivity. *Management Science*, *66*(5), 2017–2039. https://doi.org/10.1287/mnsc.2018.3281

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, *135*, 364–381. https://doi.org/10.1016/j.future.2022.05.014

Yu, L., Zhou, R., Chen, R., & Lai, K. K. (2022). Missing Data Preprocessing in Credit Classification: One-Hot Encoding or Imputation? *Emerging Markets Finance and Trade*, *58*(2), 472–482. https://doi.org/10.1080/1540496X.2020.1825935

APPENDIX A: CASE BACKGROUND

Traditional situation

Before the introduction of the prioritization method (the DDDM tool), decisions on which companies should undergo re-inspection were made in two ways: through direct coordination between the underwriting department and the risk expertise department, and through the re-inspection policy from the Product and Portfolio Management (PPM) department. The selection of companies through direct coordination between the underwriting and risk expertise departments applied to a specific group of companies. Underwriters and risk experts have knowledge of ongoing developments by delving into specific companies and by possibly re-inspecting them. A Business Process Model for the process of selecting companies to re-inspect through direct coordination can be found in Figure 13. Here, underwriters and risk experts are the decision makers, and the risk expert performs the re-inspection. During the re-inspection, risk experts provide the customer with advice on how to mitigate risks. After a (re-)inspection, the risk experts report and process their findings in a system called Arena. They also fill in an inspection usefulness score. Risk experts can assess the usefulness of the re-inspection by selecting one of the five categories: crucial, very useful, useful, unnecessary, and useless. Within a category, they can also indicate to what extent the usefulness belongs in that category. This makes usefulness a numerical score.

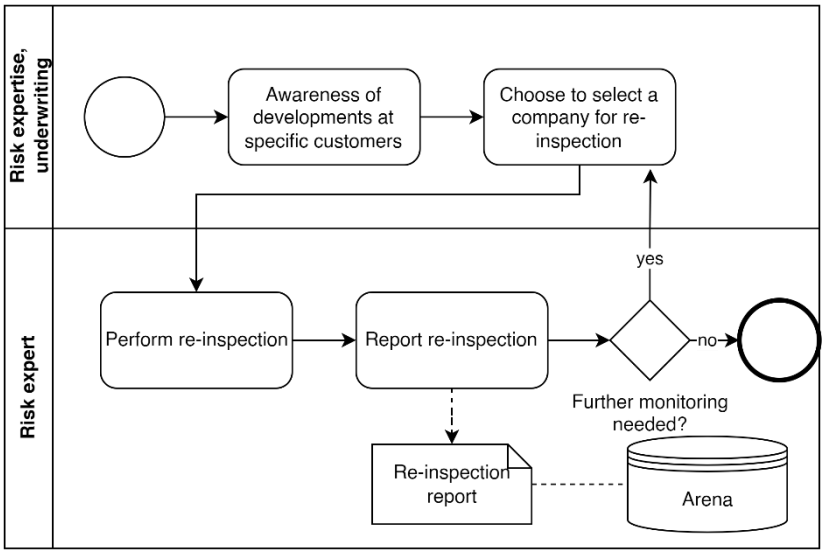

Figure 13: Business Process Model for the process of selecting companies for a re-inspection through direct coordination

The re-inspection policy from PPM was the second way for determining which companies need to be re-inspected and results in a list of buildings that need to be re-inspected. This list was compiled by PPM in consultation with knowledge teams based on signals and insights they observe in practice. Expert judgment from various knowledge teams contributed in establishing priorities and focal points. This list was formed based on the entire portfolio and primarily focuses on trends within clusters of policies, such as specific business activities or industries where risks are significant. A Business Process Model for the process of prioritizing re-inspections according to the traditional policy from PPM can be found in Figure 14.

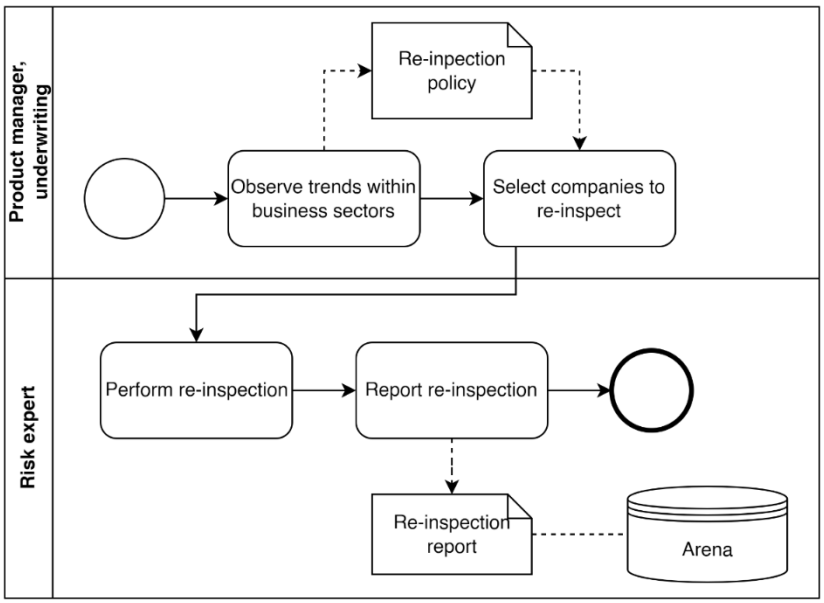

Figure 14: Business Process Model for the process of prioritizing re-inspections according to the traditional policy from PPM

The re-inspection policy from PPM needed to become data-driven because of the difference between the great need for re-inspections and the limited number of available risk experts, leading to operating under conditions of limited resources. It was therefore essential to find a better alternative to determine where the available expertise of the risk expertise department can achieve the most within the portfolio. The goal was to select locations with the highest risk of damage, aiming to intervene preventively to avoid damages. The driving force behind this was to anticipate damage and prevent it rather than conducting post-damage repair. The assumption here was that re-inspection capacity could be best utilized where damages occur because these damages can be prevented by risk experts. The prioritization method is initially and currently intended to replace the re-inspection policy from PPM and not to replace the coordination between underwriting and risk expertise, hence the focus of this case study is on the re-inspection policy process from PPM given in Figure 14.

Current situation

The data scientists have developed a method that prioritizes damage based on predicting the damage probability and the damage burden. The list of prioritized companies resulting from this prioritization method is forwarded to employees of the underwriting department, employees of the risk expertise department, and to the product manager. They make changes to the list where necessary, supplement the list, and authorize the list. The risk experts carry out the re-inspections according to the list. The Business Process Model for the process of selecting re-inspections using the prioritization method can be found in Figure 15. Within the business process, the prioritization method is a comprehensive term for the damage probability model and a damage burden model, for merging these results into a final prioritization, and for formatting a list representing the buildings that should be re-inspected according to the prioritization. In the process, the data scientist is responsible for the prioritization method, the product manager together with the risk expertise department and acceptance department for making decisions about re-inspections, and the risk experts for carrying out the re-inspections. The process of forming a new priority list of buildings to re-inspect buildings repeats yearly, with the data scientist and the product and risk expertise managers ensuring that the list of re-inspections can be carried out at the beginning of a new calendar year, and the risk experts conducting the inspections of buildings on the list throughout the calendar year.

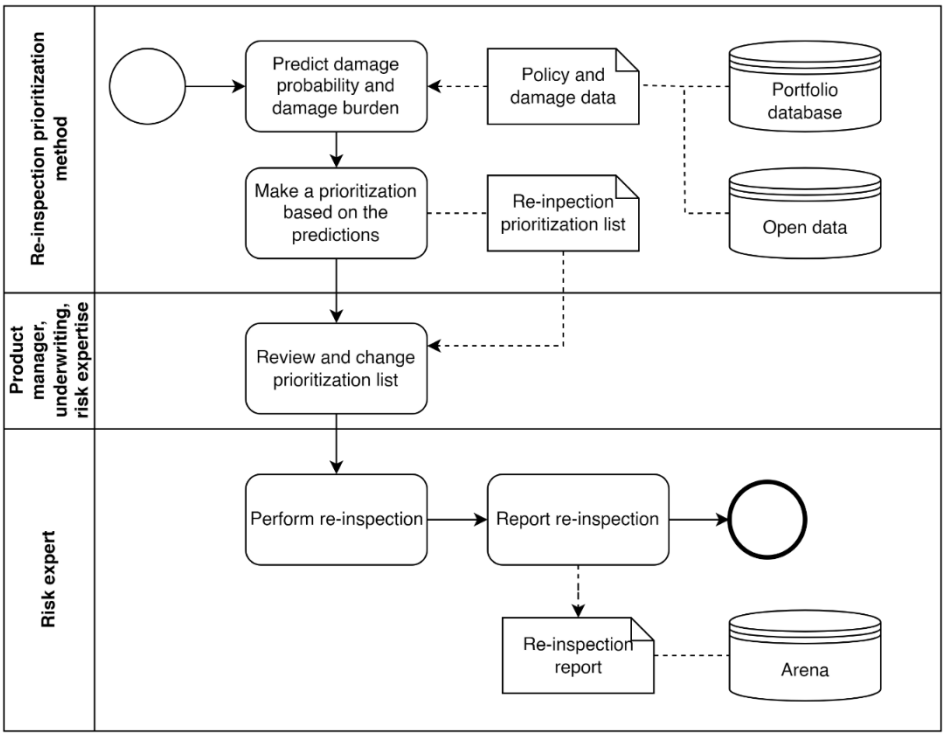

Figure 15: Business Process Model for the process of selecting re-inspections using the prioritization method

The computation with which the re-inspection prioritization is made ('Make a prioritization based on the predictions' in Figure 15) differs depending on the scale of the customer's company. The prioritization method for small and medium-sized enterprises (SMEs) is created by multiplying the predicted probability and burden of the damage and ranking this outcome from high to low. The prioritization method for large enterprises is created by first ranking the outcomes of the predicted probability of damage and the predicted damage burden separately and then taking the inverse of this ranking number for both the predicted probability and burden. These two inverse numbers are added together and then ranked again from high to low to form the final prioritization for large enterprises. As a result, large enterprises that have a high score in the damage probability model or a damage burden model also rank high in the final prioritization method. This prioritization method is considered the most advantageous because it identifies the companies that submit the largest sum of claims over the years. The prioritization method for SMEs and large enterprises is visualized in Figure 16.

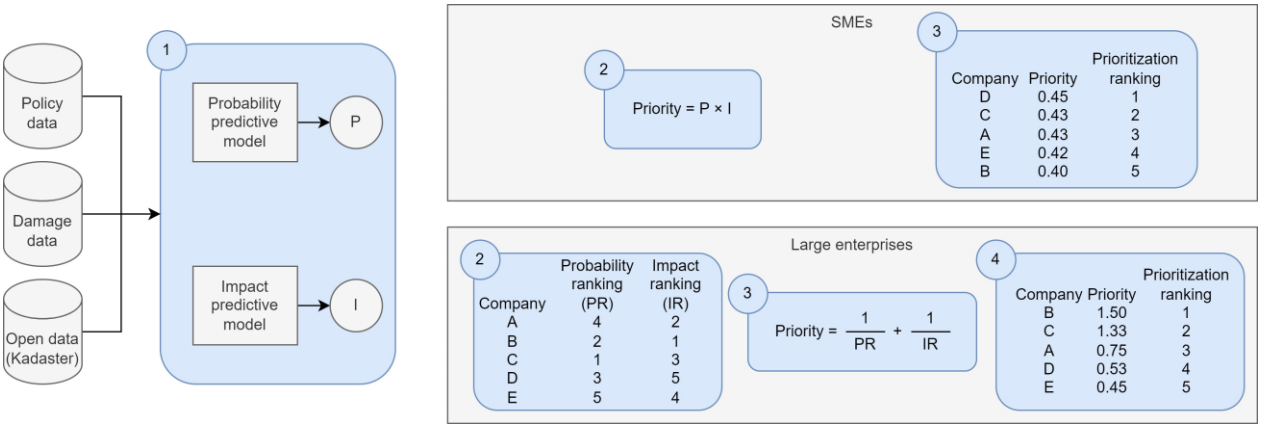

Figure 16: Visualization of the current re-inspection prioritization method within Achmea

Anticipated situation

To solve the lack of interaction between the model and risk experts in the current (re-)inspection prioritization, a usefulness predictive model is developed in this study. Implementing this predictive model should lead to a new Business Process Model, which can be found in Figure 17. In contrast to the Business Process Model in the current situation, the anticipated situation incorporates a feedback loop. In this loop, a usefulness prediction is generated using the usefulness scores provided by risk experts, facilitating the processing of feedback and soft information into the prioritization method in order to arrive at re-inspections that are more useful. Such a feedback loop not only contributes to the improvement of the tool but also provides opportunities for users to gain insights and learn from. Moreover, the incorporation of risk experts' assessments into the prioritization method contributes to a sense of user influence. This perception among end users enhances the tool's acceptance among end users and other stakeholders.

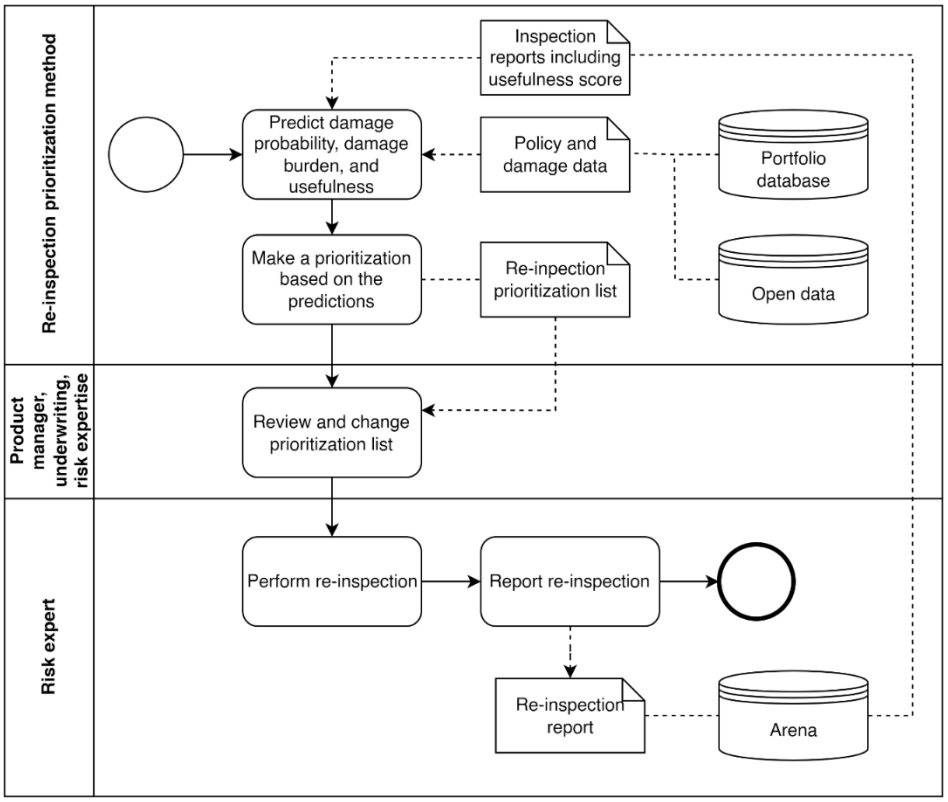

Figure 17: Business Process Model for the process of selecting re-inspections where a usefulness prediction is incorporated in the re-inspection prioritization method

APPENDIX B: LITERATURE SEARCH METHODS

To provide the research with a relevant theoretical framework, literature is collected. An overview of the relevant literature topics, the corresponding search terms, and the number of papers that are included in the research can be found in Table 10. The literature search is conducted using Scopus, FindUT, and Google Scholar. The snowball methodology has also been used, meaning that the bibliography in some books or journal articles is used to find other relevant sources. The literature has to answer the following questions:

1. What is usefulness in the context of re-inspections of commercial property?
2. What predictive modelling algorithms exist and what are their characteristics?

3. How can predictive models become more interpretable?
4. What are the essential steps in the development of predictive models?
5. How can an organizational learning approach enhance data-driven decision-making?

At first, a literature review is conducted on insurance inspections and the concept of usefulness to investigate whether previous research has been done on this concept, and to arrive at a definition based on the state-of-the-art literature of the concept.

Given that the research involves developing a predictive model, research is conducted into the predictive model algorithms used in this research. Special attention is given to the concept Explainable AI and to the extent to which the various algorithms are explainable. Explainable AI seems to be important in order to make a predictive model understandable for its users. Moreover, research is conducted into the different development phases of a predictive model. This is necessary to ensure a systematic approach to the model's development.

This research attempts to develop a predictive model from an organizational learning approach because the research problem is at the intersection of people and computers where both can learn from each other and is therefore not exclusively a data science or computer science research problem. Therefore, literature search on organizational learning is conducted. The research focuses on data-driven decision-making for prioritization, so literature will also be examined on this topic. Attention is given to the concept of "human in the loop", which emphasizes the collaborative role of human decision-makers in conjunction with data-driven insights. The literature review on organizational learning, data-driven decision-making, and the development of a predictive model together should contribute to a proposed methodology for creating a predictive model through an organizational learning approach that enables triple-loop learning.

By conducting these literature research methods, the aim is to build a relevant theoretical framework for this research about predictive modelling algorithms and methods and about effective application of data-driven decision-making within organizations in order to arrive at a method with which a predictive model can be developed using an organizational learning approach.

Table 10: Literature research topics and corresponding search queries

Subject	Search terms	References
Predictive model algorithms	• regression analysis • neural network • deep learning • decision trees • random forests	(Stulp & Sigaud, 2015) (Uyanık & Güler, 2013) (de Ville, 2013) (Biau & Scornet, 2016) (Altman & Krzywinski, 2017) (Prieto et al., 2016) (Shrestha & Mahmood, 2019)
Explainable AI	• (explainable OR interpretable) AND (artificial intelligence OR machine learning OR AI)	(Arrieta et al., 2019) (Nauta et al., 2023) (Andrews, Diederich, & Tickle, 1995) (Lundberg & Lee, 2017) (Ribeiro et al., 2016)
Development phases of a predictive model	• development OR development phases OR development steps AND predictive model OR machine learning model • feature selection	(Waljee et al., 2014) (Jakeman et al., 2006) (Steyerberg & Vergouwe, 2014) (Lukyanenko et al., 2019) (Jackson & Keys, 1984) (van der Spoel, 2016) (J. Li et al., 2018)
Data-driven decision making	• decision OR decision making OR concept decision OR definition decision • data driven decision making OR data decision making	(Hutton & Klein, 1999) (Eisenhardt & Zbaracki, 1992) (Lunenburg, 2010) (Brynjolfsson et al., 2011) (L. Wu et al., 2020) (Cech et al., 2018) (Fu et al., 2021) (Provost & Fawcett, 2013) (L. Li et al., 2022) (Raisch & Krakowski, 2021) (Kokina & Davenport, 2017)
Human in the loop	• human in the loop OR human-in-the-loop	(X. Wu et al., 2022) (Mosqueira-Rey et al., 2023) (Jarrahi, 2018)
Organizational Learning	• organizational learning OR organizational knowledge OR organizational learning AND triple-loop learning	(Jarrahi, 2018) (Argote & Miron-Spektor, 2011) (Wijnhoven, 2022) (Nonaka, 1994) (Seidel et al., 2018) (Metcalf et al., 2019)

APPENDIX C: PREDICTIVE MODELLING ALGORITHMS

Multiple regression

Multiple linear regression, in short multiple regression, is a statistical method used to examine the linear relationship between two or more independent variables having a relation (Stulp & Sigaud, 2015; Uyanık & Güler, 2013). In contrast to simple linear regression, which involves only one independent variable, multiple linear regression incorporates several predictors to better model the complexity of real-world relationships. The general form of multiple linear regression is represented by equation (6).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \quad (6)$$

In equation (6),

- Y is the dependent variable,
- $X_1, X_2, \ldots, X_n$ are the independent variables,
- β_0 is the y-intercept and represents the value of Y when all independent variables are zero.
- $\beta_1, \beta_2, \ldots, \beta_n$ are coefficients that represent the change in Y associated with a change in the corresponding independent variable, assuming all other variables remain constant,
- $X_1, X_2, \ldots, X_n$ are the independent variables,
- ε is the error term, representing factors that affect Y but that are not accounted for by the model.

The goal in training a multiple linear regression is to estimate the coefficients ($\beta_1, \beta_2, \ldots, \beta_n$) that minimize the sum of squared differences between the predicted values and the actual values of the dependent variable (Uyanık & Güler, 2013). The assumptions of multiple linear regression include linearity, independence of errors, homoscedasticity (constant variance of errors), and normality of errors. Additionally, multicollinearity (high correlation between independent variables) can impact the reliability of coefficient estimates. In the context of prediction, the multiple regression model establishes a linear relationship between the target variable and features by estimating coefficients that quantify the strength and direction of these relationships. Multiple regression allows for a more nuanced understanding of the impact of multiple factors on the predicted outcome, enabling effective predictions in scenarios where various variables contribute to the overall outcome.

Random forest

A decision tree is a machine learning algorithm used for both classification and regression tasks. They split the data based on different features to create a set of if-else conditions that lead to a prediction (de Ville, 2013). Decision trees use a tree-like structure of nodes, where each internal node represent a feature or attribute, each branch connects nodes and shows the flow from question to answer, and each leaf node represents a final prediction. Decision trees are attractive due to their interpretability and ease of visualization. However, they can be prone to overfitting, especially when the tree is deep.

Random Forests are so-called ensemble learning methods that build multiple decision trees and merge their predictions to improve accuracy and reduce overfitting (Biau & Scornet, 2016). The “random” in Random Forest comes from the fact that each tree is trained on a random subset of the data, also called a bootstrap sample, and at each split, a random subset of features is considered. The predictions from the different trees are aggregated to one outcome, which is called bootstrap aggregating.

Random Forest is well-suited for both classification and regression tasks (Altman & Krzywinski, 2017). In classification, the predicted class is determined through a process known as voting. Each tree independently predicts the class, and the class with the majority of votes across all trees is assigned as

the final prediction. In regression tasks, Random Forest uses averaging. Each tree provides a numeric prediction, and the final regression prediction is obtained by averaging these individual predictions.

Random Forests are less sensitive to overfitting compared to individual decision trees. However, the combined output of multiple trees in Random Forests can be challenging to interpret. Identifying and understanding the specific decision path or set of rules that lead to a particular prediction in a Random Forest is not as straightforward as in a single decision tree.

Neural Network

Neural networks are a type of machine learning models inspired by the structure and functioning of the human brain: they consist of layers of interconnected neurons that process information in parallel (Prieto et al., 2016). The three main types of layers are the input layer, hidden layers, and output layer. The input layer receives the initial data, the hidden layers process this data through weighted connections, and the output layer produces the final result. A neuron takes multiple inputs, applies weights to these inputs, sums them up, and passes the result through an activation function to produce an output. Activation functions introduce non-linearity to the model, allowing it to learn complex, non-linear patterns.

Weights are parameters associated with the connections between neurons and determine the strength of the influence of one neuron on another. Biases are additional parameters in each neuron that allow the model to account for variations and to make the model more flexible. In the training phase, data is fed through the network in a process known as feedforward. The predicted output is then compared to the actual output to calculate the error. This error is then propagated in reverse through the network, initiating the adjustment of weights and biases—a process commonly referred to as backpropagation.

In a neural network designed for regression, the output layer typically consists of a single neuron that has a linear activation function. In a neural network designed for classification, the output layer usually has one neuron per class. In binary classification, the output layer typically uses the sigmoid activation function for a probability output between 0 and 1, while in multi-class classification, the softmax function is employed to normalize outputs into class probabilities summing to 1, and the class with the highest probability is predicted.

Neural networks with multiple hidden layers are referred to as deep neural networks (Shrestha & Mahmood, 2019). Deep learning leverages the power of deep neural networks to learn intricate patterns and representations from data. Convolutional neural networks are specialized for processing grid-like data, such as images. They use convolutional layers to automatically and adaptively learn spatial hierarchies of features. Recurrent neural networks are designed for sequence data, like time series or natural language, and can be seen as forecasting algorithms. They have connections that form directed cycles, allowing them to maintain a memory of previous inputs.

Neural networks excel in learning complex patterns and nonlinear relationships in the data. However, neural networks operate as black boxes, meaning that their internal process is hard to interpret and explain. This lack of transparency can be a crucial limitation in domains such as decision-making.

APPENDIX D: INTERVIEW QUESTIONS

Questions

1. How are you involved with the re-inspections? *Hoe bent u betrokken bij de herinspecties?*
2. When do you consider a re-inspection useful? *Wanneer beschouwt u een her-inspectie als nuttig?*
3. What factors influence the usefulness of re-inspections? *Welke factoren hebben invloed op de nuttigheid van her-inspecties?*
4. Suppose we create a predictive model that predicts the usefulness of the re-inspection, what functionalities should be considered "must-haves"? *Stel dat we een voorspelmodel maken die de nuttigheid van de her-inspectie voorspelt, over welke functionaliteiten moet dit nuttigheidsmodel beschikken?*
5. How can the usefulness predictive model result in an improvement of (the prioritization process of) the re-inspections? *Hoe kan het nuttigheidsvoorspelmodel ervoor zorgen dat (het prioriteringsproces van) de herinspecties verbeterd wordt?*
6. Are there any specific data from re-inspections that you believe should be documented but currently aren't? How do you think this could be improved? *Zijn er gegevens van her-inspecties waarvan u graag had gezien dat deze vast zouden kunnen worden gelegd, maar die op dit moment nog niet worden vastgelegd? Hoe zou dit volgens u verbeterd kunnen worden?*
7. What are the advantages of the currently used re-inspection prioritization method compared to expert judgment for prioritizing re-inspections? *Wat zijn voordelen van het huidige her-inspectie prioriteringsmodel ten opzichte van expert judgement voor het prioriteren van her-inspecties?*
8. What are the disadvantages of the currently used re-inspection prioritization method compared to expert judgment for prioritizing re-inspections? *Wat zijn nadelen van het huidige her-inspectie prioriteringsmodel ten opzichte van expert judgement voor het prioriteren van her-inspecties?*
9. Where do you expect and hope to be in 5 years when prioritizing re-inspections? *Waar verwacht en hoopt u te staan over 5 jaar bij het prioriteren van herinspecties?*

Question	Risk expert	Manager risk expertise	Underwriter	Manager underwriting	Data scientist	Product manager
1	X	X	X	X	X	X
2	X	X	X	X	X	X
3	X	X	X	X	X	X
4	X	X			X	X
5					X	X
6	X	X			X	X
7	X	X	X	X	X	X
8	X	X	X	X	X	X
9		X		X	X	X

APPENDIX E: DATA UNDERSTANDING AND PREPARATION

The data understanding and preparation process is structured as follows: first, research was conducted into the relevant enterprise systems; then, descriptions were provided for the relevant databases of the enterprise systems; next, an investigation was conducted into the optimal methods for linking the datasets to each other; finally, the total data preparation was described and visualized.

E.1 ENTERPRISE SYSTEMS

This section discusses the enterprise architecture of the systems used for the re-inspection prioritization method. Figure 18 illustrates the architecture model, depicting the interconnection of applications and technologies, as well as their utilization within the re-inspection business process. The figure is made using the ArchiMate Enterprise Architecture Modeling Language[2]. The overview is based on the current state of the re-inspection prioritization method. Consequently, it serves to give the right context to the systems within the organization and as an introduction to the subsequent data preparation.

The business process in Figure 18 starts with the (re)development of the re-inspection prioritization method. A data scientist prepares historical policy and its accompanying damage data for buildings. Currently, data is only prepared from the Kameleon database, as the prioritization is exclusively conducted for Centraal Beheer at present. Subsequently, the data scientist trains the model with the prepared historical policy and damage data. The data scientist then runs the prioritization method on the current portfolio of policy data for buildings. The model prioritizes based on predicted chances of damage and damage cost, creating a list of buildings that need to be re-inspected. After inspection and approval of the list from the product manager and the manager of the risk expertise department, risk experts carry out the re-inspections from the prioritization list and generate a report through the Arena application.

The figure describes three databases relevant to the re-inspection: the Kameleon database, which stores policy data for Centraal Beheer, the BCP database, which stores policy data for Interpolis, and the Arena database, where the reports of re-inspections are stored. Only data from the Kameleon database has been used for the prioritization method up to now. The link between data from Kameleon and data from BCP, and the link between these datasets and Arena, has not been made before. However, to calculate usefulness, a link between the data from all databases is needed. Linking policy data with Arena is necessary to associate the usefulness score with specific policies, enabling predictions based on policies. Linking Kameleon data with BCP data is necessary to expand the policy dataset. This is because the linked policy – re-inspection dataset for Centraal Beheer is too small to train a predictive model for the usefulness.

There is a desire within Achmea to consolidate the systems of all business insurance policies from various brands into one system: SKB+. In the business insurance chain of Achmea, inefficiency is experienced due to differences in the backend between labels, meaning that the same processes have to be repeated for different labels instead of doing it all at once. Therefore, the Management Team of the Non-Life Companies division has instructed the development of a unified backend for all labels, aiming to promote efficiency and thereby reduce workforce costs. This is symbiotic learning, as the learning outcome that there is a need for increase in data quality is being implemented. This migration is a multi-year project. An essential part of this migration for the prioritization method is data integration, specifically the integration of the policy portfolio for Kameleon and BCP. All data from

[2] https://pubs.opengroup.org/architecture/archimate3-doc/index.html

different brands must be migrated to a unified format. The complexity lies in the fact that data from different labels have their own semantics, as will become clear in the next chapter.

Due to the migration to SKB+, many changes are expected for re-inspections and their prioritization in the future. Prioritization based on the re-inspection prioritization method will not only occur for Centraal Beheer but for all labels in the future. The procedures for re-inspections, which currently vary per label, will be standardized through the migration to SKB+. Additionally, with the migration to SKB+, data quality is expected to improve, making data preparation for the re-inspection prioritization method much easier. However, the SKB+ migration is currently in development, which means a provisional solution will need to be found to link the data from Kameleon and BCP to each other and to Arena.

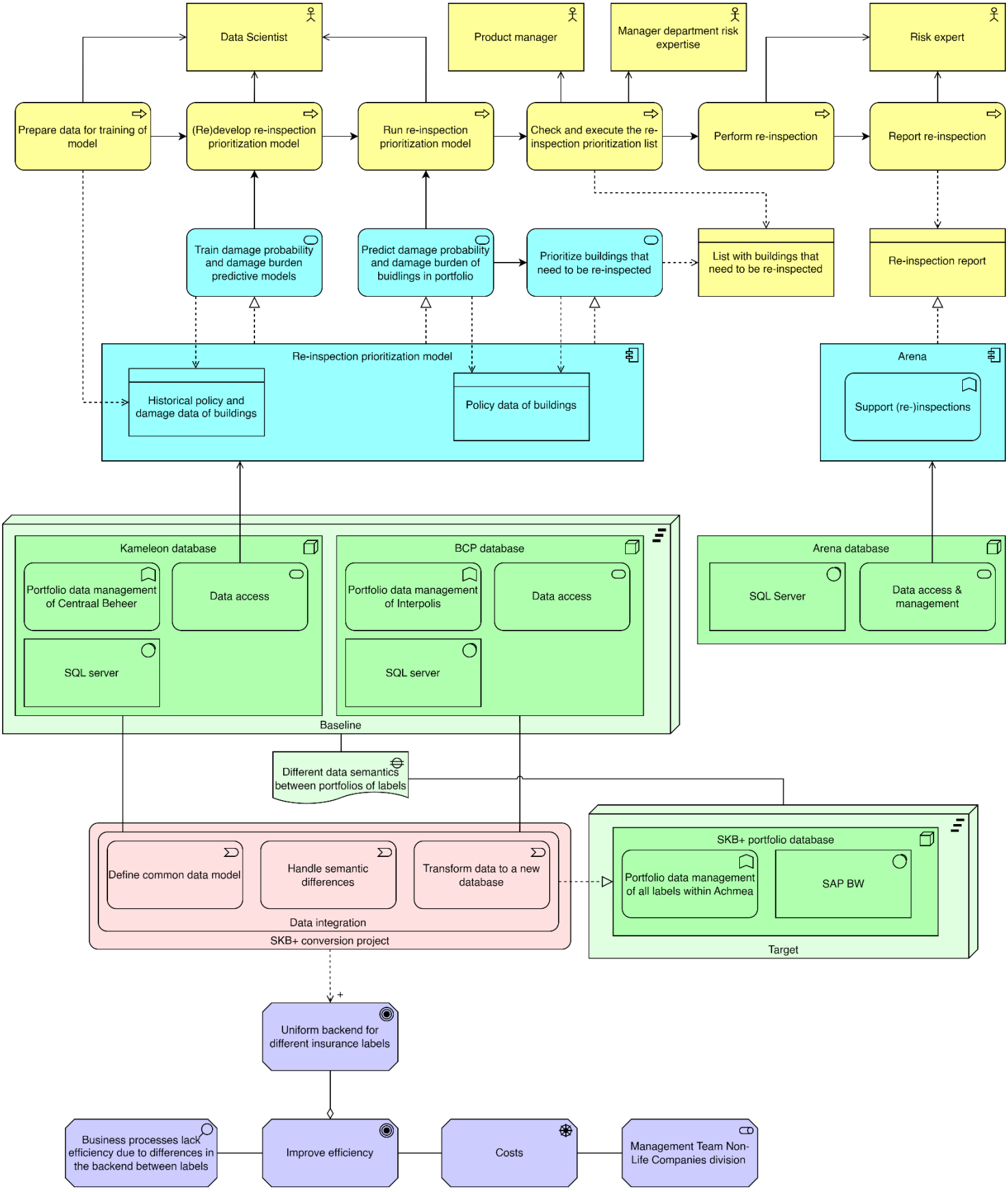

Figure 18: Architecture of the systems and migration relevant for the re-inspection prioritization method

E.2 INITIAL DATABASES

This section describes the three initial databases that contains data needed to develop a re-inspection usefulness predictive model. The Kameleon database encompasses policy and damage data for Centraal Beheer's insured policies, while the BCP database covers the same data for Interpolis' insured policies. Additionally, the Arena database encompasses re-inspection data for all labels within Achmea, including Centraal Beheer and Interpolis.

Database for policy data of Centraal Beheer

Centraal Beheer's policy and claims data are recorded via a system called Kameleon and stored in a relational database. A simplified overview of the relational database can be found in Figure 19.

The core of this database is the class Insured object, in which the policy data is stored at the smallest level, which is a business object. The contract number represents a contract between Centraal Beheer and a policyholder. Within a contract, insurance policies can be issued for various business objects, for example for multiple buildings. The object number serves to differentiate between these insured business objects. Therefore, a combination of an object number and a contract number constitutes a policy.

The class Insured object contains an object code, which is used to categorize the type of the business object that is insured via the class Object type. Additionally, the sector code and sector sequence number enable the determination of the business sector to which the insured object belongs via the class Sector. Furthermore, the class Insured object contains the address details of the insured object and a code for the commercial product, by which the type of company (SME, large business, government, healthcare, exploitation, etc.) can be determined.

The class Insured object is connected to Contract through the combination of the contract number, the contract sequence number, and the object number. The class Contract contains the insured amount of the object per contract number and insured object. In addition, the class contains a relation number. The relation number indicates under which relation or company the contract is insured, as multiple contracts may have been concluded per relation. The class also contains a start and an end date of the contract and the status of the contract, which indicates whether the contract is still active. The dates and contract status are specific to a unique combination of a contract number and a contract sequence number. Whenever alterations are made to the contract, a new contract sequence number is generated to accurately document these changes in the database.

In summary, the relation number in this database represents a policy holder. Because a policy holder can conclude multiple contracts, a relation number can have multiple contract numbers. Insurance policies for one or more business objects are recorded in contracts, meaning that a contract number can contain multiple object numbers. Also, due to adjustments or renewals of the contracts, multiple contract versions are stored per combination of a contract and an object. These versions are identified by the contract sequence number.

The damage data is stored in the database in the class Damage. Damage is recorded at object level and can be linked to the class Insured object through the combination of the contract number and the object number. The date of the damage can be used to link damage to a contract sequence number. For each damage, further information is known about the damage burden and the cause of damage. The cause of damage can be determined using the cause of damage code in the class Cause of damage. The clause data is stored in the database in the class Clause, recorded at object level and can also be linked to the class Insured object through a combination of the contract number and the object

number. The clause name indicates the name of the clause and the date of clause indicates the date on which the clause is made active.

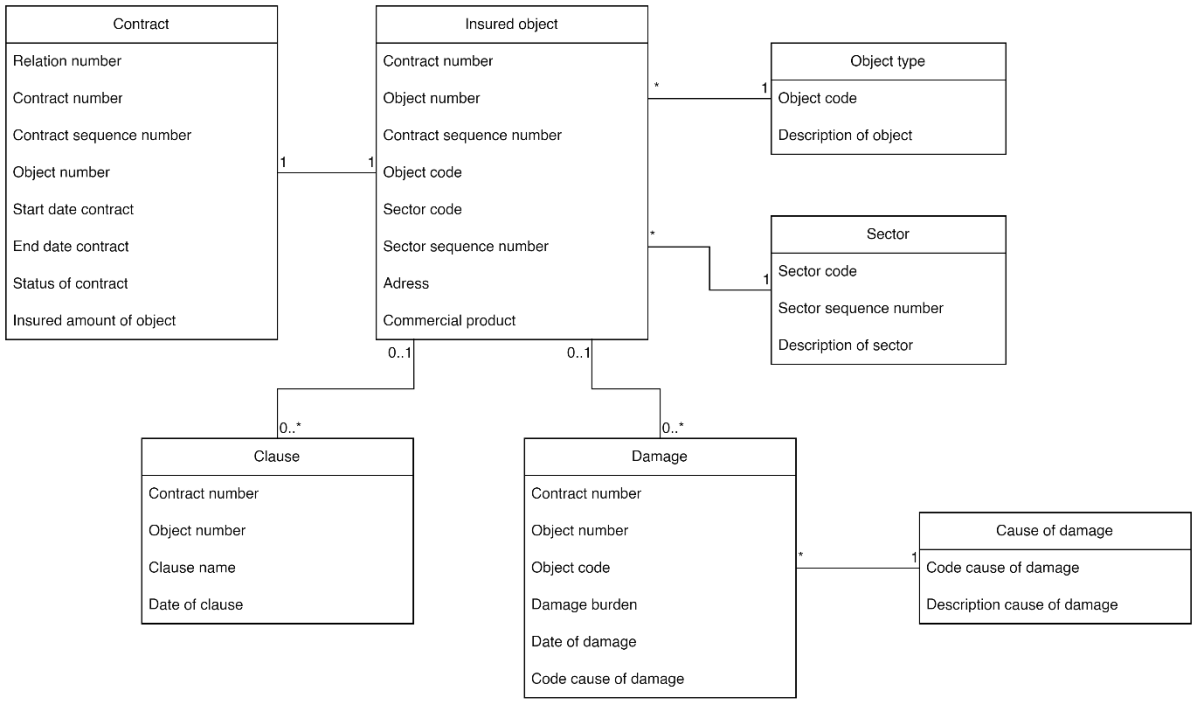

Figure 19: Class diagram of storage of policy and claims data from Centraal Beheer (the Kameleon database)

Database for policy data of Interpolis

The policy and claims data of Interpolis are registered in a system called BCP and stored in a relational database, of which a simplified overview can be found in Figure 20. The database shows similarities with the Centraal Beheer policy database. However, there are some key differences between the Interpolis policy database and the Centraal Beheer policy database.

The core class of Interpolis is the Policy class, in which policy data is described at the level of business objects. The highest level of policy data in the Interpolis database is the policy itself, which can be recognized by the policy number, and the lowest level of policy data is a business object, which can be recognized by an object sequence number. The object's type can be determined by referencing the object code within the Object class, while the sector's type can be identified by the activity code within the Activity class. The sector in the Activity class is comparable to the description of sector within the Centraal Beheer dataset. Nonetheless, they possess distinct semantics in the category of sector given to a business.

Interpolis' claims data is described in two classes, with one class containing Interpolis' claims data up to and including 2018 and one class containing claims data from 2019 onwards. Essentially, both classes store similar types of data. However, starting from 2019, the product coverage and cause of damage are stored within the damage class, whereas up until 2018, these attributes are referenced in a distinct class.

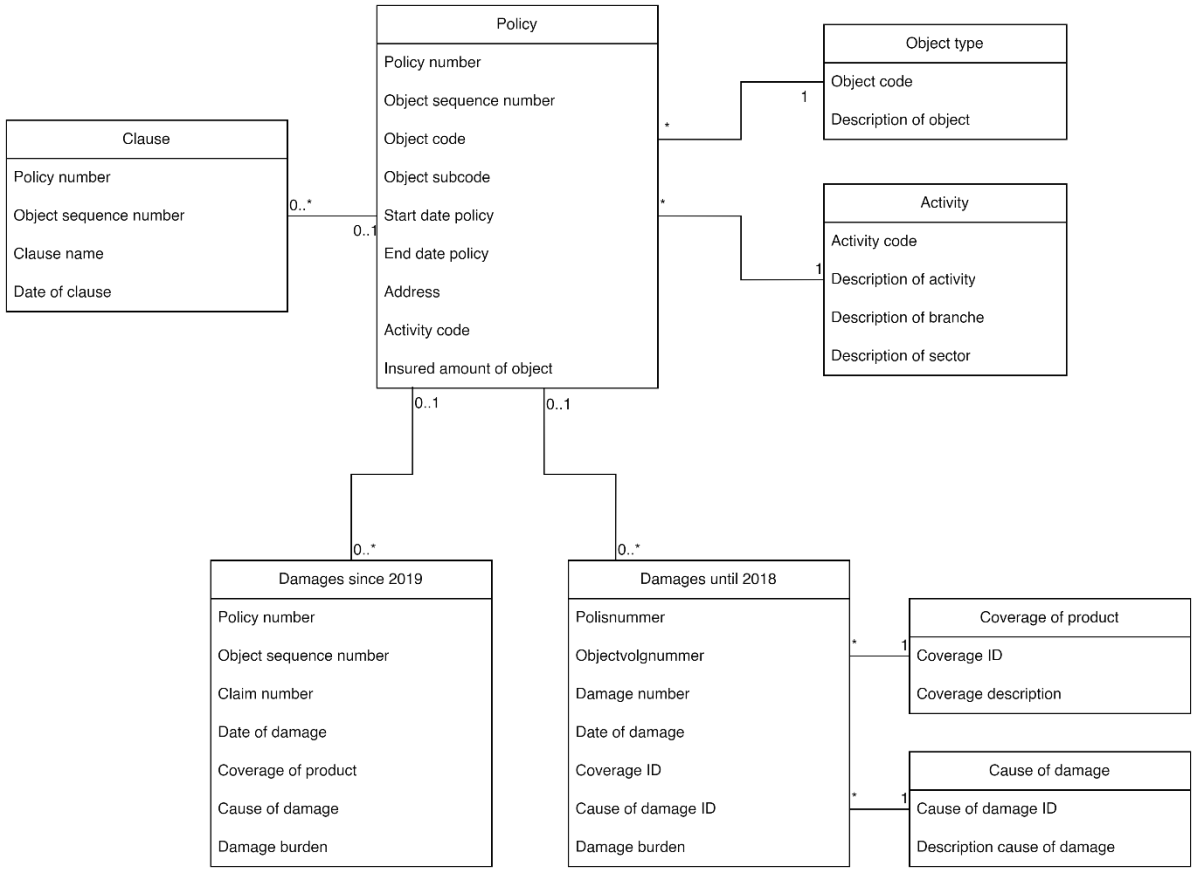

Figure 20: Class diagram of storage of policy and claims data from Interpolis (the BCP database)

Difference of registering policy data between Centraal Beheer and Interpolis

The difference between Centraal Beheer and Interpolis datasets is evident in their distinct ontologies devised for insurance policies. The ontologies have not undergone the combination process as shown at number four in Figure 5, which results in inconsistencies in the machine's mental model. Those inconsistencies have to be solved during the data preparation. The process of combining the databases and finding a sustainable solution for the inconsistencies is currently in progress with the SKB+ conversion project. In Centraal Beheer, data is organized with the relation number level at the highest level and the policy number at the lowest level, whereas Interpolis operates at the policy number as the highest level and object number as the lowest level of data. Additionally, there is a contrast between de datasets in how the sector assigned to a company is interpreted semantically. Those dissimilarities present no inherent issue, as their ontologies are closely aligned and there is a mapping available for the differences in semantics of the business sector. However, it is crucial to be mindful of this when integrating these datasets, as they must be linked to Arena in a different manner.

Database for re-inspections

The information that a risk expert enters during a re-inspection is processed by the Arena system and stored in a database. A simplified class diagram of this database can be found in Figure 21. The classes are organized by data categories. The classes each contain information at the level of a re-inspection and can be linked together via a re-inspection ID. Although the division into classes suggests that the database is relational, the database is a flat file database and the data is categorized according to classes.

Particularly important for this investigation are the usefulness of a re-inspection, the date on which the re-inspection took place and the data with which a re-inspection can be linked to a policy. These are represented in the classes Summary and Adress and policy details.

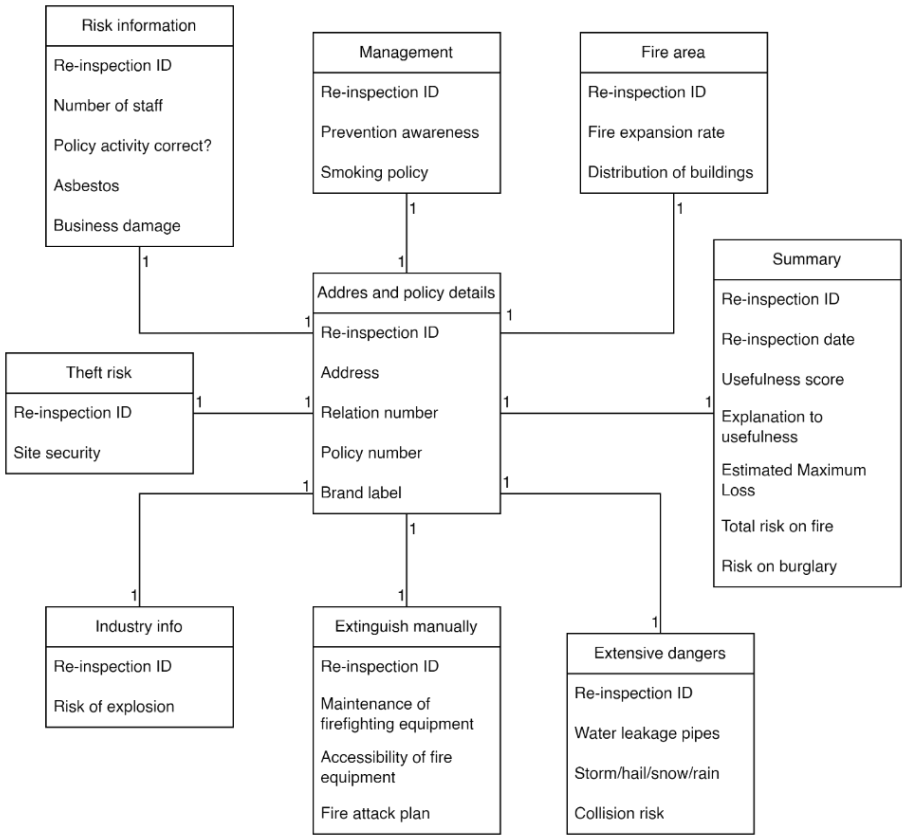

Figure 21: Class diagram of re-inspection data storage (the Arena database)

E.3 DATA CONNECTION BETWEEN DATABASES

For the datasets of Centraal Beheer and Interpolis, establishing a one-to-one connection with Arena proves challenging, as only the highest level of the insured company is registered within Arena, which is either the relation number within Centraal Beheer or the policy number within Interpolis. Not all business objects of the insured company are re-inspected, and it remains unclear which objects are re-inspected and which are not. In some cases, the object code within Arena is entered in a free-text field and can be utilized as link to the re-inspected business objects. However, in most cases, the data needs to be linked in a different manner. Therefore, this chapter explores how Arena data can be linked with data from Centraal Beheer and Interpolis. This linkage is based on the registered business objects when available, and otherwise on the relation number or the policy number and the postal code or the address.

Preparation of re-inspection data for Centraal Beheer

The Arena system records data on re-inspections for the insurance brands Centraal Beheer, Interpolis and Avero. The total number of registered re-inspections in Arena is 6710. Initially, the data from Arena will be linked to that of Centraal Beheer. The number of re-inspections carried out for Centraal Beheer and recorded in Arena is 1011. The usefulness of re-inspections was initially not recorded in Arena, which means that the Arena database contains re-inspections that do not contain a usefulness score. However, for this study, only those re-inspections are useful for which a usefulness score is available.

The number of re-inspections carried out for Centraal Beheer, during which the usefulness was recorded, is 420.

However, not every recorded usefulness score turns out to be useful. Over time, it was found that some risk experts did not enter a usefulness score and did not provide an explanation of the usefulness of the re-inspection. As a result, the usefulness score in the system was set to 50 by default. Therefore, re-inspections with a usefulness score of 50, where no explanation of usefulness was provided, were excluded. If the usefulness score is 50, but an explanation of the usefulness of the re-inspections is provided, it is assumed that the score of 50 has been entered by the risk expert. The number of re-inspections with a non-standard usefulness score is 342. A schematic overview of the preparation and selection of the Arena data is shown in Figure 22.

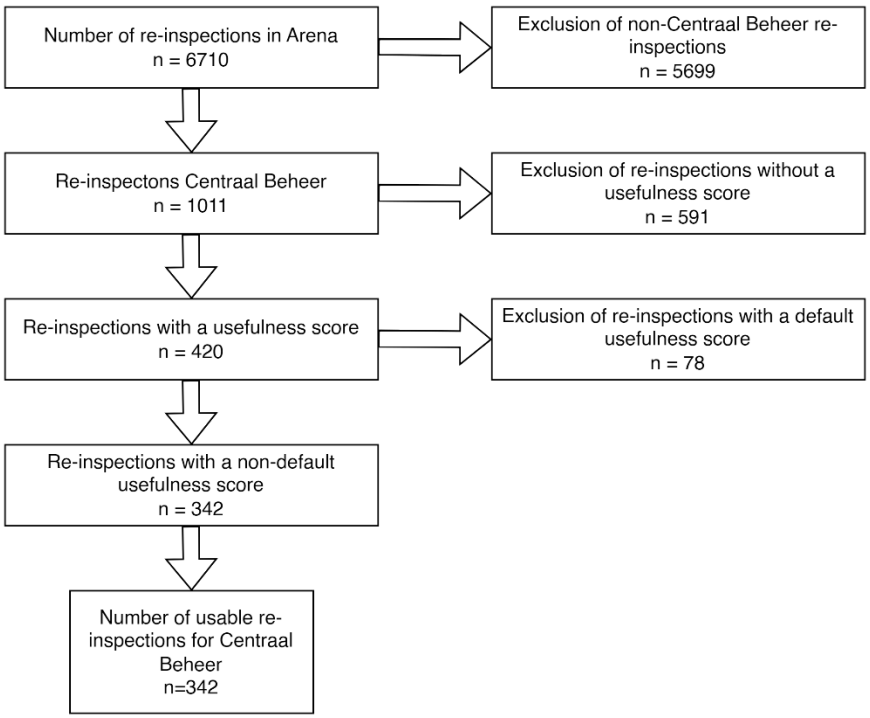

Figure 22: Selection of usable re-inspection data for Centraal Beheer policy re-inspections

Joining re-inspection data to the policy data of Centraal Beheer

When designing Arena, the technical link between re-inspection data and policy data was not taken into account. At the time, there was no need to design Arena in such a way that this data could be directly linked to each other. As a result, the Arena database does not have a built-in key to connect re-inspection data with Centraal Beheer's policy data. However, there are open fields in the Arena database that in 121 of the 342 cases have been used to record a contract number and object number during re-inspections for Centraal Beheer, making a direct join to re-inspected business objects possible for those cases. Several alternatives are available for the 221 re-inspections where direct joining to the re-inspected business objects is not possible. An overview of the available alternatives can be found in Figure 23.

The first alternative is to join the re-inspection data to the policy data using the relation number. Of the 221 re-inspections that have to be joined, only 114 re-inspections can be joined based on a relation number. This is because the relation number in Arena is a free field and is not always filled in consistently. A disadvantage of joining the data using the relation number is that there are considerably more business objects joined to the re-inspection than the business objects that have been re-inspected as a relation can have numerous insured business objects. For this reason, joining

re-inspection data to the policy data of Centraal Beheer based on the relation number is not a suitable solution.

The second alternative is to connect the re-inspection data to the policy data based on the address. This allows 123 of the 244 re-inspections to be joined to a contract. This method guarantees that the objects on a contract are located at the re-inspected address. There are also disadvantages to this method. First, multiple companies may be confirmed at one address, leading to an incorrect join between contracts and a re-inspection. Secondly, the house number is not entered consistently, which means that some of the re-inspections cannot be joined to the policy data based on the address. Therefore, this alternative is also not the best solution.

The third alternative is to join the re-inspection data to the policy data by using the relation number in combination with the zip code. This joins 111 of the 221 re-inspections to the policy data. The advantage is that the re-inspection and the policy have a match in the policy holder and the zip code, which means there is a great chance that the business objects have actually been re-inspected. The disadvantage is that it is not certain whether a business object belongs to the re-inspected address or whether it is located at another address within the same zip code, which reduces the reliability of this join.

The fourth alternative is to join the data from the re-inspections to the policy data by using the contact number in combination with the address. 86 re-inspections are joined to the Centraal Beheer policy data. An advantage over the other alternatives is that the policyholder and address match, making it likely that the insured business properties on that contract have been re-inspected. However, both the address and the relation number in Arena are not entered consistently, resulting in a few joinable re-inspections.

When joining the re-inspection data of Centraal Beheer to the policy data of Centraal Beheer, it turned out that the policy data is correct, but the re-inspection data contains imperfections. This is partly due to errors in completing the re-inspection data, but also because the level of re-inspection can vary: a re-inspection can take place over business objects from multiple contracts, but a contract can also contain business objects from multiple addresses. This can be taken into account in the analysis by only joining the business objects within a contract to a re-inspection that are located at the same location. Therefore, the alternatives that are joined on relation number and a location are both possible alternatives for joining data. Given the data loss resulting from address inconsistencies, the alternative of joining based on the relation number and the postal code emerges as the most fitting option for establishing connections with data that cannot be directly joined to a specific business object. This makes that the total number of joined re-inspections for Centraal Beheer is 232.

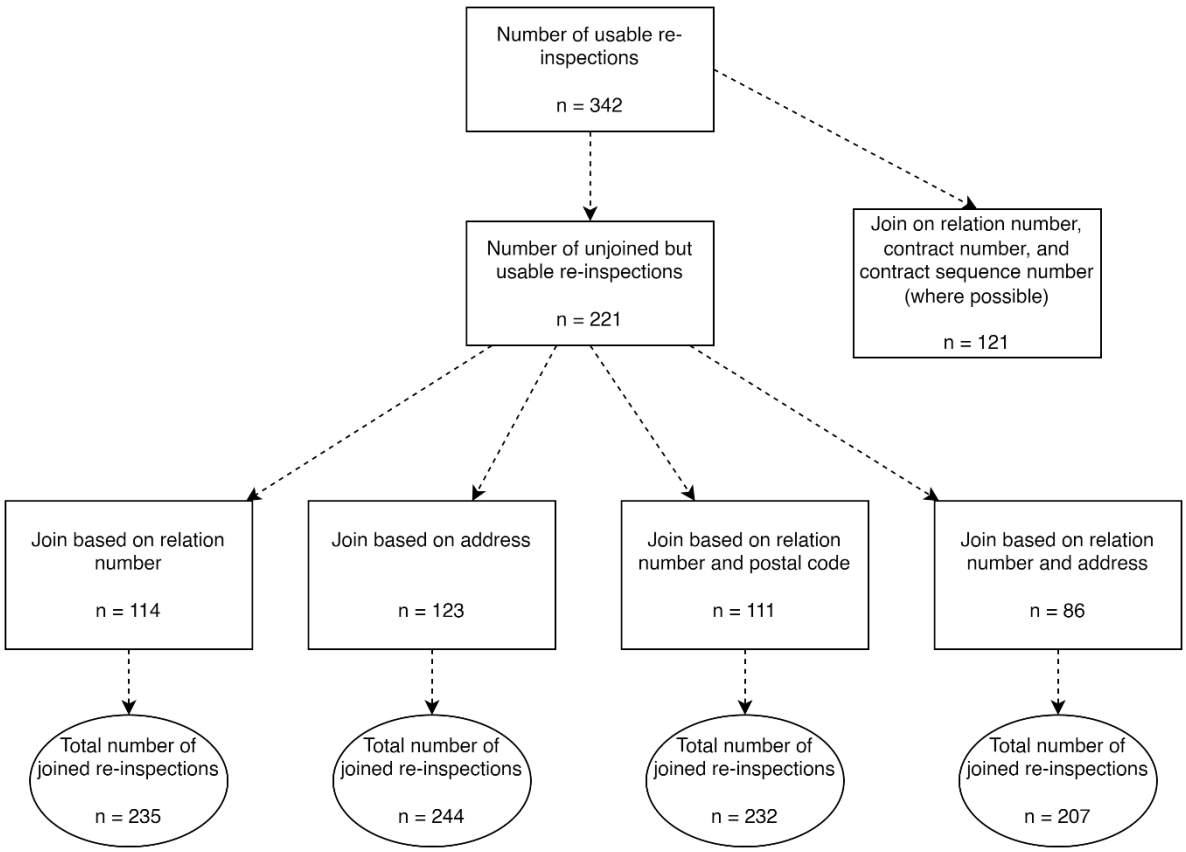

Figure 23: The number of joined re-inspections to policy data for different alternatives for Centraal Beheer

Preparation of re-inspection data for Interpolis

The sample size of 232 is derived from the data linkage between the arena data and Centraal Beheer will be far from sufficient to develop a predictive model. More re-inspections have been carried out for Interpolis in recent years than for Centraal Beheer. Therefore, the re-inspection data from Arena will also be joined to that of Interpolis to have a greater sample size for the purpose of the training of a predictive model. A total of 4601 re-inspections took place for Interpolis, of which 2286 had a usefulness score and 1606 had a usable usefulness score. A schematic overview of this can be found in Figure 24.

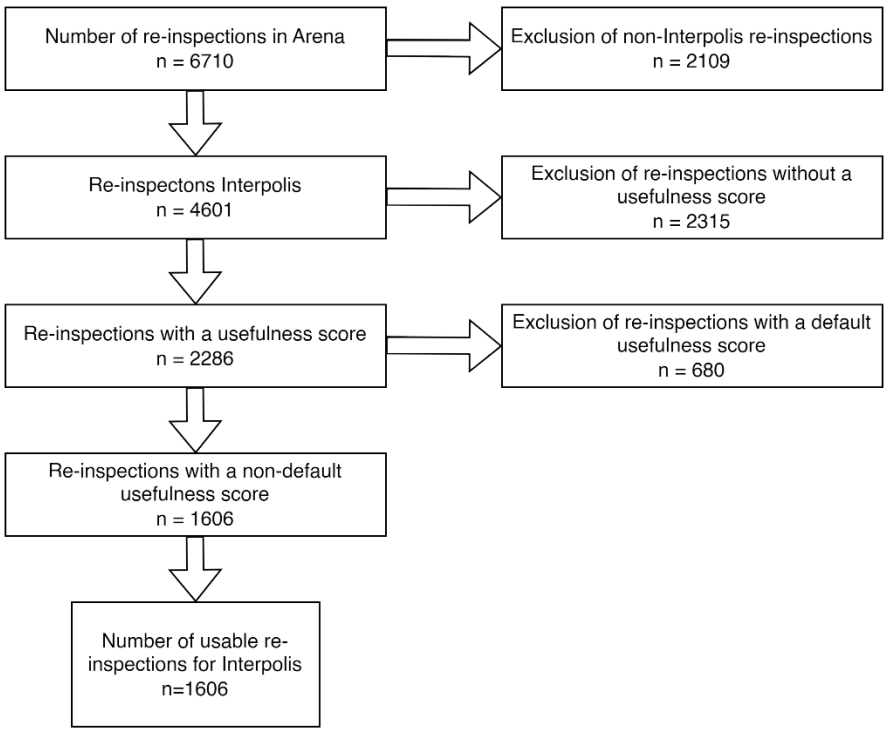

Figure 24: Selection of usable re-inspection data for Interpolis policy re-inspections

Joining re-inspection data to the policy data of Interpolis

The process of joining the re-inspection data to Interpolis policy data is similar to the process of joining the re-inspection data to Interpolis policy data. However, in the case of Interpolis, there is no field in which object sequence numbers are filled in, making a direct join impossible. Consequently, all re-inspection data for Interpolis re-inspections must be linked to Interpolis policy data using alternative identifiers. Figure 25 provides an overview of these alternatives. The alternatives are similar to those for joining Centraal Beheer re-inspection data to Centraal Beheer policy data. This means that the advantages and disadvantages of the alternatives also correspond. The most effective method for joining Interpolis re-inspection data to policy data is by utilizing the policy number and postal code, resulting in a sample size of 817.

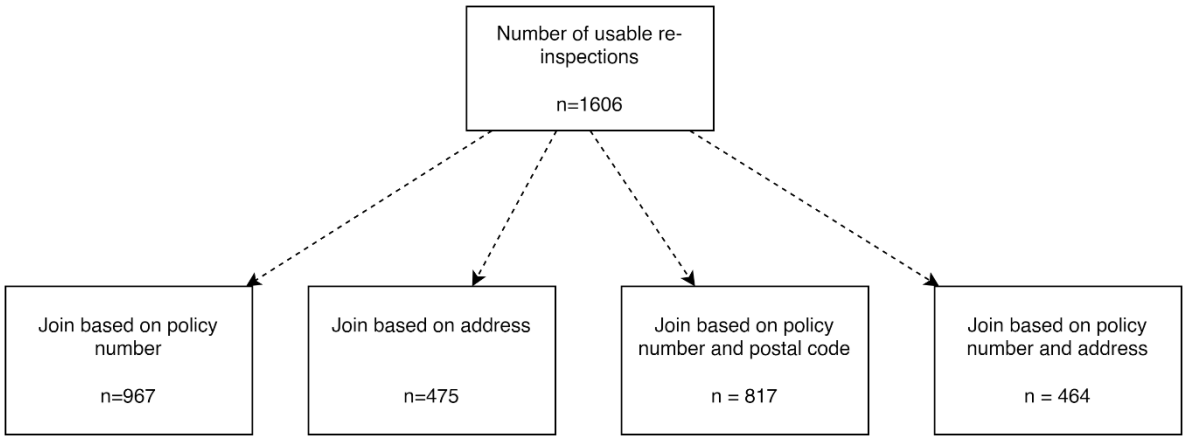

Figure 25: The number of joined re-inspections to policy data for different alternatives for Interpolis

Sample size of re-inspection data

By including data from both Centraal Beheer and Interpolis re-inspections in the development of the

re-inspection usefulness predictive model, the size of the dataset for training and testing the predictive model becomes 1049.

E.4 DATA PREPARATION

In Figure 26, the data pipeline is illustrated that is used to extract data from the databases and transform it to a dataset that can be used to train the predictive model. Figure 26.a displays the preparation of re-inspection data, Figure 26.b demonstrates the preparation of policy, damage, and clause data, and Figure 26.c illustrates the final data preparation where different datasets are joined together.

The Arena system records inspection data for the insurance brands Centraal Beheer, Interpolis, and Avéro Achmea. Initial inspections are also recorded, so not all recorded inspections are re-inspections. The non-re-inspections are filtered out. Initially, the usefulness of re-inspections was not recorded, resulting in the database containing re-inspections without a recorded usefulness score. For this study, only re-inspections with an available usefulness score are considered useful, and re-inspections without a usefulness score are filtered out. Moreover, not all recorded usefulness scores prove to be valuable. It was discovered over time that some risk experts did not input a usefulness score or provide an explanation for the re-inspection usefulness. Consequently, the system defaulted the usefulness to a score of 50. Re-inspections with a usefulness score of 50 and no provided explanation are therefore excluded. If a usefulness score is 50 but an explanation is provided, it is assumed to be entered consciously by the risk expert.

The filtered re-inspection data is then categorized into Centraal Beheer and Interpolis re-inspections. Centraal Beheer re-inspections are included due to the initial focus on Centraal Beheer in the re-inspection prioritization pilot phase. Interpolis re-inspections are included to increase the sample size.

For certain re-inspections conducted by Centraal Beheer, so-called object codes are available for the specific objects that underwent re-inspection, meaning that insured objects can be linked to a re-inspection accurate. This allows for a secure connection between the policy and the re-inspections. These object codes are therefore incorporated into the Centraal Beheer re-inspection data where applicable. Thereafter, the re-inspections of Centraal Beheer and Interpolis are concatenated. Then, for each re-inspection, the closest previous (re-)inspection is sought per re-inspection and, if found, merged. This final step concludes the specific data preparation process for re-inspections.

The policy, damage, and clause data are extracted from the respective policy databases of Centraal Beheer and Interpolis. There are differences in sector categorization semantics between Centraal Beheer and Interpolis policy data. Therefore, a mapping table is employed to achieve standardization. This mapping table is derived from the ongoing SKB+ project and is still in development. Consequently, the reliability of data quality within this mapping table is limited.

The policy data is then joined with the re-inspection data. The specified object numbers are utilized for the join whenever possible, and otherwise, a combination of the relationship number for Centraal Beheer or policy number for Interpolis and the zip code is used for the join. This is because re-inspections that pertain to specific relationship numbers or policy numbers are likely to cover a cluster of buildings with potentially distinct addresses but the same zip code. However, this cannot be determined with certainty, which affects the quality of this data join. Furthermore, the join also incorporates the re-inspection date within the policy timeframe, ensuring that the correct historical context of the policy is joined to the re-inspection.

After the policy – re-inspection join, calculations are made to determine the damage burden, the total amount of damages incurred, and the presence of a clause within a specific timeframe until the re-inspection. These calculations are then merged with the re-inspection and policy data, resulting in a comprehensive dataframe containing integrated re-inspection, policy, damage, and clause data.

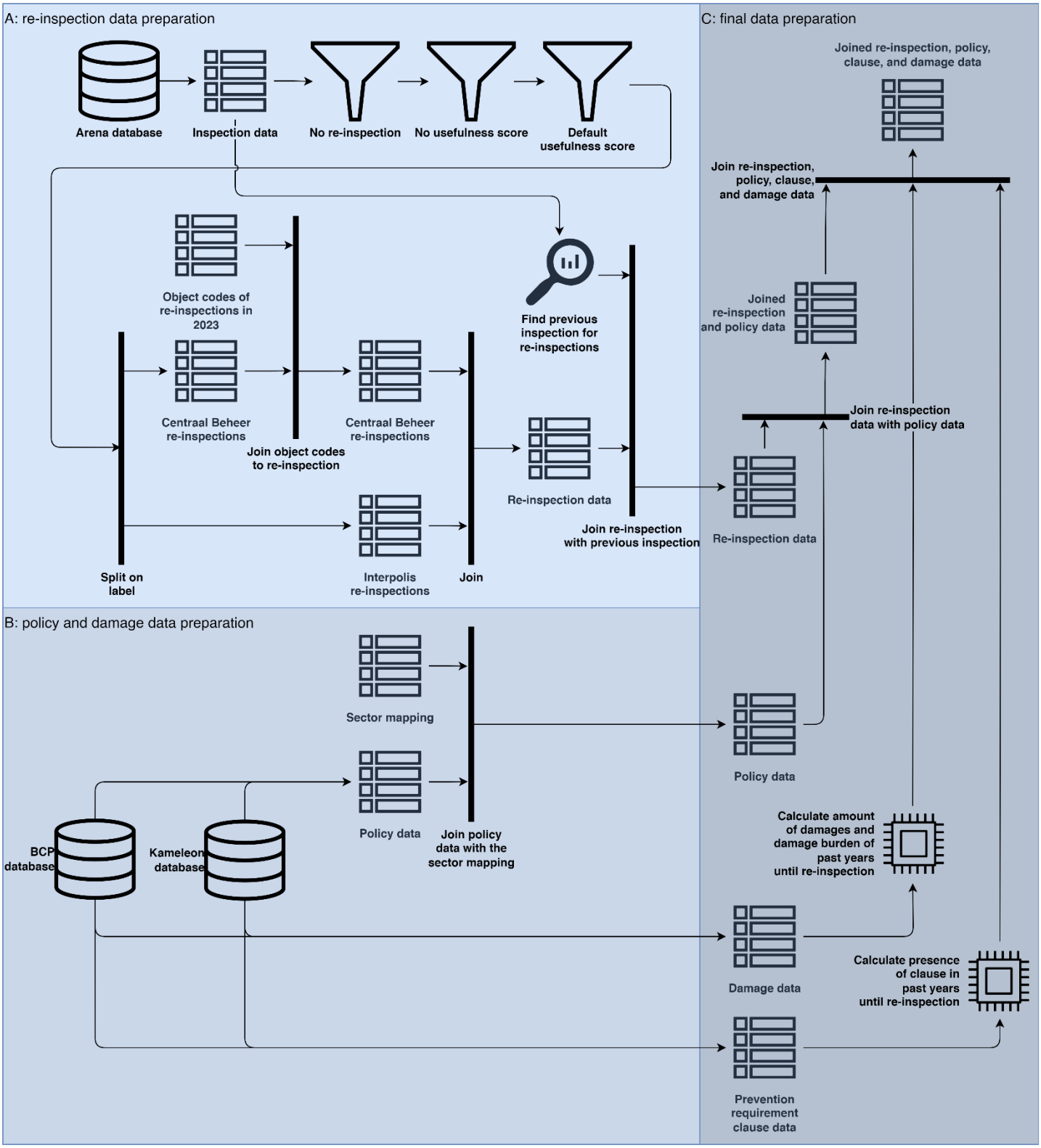

Figure 26: Visualization of data pipeline for the collection and preparation of the data

APPENDIX F: MODEL DEVELOPMENT RESULTS

	Algorithm	Dataset	Accuracy Mean	Accuracy SE	Macro-averaged F1-score Mean	Macro-averaged F1-score SE	Mean Squared Error Mean	Mean Squared Error SE
Analysis with three caterogies	Average as prediction	1	0.5427		0.2345		259.75	
		2	0.5372		0.2330		248.28	
		3	0.5300		0.2310		250.37	
		4	0.5300		0.2310		250.37	
	Multiple regression	1	0.5372	0.0174	0.3711	0.0185	278.17	23.66
		2	0.4814	0.0507	0.3733	0.0510	316.49	34.68
		3	0.5167	0.0100	0.3684	0.0175	268.04	20.87
		4	0.4919	0.0149	0.3660	0.0177	276.93	53.53
	Random Forest regressor	1	0.5153	0.0280	0.3559	0.0311	267.22	26.60
		2	0.4898	0.0626	0.3098	0.0426	256.50	10.78
		3	0.4671	0.0247	0.3749	0.0449	284.71	20.22
		4	0.4824	0.0153	0.3540	0.0286	273.14	28.52
	Random Forest classifier	1	0.4650	0.0125	0.3552	0.0255		
		2	0.4754	0.0466	0.3475	0.0282		
		3	0.4576	0.0240	0.3830	0.0255		
		4	0.4671	0.0191	0.3597	0.0186		
	Neural Network regressor	1	0.5263	0.0338	0.3430	0.0160	273.30	26.83
		2	0.4547	0.0796	0.3631	0.0674	464.94	139.88
		3	0.5043	0.0167	0.3800	0.0413	379.75	125.94
		4	0.4442	0.0133	0.3575	0.0275	386.30	39.19
	Neural Network classifier	1	0.3732	0.0953	0.3212	0.0484		
		2	0.4196	0.0610	0.3428	0.0494		
		3	0.3956	0.0368	0.3680	0.0240		
		4	0.4328	0.0215	0.3852	0.0030		
Analysis with five caterogies	Average as prediction	1	0.5427		0.1407		259.75	
		2	0.5372		0.1389		248.28	
		3	0.5300		0.1386		250.37	
		4	0.5300		0.1386		250.37	
	Multiple regression	1	0.4311	0.0124	0.2318	0.0323	385.03	36.63
		2	0.4421	0.0601	0.2178	0.0306	327.24	42.65
		3	0.4385	0.0129	0.2149	0.0358	432.93	148.22
		4	0.4042	0.0317	0.2286	0.0205	562.40	112.20
	Random Forest regressor	1	0.4573	0.0375	0.2000	0.0217	276.34	32.48
		2	0.4463	0.0361	0.1910	0.0239	275.64	23.32
		3	0.4423	0.0154	0.1965	0.0079	283.99	23.69
		4	0.4614	0.0172	0.2123	0.0325	273.14	24.01
	Random Forest classifier	1	0.4551	0.0279	0.2256	0.0226		
		2	0.4381	0.0515	0.1952	0.0449		
		3	0.4233	0.0184	0.2052	0.0075		
		4	0.4576	0.0208	0.2085	0.0310		
	Neural Network regressor	1	0.4345	0.0880	0.1873	0.0358	296.93	36.64
		2	0.4381	0.0332	0.2201	0.0242	484.52	156.71
		3	0.4747	0.0238	0.2340	0.0173	335.78	55.86
		4	0.4271	0.0272	0.2210	0.0260	416.90	104.28
	Neural Network classifier	1	0.2998	0.0656	0.1990	0.0250		
		2	0.4445	0.0823	0.2668	0.0721		
		3	0.3327	0.0172	0.2203	0.0054		
		4	0.3775	0.0297	0.2193	0.0175		

APPENDIX G: INTERPRETABILITY SCENARIOS

Interpreteerbaarheid

Scenario 1: je krijgt alleen de prioritering te zien

Bedrijf	Prioritering
X	1.
Y	2.
Z	3.
...	...

De prioriteringslijst

Interpreteerbaarheid

Scenario 2: je krijgt de prioritering en de voorspellingen te zien

Bedrijf	Prioritering	Voorspelde schadelast	Voorspelde schadekans	Voorspelde nuttigheid
X	1.	€3.900.000	10 schades	Nuttig
Y	2.	€1.950.000	6 schades	Nuttig
Z	3.	€400.000	16 schades	Zeer nuttig
...	...	...	...	...

De prioriteringslijst | De voorspellingen in het prioriteringsmodel

Interpreteerbaarheid

Scenario 3: je krijgt de prioritering, de voorspellingen en de factoren die aan de voorspelling bijdragen te zien

Bedrijf	Prioritering	Voorspelde schadelast	Voorspelde schadekans	Voorspelde nuttigheid	Vorige nuttigheid	Vorige score management	Vorig brandrisico	...	Schadelast afgelopen jaar
X	1.	€3.900.000	10 schades	Nuttig	onbekend	A	B	...	€850.000
Y	2.	€1.950.000	6 schades	Nuttig	onbekend	B	A	...	€0
Z	3.	€400.000	16 schades	Zeer nuttig	35	A	C	...	€0
...	...	...	...	...	...	...	...	...	...
De prioriteringslijst		De voorspellingen in het prioriteringsmodel			De gebruikte parameters voor de nuttigheidsvoorspelling				

Interpreteerbaarheid

Scenario 4: je krijgt de prioritering, de voorspellingen en de berekening voor de voorspelling te zien

Bedrijf	Prioritering	Voorspelde schadelast	Voorspelde schadekans	Voorspelde nuttigheid	Vorige nuttigheid	Vorige score management	Vorig brandrisico	...	Schadelast afgelopen jaar
X	1.	€3.900.000	10 schades	46 = Nuttig	0.3 x 0	A: 10	B: 12	...	0.001 x €850.000
Y	2.	€1.950.000	6 schades	53 = Nuttig	0.3 x 0	B: 15	A: 8	...	0.001 x €0
Z	3.	€400.000	16 schades	34 = Zeer nuttig	0.3 x 35	A: 10	B: 12	...	0.001 x €0
...	...	...	...	...	...	...	...	...	...
De prioriteringslijst		De voorspellingen in het prioriteringsmodel			De modelberekening voor de nuttigheidsvoorspelling				

Interpreteerbaarheid

Scenario 5: je krijgt de prioritering, de voorspellingen en een voorspelde reden voor de voorspellingen te zien

Bedrijf	Prioritering	Voorspelde schadelast	Voorspelde schadekans	Voorspelde nuttigheid	Voorspelde reden voor de nuttigheid
X	1.	€3.900.000	10 schades	Nuttig	Hoog risico
Y	2.	€1.950.000	6 schades	Nuttig	Risicobewustzijn van klant kan worden vergroot
Z	3.	€400.000	16 schades	Zeer nuttig	Preventie is noodzakelijk
...	...	...	...	...	...
De prioriteringslijst		De voorspellingen in het prioriteringsmodel			De voorspelde reden

Interpreteerbaarheid

Scenario 6: je krijgt de prioritering, de voorspellingen en de factoren die het meest hebben bijgedragen aan de voorspellingen te zien

Bedrijf	Prioritering	Voorspelde schadelast	Voorspelde schadekans	Voorspelde nuttigheid	Belangrijkste factoren in voorspelde nuttigheid
X	1.	€3.900.000	10 schades	Nuttig	Vorig brandrisico = B Verzekerd bedrag = €14.500.000
Y	2.	€1.950.000	6 schades	Nuttig	Vorige managementscore = B
Z	3.	€400.000	16 schades	Zeer nuttig	Vorig brandrisico = C Vorige nuttigheid = 35
...	...	...	...	...	...
De prioriteringslijst		De voorspellingen in het prioriteringsmodel			De belangrijkste factoren