Universiteit Leiden

Designing a holistic method for enhancing data quality with the use of machine learning:

A master thesis for ICT in Business & the Public Sector at Leiden University

Master Thesis

Vincent Buitenhuis

July 25, 2023

Supervisors:
Drs. J.B. Kruiswijk (LIACS)
Drs. N. van Weeren (LIACS)
LIACS, Leiden University

Abstract

With the dawn of Big Data and data-driven businesses, the concept of data quality is becoming increasingly important. It is an important element within the process leading up to data-driven decision making (DDDM), potentially being a decisive factor in the outcome of the decision. A multitude of methods exist for assessing, improving and controlling data quality. One of these methods is through the use of Machine Learning (ML). As common researches on this topic investigate the use of ML for isolated data quality dimensions or issues, we have designed a holistic method for enhancing multiple aspects of data quality with the use of ML-based techniques.

As part of this thesos research, we have first conducted an extensive literature review on the topics of data-driven businesses, Big Data, data quality and ML. Analysing an assembly of different scientific sources has allowed us to elicit requirements for our proposed method design. We have subsequently developed a proof-of-concept in the form of a working prototype, HAQIM (**H**olistic **A**ssessment of **Q**uality and **I**mprovement **M**ethod).

We have validated our proposed method by conducting experiments with 6 different datasets and 3 scenarios: the dataset including the raw data quality issues, a dataset enhanced through manual methods and a dataset enhanced with HAQIM. All datasets and corresponding scenarios have been used within predictive analytical modelling. We have measured the performance metrics of the models using each dataset, including RMSE and r^2. The majority of datasets have experienced an increase in predictive performance after being processed by HAQIM. Four datasets have seen a positive change in RMSE, and five a positive change in r^2. These experiments have validated our proposed method to be effective for enhancing data quality.

Acknowledgements

As adding a page dedicated to acknowledgements appears to be some sort of tradition, I will join in on the fun.

The finalization of this thesis research project marks the end of my education. After eight years of primary school, seven years of high school and seven years of university I have spent a total of twenty-two years being educated. During this time, a multitude of people have spent time, energy and care into getting me through it all. I want to start off by thanking my teachers in primary school. Honestly I do not remember many teachers, except for Henk. So I would like to thank Henk for being a great and kind teacher.

Contrary to my memories of primary school, I remember my time in high school a lot better. I have attended Het Vlietland College in Leiden, in my opinion one of the most beautiful cities to ever exist upon the face of the earth. During my time in high school, I have had a great deal of memorable experiences, met a lot of good people and got taught by great and inspiring teachers. One of these teachers that really stood out has been Pim Versteeg, from geography class. His way of telling stories to teach different subjects has activated a general curiosity in me about the world and how things worked. Typically during high school one has the opportunity to explore many aspects of the possibilities of your educational adventure, as well as other experiences in life beyond school. For me, I have cultivated my curiosities and understanding of what sciences I want to pursue after I would have graduated from high school.

I started my university career at TU Delft, spending merely three months studying Civil Engineering. After working a few different jobs I picked up studying again in the next year. This time I enrolled in the study Computer Science & Economics at Leiden University. During this bachelor study I have had the chance to delve deep into a lot of different subjects and ideas. At times, this study has been quite tough, having spent countless hours studying, programming or researching. Sometimes stretching out far into the night. For this, I want to give a special shout-out to Indian YouTubers for being able to explain so many (even awfully specific) subjects when I struggled to comprehend certain material. Also, I have had the chance to gain a lot of new friends and meet inspiring people during this time. After officially completing my bachelor and gaining the honour to place a framed diploma in my home, I continued my education with the master study ICT in Business & the Public Sector, at Leiden University as well. During this final phase of my entire education I have yet again had the chance to meet many

inspiring and unforgettable people. This master study met its conclusion with a thesis research project, of which this pdf is the trophy.

In relation to this thesis itself, I would like to give thanks to my university supervisors: Drs. Bas Kruiswijk and Drs. Niels van Weeren. Bas and Niels have challenged me on various aspects of my thesis, which has allowed me to perfect the definition of my topic and research overall. Additionally, I would like to thank my thesis supervisor at PwC, where I have written the thesis as part of an internship: Reinier van Doggenaar. Furthermore I would also like to give thanks to the CIO Advisory team for their hospitality, help and opportunities to assist in various projects during the internship. During my time at PwC I have learned a lot about various aspects related both to the professional field and personal growth, for which I am both thankful and grateful.

I want to give a special thank you to everyone I have had the opportunity to meet and got to know during these entire twenty-two years. I want to thank my family members, friends and girlfriends during this time. Reflecting on the good times and on the bad times, life comes as it is, as life itself is poetry. Some times you are the poet, and some times you are not.

"It's the possibility of having a dream come true that makes life interesting."
- Paulo Coelho, The Alchemist

Contents

Chapter 1

Introduction

Data quality as a concept is an important aspect in keeping data healthy. It can be defined as "Fitness for use"[1], which is also relative to the purpose of the data in question. Data quality is a multi-dimensional concept, with on one hand the subjective quality measurement, representing the manner in which data fulfils the needs of stakeholders. While on the other hand, there is the objective quality measurement, which refers to the state of data which can be assessed regardless of its context in terms of business objectives.[2]

In the year of 2007, the world was shook by a global financial crisis. It has caused economic distress in a great number of countries across the world, with long-lasting effects on international and local markets.[3] A series of events triggered by several large financial firms have led to a meltdown of the global financial system.[4] Notable events included the burst of the housing bubble and the bankruptcy of the Lehman Brothers. The root cause can be attributed to a multitude of factors. However, one major contributor of the financial crisis was poor data quality. An examination of the role of data quality by Francis & Prevosto (2010) adduce the example of incomplete and inaccurate records within mortgage data.[5] Due to a lot of data being missing and inaccurate, the risk associated with a number of mortgage loans could not be assessed properly. This examination also states that complete and accurate data could have been used in predictive modelling which could have forecast the crisis developing.

The financial crisis of 2007 has been one example showing how bad data quality could have catastrophic effects on the decision making process in global financial systems. It is also vital for the typical enterprise, as strategic decision making based on data analytics has become a commonplace practice in modern businesses.[6] The approaches taken by these businesses usually take on the form of exploratory, inferential, causal, descriptive, predictive

and prescriptive data analytics.[7,8] These approaches as part of the decision making process are one of the key activities typically found in data-driven businesses,[9] and if used successfully, yield strategic business value.[10] However, insights derived from analytical capabilities can only be as good as the data it consumes.[11] The quality of data used for strategic analytical capabilities is vital for the decision making procedure of a business, as a high degree of data quality leads to better decision effectiveness.[12] On the other side, poor data quality leads to less effective decision making and a reduced ability to execute strategies.[13]

To combat the negative consequences of poor data quality, it should be improved accordingly. There are many challenges in improving data quality, such as the large volume of data streams, increasing diversity of data sources and a lack of unified data quality standards.[14] When looking at the technical aspect of improving data quality, a common approach is to tackle data quality issues manually. Gudivada & Ding (2017)[15] state that many organizations still rely on this approach, which can be very time consuming and labour intensive. Another approach is to apply Artificial Intelligence (AI), and more specifically Machine Learning (ML) based methods. Practical examples have been described by a research conducted by Shi et al. (2015)[16] who have applied a Support Vector Machine (SVM) to tackle the data quality issue of missing entries on power grid monitoring data. As ML has many practical uses, it also offers a lot of potential in the field of data quality improvement.

1.1 Research Problem

Data is everywhere. It is estimated that the global amount of data being created and stored in 2025 is 175 Zettabytes,[17] which is equal to 175 trillion Gigabytes. This widespread prevalence of data creation is nothing new to businesses: a topic surrounding most major businesses in all industries is data-driven business transformations. A data-driven business transformation is a process in which an organization integrates data within their (core) business processes, including analytical capabilities.[18] Such transformations are evolving from a state of novelty to becoming the norm. This can be supported by the fact that in 2006, only one of the ten largest companies in terms of market capitalization took a data-driven approach, to six of these companies being data-driven in 2017.[19]

To enable analytical insights, organizations are making use of large-scale data, also known as Big Data. Big Data is a concept, while its formal definition is not agreed on by scholars,[20] which is described by De Mauro (2015) as "*the Information asset characterized by such a High Volume, Velocity and Variety*

to require specific Technology and Analytical Methods for its transformation into Value."[21] The use of Big Data within core business processes to create value, makes this data a strategic asset. The use of big data analytics is becoming more widespread within businesses. The global market value of big data analytics solutions has been estimated at 29 billion USD in 2016, and has been expected to grow to a total of 67 billion USD in 2021.[22] A survey conducted by Harvard Business Review, sent to Fortune 1000 companies, showed that 40.8% of business executives claim their businesses are competing on big data analytics.[23] Besides this, the use of big data analytics has shown to have a positive impact on the overall firm performance,[24] as well as being a positive contributor to the valuation of firms.[25]

Poor data quality can lead to negative business results as well. On an operational level, it is estimated from research in 1998 that the cost of poor data quality ranges between 8% and 12% of an organization's revenue.[26] In the United States alone, poor data quality has been associated with a direct annual cost of billions of dollars.[27] Modern day estimates suggest the current-day loss in revenue in the United States total at around 3.1 trillion US Dollars.[28] On a strategic level, poor data quality leads to less effective decision making and a reduced ability to execute strategies.[13] Besides being a large issue for commercial businesses, data quality is an important factor in other fields as well. For example, health care organizations experiences problems with bad data quality as well. Charnock (2019)[29] states that data quality issues such as missing, illegible and incomplete data results in a decrease in the quality of care being delivered by health care professionals.

In consideration of the aforementioned, we can see that more businesses are adopting the concept of being data-driven, in which big data analytical capabilities are utilised for strategic decision making. A significant hindrance in the effectiveness of this process is poor data quality. Thus, it is vital to improve the overall quality of strategic asset data. AI solutions have opened up new possibilities for using ML techniques to enhance data quality, which will be explored during this research.

1.2 Research objective

The aim of this research is to explore the possibilities of ML techniques for enhancing data quality. In order to reach this goal, the research stage of this thesis can be divided within two segments. Firstly, the data quality issues existing within the given context will be studied. This includes establishing the behaviour and characteristics of data within this context. Secondly, the application of different ML techniques to enhance data quality will be inves-

tigated. These efforts will culminate in the main deliverable resulting from this thesis project, which is a method design. The primary goal is to create a working prototype, which can be applied towards a real data set for data quality assessment and enhancement, mainly based on ML methods.

In order to translate the research objectives into concrete points, the following research question (RQ) and related subquestions (SQ) are formulated:

- **RQ:** How can we develop a novel method based on machine learning techniques to effectively improve multiple data quality issues at once, for data found in a data-driven context mainly used for analytical purposes?
  - **SQ1:** What are the characteristics of data typically used for analytical, specifically for predictive purposes within a data-driven setting?
  - **SQ2:** What common data quality issues exist within the context of data being used for analytical and specifically predictive purposes within a data-driven setting.
  - **SQ3:** What machine learning techniques can be effectively applied to tackle data quality issues commonly found within the described context?
  - **SQ4:** How can we apply Machine Learning algorithms to address specific data quality issues in a holistic manner?
  - **SQ5:** How can a prototype be constructed which implements this approach?

1.3 Research scope

Data quality in itself is a broad concept. The many uses of data has resulted in the need to address the concept of data quality from multiple aspects, such as measurement, assessment, management, monitoring and enhancement.[15] Additionally, within scientific literature, there are different approaches to study data quality: intuitive, theoretical and empirical.[30] As data quality issues can be highly context sensitive, depending on the industry, business and use of the data in which the issues occur, it is difficult to employ an all-encompassing approach to study data quality in this thesis. In addition, time and resource constraints exist for the thesis project trajectory, which influences the research and application development phases. Thus, we will define the scope of this research accordingly.

During this research we will focus on data quality issues that occur within the context of data being used as a strategic asset by businesses, in specific strategic decisions based on the outcomes of analytical modelling. These are capabilities typically found within data-driven organizations. For this research, we will narrow the focus down to the case of predictive modelling. In addition to this, we will only take data quality issues into account which can be solved regardless of context. As the goal of this research is to investigate how ML techniques can be applied to data in any business, we must disregard quality issues which exist in the context of specific business rules or other context-sensitive issues.

The idea of this research is to generally focus on enhancing the data quality of data typically found within the pipeline leading to data-driven decision making. The design of our data quality enhancement method will have this type of data in mind. During our literature review we will study the characteristics and behaviour of such data, which will allow us to tailor the method specifically to this type of data.

The design and development of a data quality enhancement method will also be subject to a scope definition. As the primary aim of the development stage is to create a working prototype, building a comprehensive application will exceed the time and resource constraints. This implies that a limited number of features will be constructed during the entirety of this thesis project.

Furthermore, this thesis has been written as part of an internship trajectory at PwC NL. PwC is a multinational network which offers professional services to various types of organizations, including accounting, auditing and advisory. PwC NL is the denomination of the Dutch office of this network. The internship has been accommodated by the CIPS Technology CIO Advisory team, which consults on IT-related topics such as ERP systems, data management and business processes.

1.4 Academic contribution

The main academic contribution of this research is to strengthen the integration between AI and strategic decision making. As more businesses are relying on data-driven decision making, bridging this gap could potentially result in significant implications for business success. This will be done in two ways. Firstly, constructing a method to combat data quality issues, based on ML techniques, provides a more practical contribution. Secondly, communicating valuable insights on the overall quality of given data enables organizations to understand their data more thoroughly. These aspects will

advance the effectiveness of data-driven decision making of businesses. Besides commercial organizations, the principles learned from this research could also be transferred to other domains. Examples of this are health care or the public sector, as these domains typically make use of data-driven strategic decision making.[31,32] Overall, this study has the potential to contribute both theoretically and practically to the integration between AI and strategic decision-making.

Concerning the novelty of this thesis, no known researches have investigated a holistic method to enhance data quality in the context of data used for strategic decision making, specifically based on insights from analytical capabilities, making use of primarily ML techniques.

Chapter 2

Research design

This chapter outlines how the research for this thesis project has been designed; describing which methodologies have been applied, which approaches have been taken and what resources have been used. The main methodology on which the research has been based is Design Science Research Methodology (DSRM). DSRM is a problem solving approach which seeks to create innovative artifacts, with the intention to improve technology and scientific knowledge.[33] The general idea is that these artifacts have the purpose of improving and solving problems within the context they are instantiated.

Peffers et al. (2007) have provided an overview of the DSRM process specified when applied to creating information systems in a matter of activities.[34] These activities are presented in the form of stages which should be performed in order. These stages are as follows.

1. **Problem identification:** This initial stage involves the identification of a real-world opportunity or problem to which the creation of an artifact will provide a solution. This stage includes justifying the research and creating an understanding on the problem. Section 1.1 of this thesis has interpreted the research problem for the case of creating a method that employs ML techniques to enhance the data quality of data within the given context. As part of this activity, a literature review has been conducted in order to examine the properties of the environment the artifact is instantiated, as well as the current state of knowledge in the field.
2. **Defining the objectives for a solution:** This stage involves the goal- and objective-setting steps of the research. This includes giving a definition of what is possible and (technologically) feasible. Within this thesis, this activity has been covered by sections 1.2 and 1.3. It has also been

supplemented with the knowledge presented in the literature review, used to set the stage in terms of scientific knowledge and technological possibilities.

3. **Design and development:** This activity involves the creation of an artifact, in the form of any designed object that contributes to research. The functionalities and architecture of this artifact are determined before the actual development. In the case of this thesis research, we have designed a method with a corresponding architecture. A proof-of-concept has been developed in the form of a working prototype, which can be applied to any real-world dataset. This stage has been addressed in section 4.

4. **Demonstration:** To prove the idea generated from the research actually works, a demonstration of the artifact is necessary to show its functionality and effectiveness in a real-world setting. Our prototype will be tested adequately in a simulated setting on real-world data. The prototype has been described in section 5 of this thesis.

5. **Evaluation:** This activity involves observing and measuring the degree to which the artifact provides a solution to the described problem. In section 2.4 we have addressed this stage. We have used an empirical method to evaluate the performance of the prototype, by comparing the quality of the enhanced dataset to an unenhanced dataset, if used for strategic decision making.

6. **Communication:** The final activity of the DSRM process includes communicating the problem and its justification, the design of the artifact, the artifact itself, its utility, novelty and its effectiveness to all relevant audiences. The audience in question includes researchers of the topic of data quality, professionals encountering data quality subjects within their work as well as people with a general interest in big data, data quality and ML. Communicating the aforementioned aspects has been executed by making this thesis publically available, releasing the source code of the prototype, presenting (intermediate) results to professionals and presenting the results in an academic setting.

Considering the DSRM theory, a general research design has been constructed. The research starts with a literature review, in which the first three research subquestions have been addressed, followed by the design of the holistic data enhancement method, which has addressed the fourth subquestion. Then, we have developed the prototype, which has resulted in a working prototype and the answer to the fifth subquestion. Finally, after the validation phase the final thesis was delivered, along with the answer to the main research question. Figure 2.1 visualizes the general research process, including the major phases and main deliverables.

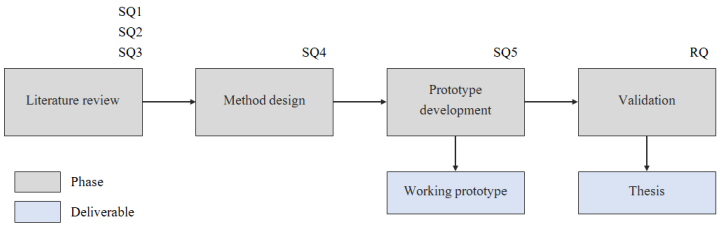


Figure 2.1: Overview of the research design

2.1 Literature review

The first phase of the research process is the literature review. During this phase, we have reviewed a number of scientific literature with the purpose of collecting data to ultimately base the method design on, and to set the background in terms of scientific knowledge. Figure 2.2 provides a visualisation of the literature review phase, including its integral steps. We have firstly identified the premises of a data-driven organization and subsequently the behaviour of data within such a context, by studying big data and how it could be defined by its corresponding aspects. This allowed us to answer the first research subquestion.

Secondly, the concept of data quality and its practical implications have been studied from literature. This has been done by consolidating which dimensions of data quality should be considered, and what data quality issues are typically encountered. This allowed us to answer the second subquestion.

Thirdly, based on the findings of the previous steps in this phase, we have investigated which ML algorithms are often used to address specific data quality issues that exist within the defined scope. Meaning, data quality issues which are typically found in a data-driven environment and ultimately used for strategic decision making, which can be solved through ML based techniques regardless of context.

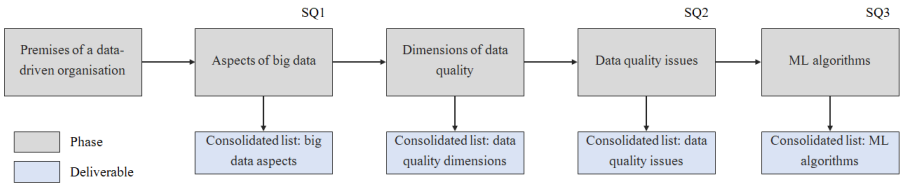


Figure 2.2: Overview of the literature review phase

2.2 Method design

The second phase of the thesis research entails designing a holistic method. For this process we have applied the practices suggested by Requirements Engineering (RE) theory. The choice to integrate RE principles in the design process was based on the desire to create a method, which lays the blueprint for quality software products. The importance of RE in upholding quality can be supported by a survey across 12 UK companies which shows that 48% of software problems can be attributed to requirements.[35]

RE is described by Aurum & Wohlin (2005) as refering to *"All life-cycle activities related to requirements. This includes mainly gathering, documenting and managing requirements."*[36] The activities which are included are typically the elicitation, analysis, validation, negotiation, documentation and management of requirements.[37] Instead of a linear process, these activities are presented in a cyclic manner, as visualized in figure 2.3. While the main principles of the RE process are integrated within the design process, the RE method as described within literature will, however, not be followed as rigorously. The main driver behind this decision lies within the time and resource constraints attributed to the entire thesis project.

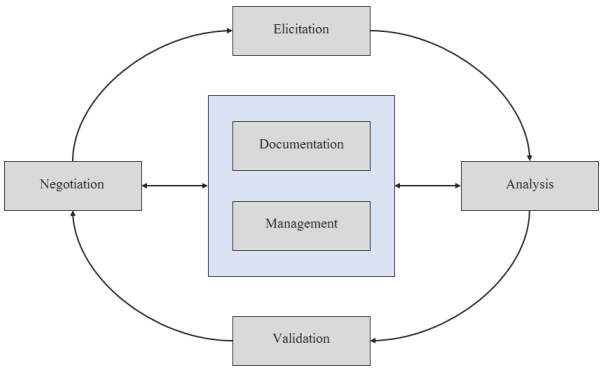


Figure 2.3: The Requirements Engineering process cycle

Thus, for this research, we have decided to take a more simplistic approach for arriving to a final architecture. Firstly, we have started by inventorizing all data gathered from the literature review. Besides eliciting requirements from analysing literary sources, we have applied a prototyping approach as well by constructing a basic program containing simple functionalities to explore the feasibility and challenges of said method. We have examined the technical and functional aspects which should be taken into account when designing the envisioned method. Based on this information all requirements

will be established, and subsequently documented by creating a consolidated list of requirements.

An architecture will be composed based on the requirements which have been established during the previous steps of the design stage. A high-level architecture of the method in its entirety will be defined using the ArchiMate framework. The lower level architecture, entailing the main components within the architecture, will be subsequently modeled using UML class diagrams.

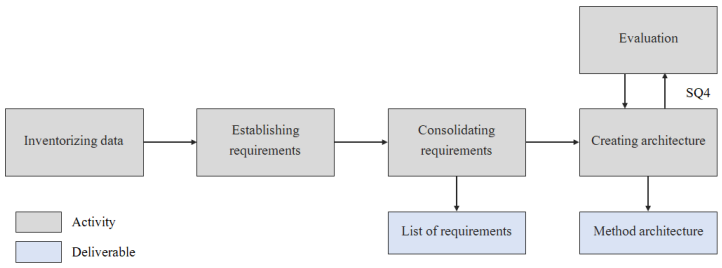


Figure 2.4: Overview of the method design phase

2.3 Prototype development

The third phase in the general research process is the development of an actual working prototype. Figure 2.5 offers an overview of the main flow of activities of this phase. As the figure shows, the design of this phase is rather simplistic. We have started with the development activities based on the architecture.

The second activity includes testing the product with curated testing data, which has been altered to simulate the environment in which the software would be deployed. After all testing has been concluded, the main deliverable of a working prototype has been produced. The programming language which has been used across the entire development of the prototype is Python. The core functionalities have been customly developed, while some others have been imported using publicly available packages, e.g. for the implementation of ML algorithms.

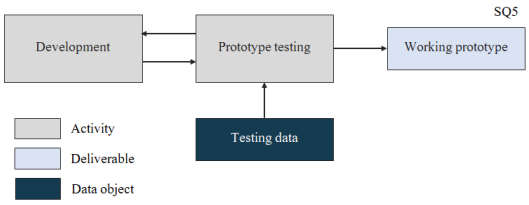


Figure 2.5: Overview of the prototype development phase

2.4 Validation

The concluding phase of the research process is validating the prototype tool which has been produced based on our method. Figure 2.6 provides a visualization of this phase. This phase involves determining the effectiveness of the prototype, by comparing different scenarios representing possible real-world usage, using 6 different datasets. Firstly, we have created three different scenarios each consisting of a types of data object for each data set:

1. **A raw dataset:** This data is based on the raw, original datasets we have retrieved. Additionally, they contain artificially created data quality issues.
2. **A manually enhanced dataset:** This dataset will be enhanced using basic, non-ML methods. Missing data will be imputed using a simple median imputing strategy. Irrelevant data will be filtered by finding the top 5 attributes correlating with the target attribute.
3. **A dataset improved by our ML data enhancement method:** This dataset will be improved on the data quality issues of missing and irrelevant data making use of ML-based techniques provided by the prototype we have created.

For all datasets available for our research, we have created 3 variants in which these scenarios are represented. In order to test the effectiveness of each method, their performances in actual analytical modelling would have to be measured. So for each scenario, of each dataset we have applied a simple linear regression to predict a target variable. The accuracy of these models, in terms of predictive performance has been measured. Based in the performance metrics, we were able to determine and compare the effectiveness of the different methods for handling data quality issues.

2.4.1 Data

In order to adequately test and validate the effectiveness of the tool, we have used real-world datasets. The simulative effort for the testing and validating phases of our research requires the use of such datasets. For these activities,

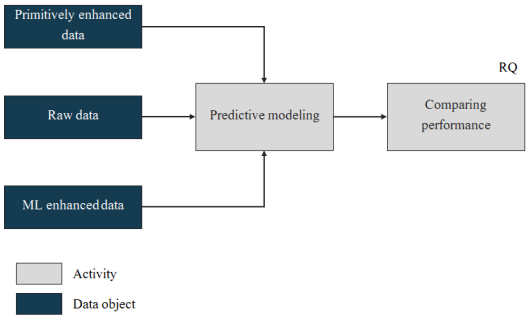


Figure 2.6: Overview of the prototype validation phase

we have acquired 6 different datasets. An elaborate description of the datasets that have been used for this research can be found in appendix A. A short description of the acquired datasets are as follows.

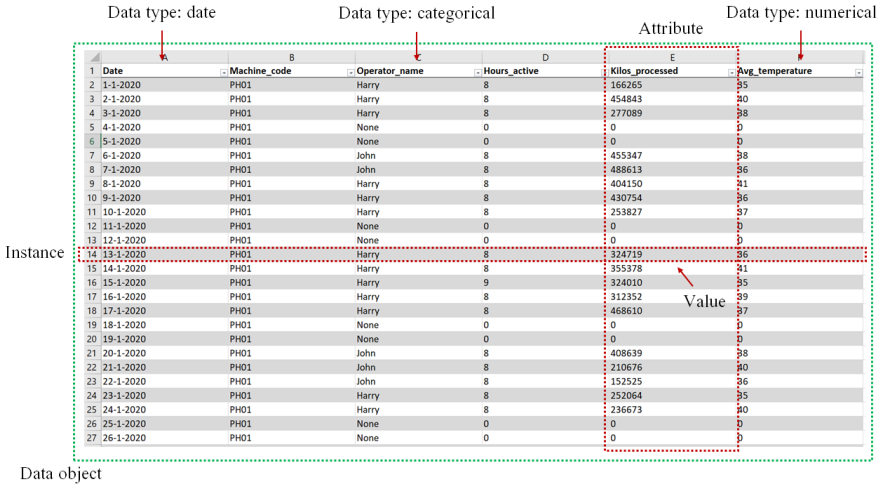
1. A dataset of a Dutch network operator containing aggregated data per postal code containing multiple attributes about the (type of) electricity connections.
2. A dataset containing attributes about the behaviour of website visitors and sales data of a webshop specialized in travel accessories.
3. A dataset containing the daily prices of different types of fuel in the Netherlands.
4. A dataset containing attributes related to the behaviour and type of website visitors of the website of a Dutch chain store specialized in selling bicycles.
5. A dataset containing daily avocado prices and related data in the United States.
6. A dataset containing information on multiple restaurants in India and their corresponding average food delivery time.

The datasets we have received for our research have been carefully curated in terms of data quality up to the point of retrieval. In order to apply and test our prototype, the input data is required to contain some form of data quality issues. Thus, we have added artificially created data quality issues to all datasets, in a structured manner.

2.4.2 Definitions

Within this thesis, we have mentioned various components within the concept of data itself. We have used a variety of terms to describe these components. In this section, we have listed a brief overview of the terms and their corresponding definitions we have used frequently throughout this thesis. In

figure 2.7 we have visualized these terms with a example dataset, which contains operating data of a factory machine.



	A	B	C	D	E	F
1	Date	Machine_code	Operator_name	Hours_active	Kilos_processed	Avg_temperature
2	1-1-2020	PH01	Harry	8	166265	35
3	2-1-2020	PH01	Harry	8	454843	40
4	3-1-2020	PH01	Harry	8	277089	38
5	4-1-2020	PH01	None	0	0	0
6	5-1-2020	PH01	None	0	0	0
7	6-1-2020	PH01	John	8	455347	38
8	7-1-2020	PH01	John	8	488613	36
9	8-1-2020	PH01	Harry	8	404150	41
10	9-1-2020	PH01	Harry	8	430754	36
11	10-1-2020	PH01	Harry	8	253827	37
12	11-1-2020	PH01	None	0	0	0
13	12-1-2020	PH01	None	0	0	0
14	13-1-2020	PH01	Harry	8	324719	36
15	14-1-2020	PH01	Harry	8	355378	41
16	15-1-2020	PH01	Harry	9	324010	35
17	16-1-2020	PH01	Harry	8	312352	39
18	17-1-2020	PH01	Harry	8	468610	37
19	18-1-2020	PH01	None	0	0	0
20	19-1-2020	PH01	None	0	0	0
21	20-1-2020	PH01	John	8	408639	38
22	21-1-2020	PH01	John	8	210676	40
23	22-1-2020	PH01	John	8	152525	36
24	23-1-2020	PH01	Harry	8	252064	35
25	24-1-2020	PH01	Harry	8	236673	40
26	25-1-2020	PH01	None	0	0	0
27	26-1-2020	PH01	None	0	0	0

Figure 2.7: Overview of the prototype development phase

- **Data object:** Also referred to as a dataset. It is a single object containing a set of grouped data. It can appear in many forms, for example in a tabular format as shown in our example.
- **Data instance:** Group of data encompassing multiple columns in a dataset, corresponding to a single point of measurement. Also known as a row.
- **Attribute:** Group of data encompassing multiple rows in a dataset, corresponding to multiple points of measurement. Also known as a column.
- **Value:** Intersection point of a data instance and attribute, containing a single sample of information.
- **Data type:** A classification of data corresponding to the sort of data. Examples are dates, categorical data or numerical data.

Chapter 3

Literature review

In this chapter, we provide an overview of a multitude of peer-reviewed scientific literature. We have reviewed literature within the field of Data Management and Machine Learning. Within these fields, the included subjects vary from data-driven organizations to big data and data quality. In addition to peer-reviewed articles, we have also included credible company reports as a data source. This literature review has a two-headed purpose. Namely first for delineating the background in terms of scientific knowledge and secondly for eliciting information relevant for the requirements included in the method design.

First, we will describe the environment and behaviour of data used as a strategic asset by investigating articles related to big data and data-driven organizations. Then, we will review the subject of data quality, so all the relevant information related to this subject within the research scope will be stated. Finally, the subject of ML will be thoroughly explained, as well as the ML techniques that are relevant for the system design.

3.1 Big Data

Big Data is a concept that has been in the spotlight the past few years. As part of the wider digitization trend, it has seen an abundance of attention from the scientific community, with 36,821 publications mentioning Big Data as a concept as of 2020.[38] It has been believed Big Data has been first mentioned as a term in the mid-nineties, and remaining relatively outside of the public eye until 2011, when its global interest took off.[39] While no clear, concise and agreed-upon definition of Big Data exists, several proposals to a formal definition have been made. In table 3.1 we can see different definitions of Big Data proposed in literature.

Reference	Definition
Bulger, M., Taylor, G., & Schroeder, R. (2014)	Data on a significant level of complexity and scale that cannot be managed with conventional analytic approaches.
Gartner. (n.d.)	Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.
de Mauro, A., Greco, M., & Grimaldi, M. (2016)	The Information asset characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value.

Table 3.1: Examples of definitions of 'Big Data' found in literature

While a general consensus exists among literary sources that Big Data does not have a single concise definition. It can be described according to several aspects. We have identified 8 references describing Big Data through these aspects. All 8 references have mentioned the traditional 3 V's, Volume, Velocity and Variety.[40,41,42,21,43,44,45,46] Additionally, 6 references have mentioned Value as an aspect[42,21,43,44,45,46] and 6 have mentioned Veracity.[40,42,43,44,45,46] The complete list can be found in appendix B.1. These V's provide a framework for understanding the characteristics of Big Data. The graph in figure 3.1 shows the number of mentions the top 12 Big Data aspects have received within scientific literature.

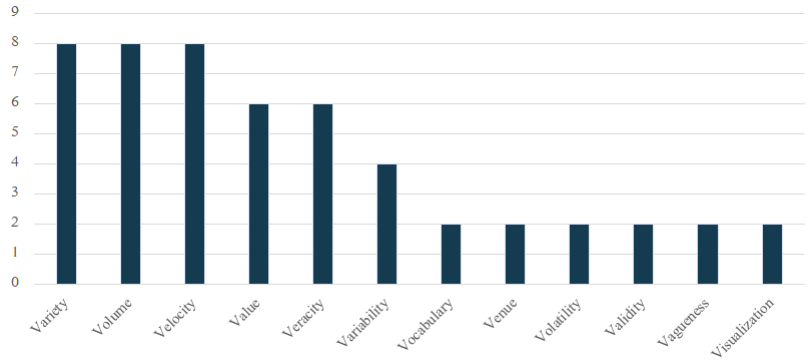


Figure 3.1: Mentions of aspects of big data found in literature (top 12)

- **Volume:** Volume refers to the vast amount of generated data available for collection, transferring and storage. This increased volume of data surpasses traditional techniques and methods for processing and analysis.
- **Velocity:** Velocity refers to the rapid generation of data, which subsequently requires the proper infrastructure to handle the timely processing of Big Data. The speed of the data flow can be as fast as real-time or near-real-time.
- **Variety:** The diversity and heterogenity is Big Data is often referred to as its Variety. This includes the different data types (structured,

unstructured and semi-structured) as well as representations (e.g. text, video, images et cetera).

- **Value:** The measurement to which the usefulness of data is determined refers to Value. In this sense, usefulness includes the benefits to the organization derived from the analyses of Big Data.
- **Veracity:** Veracity refers to the general quality of data. For proper collection, transferring and storage of data, veracity is a key aspect in maintaining the reliability of Big Data.

Practical applications of Big Data can be found in many industries, such as health care, retail, finance, manufacturing, telecommunications and many more.[47] This makes Big Data not only a trend, but an ubiquitous reality for many organizations. In terms of generation, data is extracted from many (types of) sources, for example from click streams on websites or sensors.[48] After extracting, transfering and storing data, the data is often subjected to Big Data Analytics (BDA). BDA involves uncovering trends, correlations and hidden patterns within the data, which can be used for decision making, process optimization and innovation.[44] While Big Data can be very valuable for organizations, it has also raised concerns from consumers about privacy, liability and ethical issues.[49]

3.2 Data-driven organizations

The relatively recent disruption of digital transformations has enabled businesses to increase the integration of digital technology in different aspects of the business. With the dawn of Big Data, a core component of this integration is data, which can lead to a business becoming data-driven. The concept of being data-driven has no real concise definition in literature, as it is an ambiguous concept. However, some proposals for definitions have been found in literature, in figure 3.2 some of these definitions have been presented. Although 'Data-Driven' is the commonly used term to describe this concept, other sources have made use of the term 'Information Driven'.[50] This puts a greater emphasis on the fact that facets such as data-driven innovation and analytics use information as a key resource rather than merely data, which also serves operational and processual roles within a business. A term such as 'Information Driven', or even 'Insights Driven' would refer to the fact that businesses use insights, powered by data to leverage competitive advantages. However, during this thesis we will solely refer to the concept through the term 'Data-Driven' in order to dispel confusion.

Even though no precise, agreed upon definition of a data-driven business exists in literature, Hupperz et al. (2021) have provided an overview of the key elements which make a business data-driven. According to the article, these elements are as follows:[51]

Reference	Definition
Hartmann, P. M., Zaki, M., Feldmann, N., & Neely, A. (2016)	a business model that relies on data as a key resource.[9]
Hilbig, R., Hecht, S., & Etsiwah, B. (2018, December)	A pure data-driven business model uses data as a key resource to generate any type of digital services by means of key processes such as data aggregation, data generation, data analytics, data exchange, data processing, data interpretation, data distribution and data visualization in order to create value for customers, users or stakeholders and to capture revenue.[42]
Engelbrecht, A., Gerlach, J., & Widjaja, T. (2016)	A business model of an organization is data-driven, if its core business necessarily requires digital data.[18]

Table 3.2: Various definitions of a 'Data-Driven Business'

- **Digital Transformation:** Organizations must undergo a digital transformation process in order to integrate data within their existing processes. This requires a clear vision and strategy. This process must be supported by a data-driven culture within the business.
- **Data Science:** To analyze data and extract value from it, an organization can apply Data Science. An organization is able to gain competitive advantage by leveraging data for business insights, innovation and decision making. However, it takes more than only setting up a data science team. Business analysts and IT professionals also play a critical role in this.
- **Data-Driven Business Model:** For a data-driven organization to create economic value, insights derived from data must be translated into a business model. The data-driven business model developed by Hartmann et al. (2016) contains the following elements: key resources, key activities, value proposition, customer segment, revenue model and cost structure.[9]
- **Data-Driven Innovation:** Through the vast amounts of data being generated, collected, transfered and analysed, businesses leverage the insights gained from this to discover previously unknown patterns within data and to optimize processes.
- **Data Analytics:** Data Analytics, which consists of prescriptive, descriptive and predictive analytics, is vital for extracting insights from data.

3.2.1 Data-driven decision making

The process in which data is leveraged for strategic decision making, is called 'Data-Driven Decision Making' (DDDM). This is an extension of the data-driven business through its analytical capabilities.[52] Traditionally, decisions were based on intuition rather than insights from data. Instead of decisions made based on years of experience and knowledge, DDDM allows for decisions to be informed by data analysis and its interpretation. In terms of firm performance, a research conducted by Brynjolfsson et al. (2011) using data from 179 publicly traded firms, has shown that the adoption of DDDM

resulted in a 5% to 6% increase in output and productivity.[53] DDDM can be and has been applied in different industries. Some examples noted in literature are:

- **Industry 4.0 maintenance:** In the digitized manufacturing sector, commonly known as *Industry 4.0*, operations benefit from an advanced sensor infrastructure. The enormous volume of data generated from these sensors allows for predictive maintenance, which enables proactive decision making by sensing whether a machine requires a repair ahead of its time.[54]
- **Retail:** A famous case in which a retail store has used DDDM is the one of American retail giant Walmart. It has used trillions of bytes of shopping data to predict item sales, so it could make decisions backed by data for stocking their stores during hurricane Frances in 2004.[55]
- **Finance:** Within the financial sector, DDDM is nothing new. It has seen widespread use in the stock market, using various sources of data to predict the prices of certain stocks.[56]

3.3 Data Quality

Data quality as a concept is widely interpretable. The concept of data quality is subjective and can vary depending on the context in which the data is used. Different perspectives and contexts may attribute to different levels of quality to the same data. A common definition given to data quality is *"Fitness for use"*[1], which implies the meaning varies depending on the use of the data. Within this definition, there are three viewpoints to which the entire concept can be assessed. On one hand, there is the subjective, managerial viewpoint of data quality. This viewpoint is concerned with the subjective perceptions of individuals and to what degree the data fulfills the need of stakeholders.[57] On the other hand we are dealing with the objective viewpoint of data quality.[2] An objective assessment could be task-dependant or task-independant, either reflecting the state of data with knowledge of the context or without. The third viewpoint to assess data quality entails its technical aspects, referring to data quality issues arising due to technical constraints and errors.[58]

3.3.1 Dimensions of data quality

Besides the given definition which includes an intentional vagueness about the concept, data quality can be described using a multitude of dimensions. Scientific literature within the field of data quality offers thorough classifications of data quality through its dimensions.

To understand which dimensions could be attributed to describing the concept of data quality, we have reviewed 13 references from literature, from which we have found 49 unique dimensions being analysed.[59,11,61,62,63,30,64,65,15,2,60,14] In appendix B.2 a complete overview of the dimensions per reference can be found. From this list, we have outlined the 10 most frequently highlighted data quality dimensions, by number of mentions, which have also been displayed in figure 3.2.

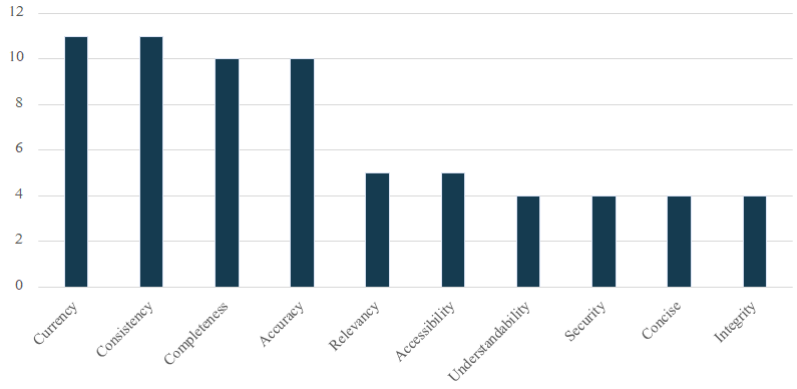


Figure 3.2: Data Quality dimensions by number of mentions in scientific literature

- **Currency:** Currency refers to the timeliness of the data. It is an assessment of how up-to-date the data is we are working with. Outdated information can lead to misleading insights, especially in an environment dealing with real-time or near-real-time data.
- **Consistency:** Consistency refers to the manner in which the data varies from representation. This includes the format in which the data is instantiated. It is assessed by the degree it has freedom from discrepancies, variations or any form of contradiction, which in turn can lead to confusion or error.
- **Completeness:** Completeness refers to the extent to which all necessary data is present or available. This includes the missing of specific values and attributes within a single dataset, to the availability of entire data objects.
- **Accuracy:** Accuracy refers to the degree to which the data represents reality. The values within the data entity is correct and reliable by being corresponding to real-world values.
- **Relevancy:** Relevancy refers to the extent to which the data is helpful, meaningful and applicable for the intended purposes. It is also an assessment of how well it aligns with the business objectives, appropriate in the context and on the technical side it includes the degree of

presence of unnecessary and redundant data.

- **Accessibility:** Accessibility refers to the availability and retrieval of data, as well as an assessment of metadata and compliance to data governance. It is related to the FAIR data principles, which is a measurement based on the findability, accessibility, interoperability and reusability of data.[66]
- **Understandability:** Understandability refers to the extent to which the data is comprehensible. Common themes within this dimensions include clarity, simplicity and ambiguity.
- **Security:** Security as a dimension of data quality refers to the extent to which access to the data entity is restricted appropriately. It assesses whether the data is secure from any threats. It corresponds to the CIA triad in security literature, including the confidentiality, integrity and availability of data.[67]
- **Conciseness:** The degree to which the representation and structure of data is compact, clear and simplistic is referred to as its conciseness.
- **Integrity:** Integrity is the extent to which a (representation of a) data entity is consistent, reliable and trustworthy over time. It assesses the changes made to the data and how it alters the data in terms of representation, values and structure.

Note that there might be overlap between specific dimensions of data quality. This can depend on the context, environment or viewpoint to which the data is assessed. The list is neither exhaustive or complete, as data quality in itself is a relative concept.

3.3.2 Data quality issues

Data quality is a spectrum, and in terms of *"Fitness for use"* its measure of quality exists somewhere between unfit for use and fit for use. Once a data entity is unfit for use, it possesses poor data quality. Once broken down, poor data quality can be attributed to specific data quality issues, which can arise in many shapes or forms. The sources of such issues can be traced to a variety of causes, such as data entries from employees or customers, changes to the source system, migration projects, user expectations, external data or system errors.[65]

This section will investigate the various data quality issues and the contexts in which they appear, by examining 13 scientific articles, which have mentioned a total of 73 unique data quality issues.[68,71,72,57,63,73,74,75,76,69,65,70,15] An overview of all references, including the data quality issues they have mentioned, has been included in appendix B.3. In figure 3.3 we have displayed

the 10 most commonly mentioned issues within the articles, appearing in a general context.

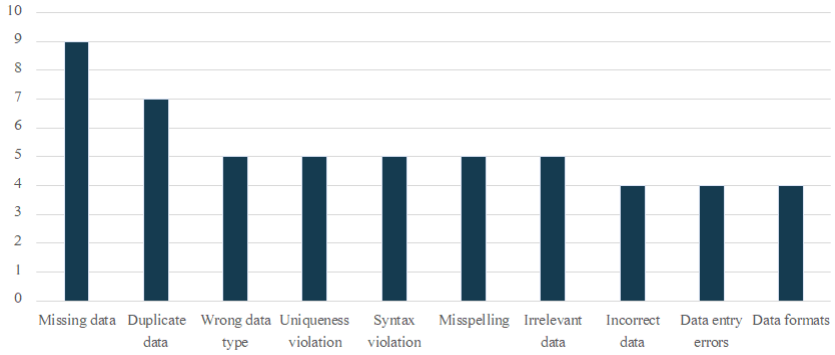


Figure 3.3: Data Quality issues by number of mentions in scientific literature

When executing ML tasks, an input is required in the form of data. It is vital for the performance of ML models that the input data does not lack in quality.[15] Biased, inaccurate or misleading predictions could be the result of poor data quality. Data quality issues commonly found within the ML process are:[15,76]

- Missing data
- Irrelevant data
- Duplicate data

A critical stage within the data integration process is ETL (Extract, Transform, Load).[77] Within this process, the data is first extracted from their sources. Secondly, it is transported to a data warehouse where it will be processed. Thirdly, the source data will be transformed and cleaned to accommodate the structure of the data warehouse. Finally, the data will be loaded into the appropriate target. Within the ETL process, the most common data quality issues according to literature are:[68,71,75]

- Duplicate data
- Wrong data types/formats
- Hardware and software constraints (computational power)

3.4 Machine Learning

As the goal of this thesis is to research the use of ML for enhancing data quality, a crucial component of the literature review is to research the ML

techniques which usable for this purpose. This section consists of three parts: first we will present a thorough definition of what ML exactly is. Then, we will examine which data quality issues are ML-solvable and within the scope of this research. At last, we will examine and evaluate which ML techniques and algorithms have been proposed by scientific literature to handle the listed data quality issues.

3.4.1 Definition

Roughly speaking, ML is a concept in which a machine (computer) learns from available input, often in the form of historical data. From this input, the machine is able to create a model which produces an output, representing the learned knowledge.[78] There are many different ML algorithms, each following a unique process to create a mathematically-based model which can be applied for its intended purposes, such as predictions. ML algorithms learn from observed data to discover patterns within to predict unobserved behaviours.[79] In ML jargon, this learning process is called *training* and the input data is refered to as *training data*. The training data we observe includes attributes, also known as *features*. These features are then utilized by the algorithms, which undergo parameter tuning, to generate an outcome.[80]

There is a general set of steps to be followed when executing ML tasks.[81] Figure 3.4 visualizes these steps within a flowchart. It begins with the extraction of the data from the correct sources, which will then be used as input for the algorithm. Secondly, the data is preprocessed, which includes subtasks such as feature selection and data cleaning. Then, the model selection takes place where the right ML algorithm will be chosen for the purpose of the ML process which will subsequently undergo a process in which the hyperparameters of the model will be tuned. Now, the model will be trained before going to the next step of model validation. During the validation step, the model is examined for its accuracy, and based on this information the model could be subjected to changes in algorithm selection and hyperparameter tuning. Finally, after a satisfactory model has been constructed, it will be deployed for its intended purposes.

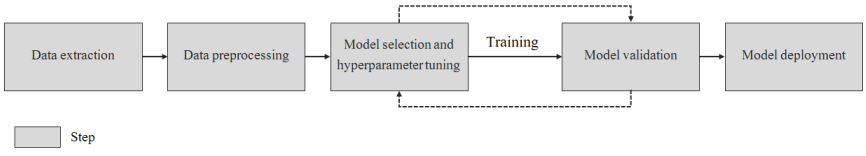


Figure 3.4: The ML process visualized in steps

There are many uses and applications for ML. For example, ML has proven

to be effective in scientific research. For example, Cesar de Sa et al. (2022)[82] have applied ML to predict the average grass length of a Dutch nature reserve. As input data, the researchers have used historical data of grass length measurements in the area, along with the corresponding spectral bands retrieved from satellite imagery which were used as training data. They were successful in creating a model that could accurately predict grass lengths based on satellite imagery data alone.

Another well known application for ML is Big Data Analytics. An example of this is an ML-based method for supporting analysis of an Intelligent Support Information Management Systems proposed by Lv et al. (2021).[83] This method integrates a Light Gradient Boosting Machine (LightGBM) algorithm for classification tasks such as predicting employee attendance for businesses.

The concept of ML contains two primary categories of learning: supervised and unsupervised learning. These categories describe the different approaches used in ML processes.

Supervised learning

First, we have supervised learning. This approach requires labeled data to be present in the training data. During the training phase, the algorithm learns by comparing the values of attributes to the corresponding output value given in the training data. The resulting model could then be used to predict the output values based on data unseen during training. If a supervised learning approach makes use of attributes to predict a continuous variable, we are dealing with a model containing a *Regression Function*. If the output variables are a discrete set of values, the model can be described as a *Classifier*.[80]

Unsupervised learning

The second category is unsupervised learning. This approach does not require labeled data for its algorithms to learn. Such types of learning find hidden patterns and relationships within the input data. Unsupervised learning algorithms are often used for tasks such as anomaly detection, clustering and dimensionality reduction.[80]

3.4.2 Machine Learning-Solvable Data Quality Issues

While a vast array of unique data quality issues exist, not all of the existing issues can be solved optimally, or are able to be solved at all with the use of ML-based techniques. We have first made a distinction on the issues being solvable with our without context knowledge. An example of of an issue being only solvable with context knowledge is *misspellings*. When

examining a dataset containing entries for last names, it is impossible to determine whether 'Jansen' or 'Janssen' is correct without the appropriate context knowledge. Among Dutch last names, both names exist, yet differ in spelling. Without knowing the true and correct name of the person in question, we cannot determine if either is a misspelling or not.

Among data quality issues which are solvable without context knowledge, we have distinguished the issues in two subcategories: non-ML-solvable and ML-solvable data quality issues. Non-ML-solvable issues typically include duplicate data, outdated data, use of special characters and syntax violations. However, some are able to be solved using ML, yet are substantially unnecessary to be handled with the use of ML. For example, duplicate data entries can theoretically be detected with the use of an ML algorithm, such as a KNN, which requires computing power scaled to the dataset at hand. Yet, the problem of duplicate data entries can be solved easily using standard operations in most programming languages. Thus, this category includes both data quality issues which are not solvable with ML at all, and those that are theoretically solvable with ML, yet are inefficient.

Then, the second subcategory covers all data quality issues that are solvable with ML, in which ML-based techniques offer efficient solutions. These typically include missing values, wrong data, irrelevant data and outliers. By training ML algorithms on the data at hand we can apply these algorithms to offer tailor-made solutions.

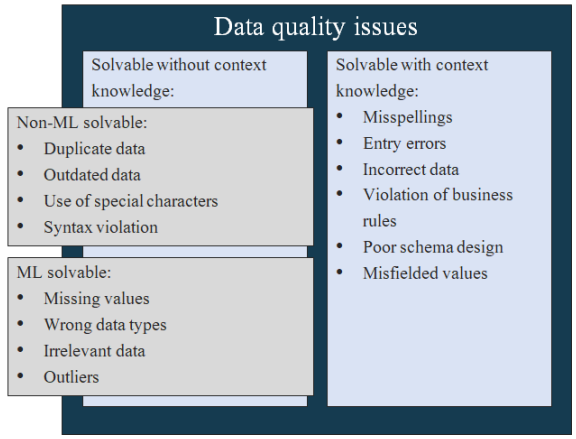


Figure 3.5: Distinction of data quality issues on ML-solvableness

In figure 3.5 we have visualized these distinctions with actual examples of data quality issues. For the selection of ML algorithms examined in this literature review we have focused on this issues in the lower-left quadrant of

the figure. These issues, missing values, wrong data types, irrelevant data and outliers will be considered in section 3.4.3.

3.4.3 Algorithms used for enhancing data quality

For this section, we have examined different literature which have studied or proposed the use of specific ML algorithms for dealing with data quality issues. We have selected references which have dealt with missing data,[84,85,86,87,88] wrong data types,[89] outliers[90,91] and irrelevant data.[92,93,94] The full table including the references and proposed algorithms can be found in appendix B.4. The primary algorithms named by the references will be described in this section.

Random Forrest

The Random Forrest (RF) algorithm is an ML technique introduced by Breiman (2001).[95] The idea behind RF is that it uses a random selection of attributes and instances from the data to create binary decision trees, either used for regression or classification tasks. The algorithm will then compare these binary decision trees based on performance, and selects the best performing trees to be used for the corresponding model.

K-Nearest Neighbour

K-Nearest Neighbour (KNN) has first been proposed by Altman (1992).[96] This algorithm can be used for both supervised and unsupervised learning, as well as for classification and regression tasks. KNN finds the k-nearest residing data points from a single data point by measuring its distance between each other, for example using the euclidean distance. By finding the closest neighbours of a data point, we can create clusters of data points, from which we can classify data instances to specific categories or to find outliers.

Neural Network

A Neural Network (NN) is a type of ML algorithm which mimics computations made by the human brain by recreating its structure. An NN is a network of directed and weighted nodes, inspired by neurons found in the brain. It works by passing signals from node to node, which then send a summation of all received signals to an activation function on the next layer of neurons. The weights of the nodes and the value of the signals it receives determines the output. Many applications for NN's exist, such as image and speech recognition, or simple classification and regression tasks.[97]

AutoEncoder

AutoEncoders are a type of NN, which can be used to compress and decompress data. The attributes of a dataset enter the NN through the first layer and subsequently to the next, smaller layer, representing a bottleneck in the NN. This entails a lower-dimensional representation of the data. During the learning phase, the encoder learns to extract the most important attributes of the input data.[98]

Chapter 4

Design

In this chapter we will describe the design of the method we have envisioned to enhance data quality. First, we will describe the considerations taken into account for the design. This will be accomplished through the use of listing all requirements which have been formed prior to creating the design. Next, the general architecture will be shown and subsequently explained using the ArchiMate modelling language. Components within this architecture will be visualized with UML class diagrams. The general idea is to develop a methodology which can be used in a holistic sense. This approach provides a blueprint for creating an actual prototype tool for practically implementing the method. Section 5 specifies how the method described in this section has been translated to a working prototype.

4.1 Requirements

Before being able to create a design, we have specified a list of requirements suited to our data enhancement and assessment method. First, we will describe the approaches we have taken for gathering the requirements. Second, we will describe the different relevant stakeholders. Then, we will list the requirements we have specified in three categories: functional-, non-functional- and data requirements.

4.1.1 Elicitation methods

The requirements have been gathered primarily through an analysis of available literature on the subjects of data-driven literature, big data, data quality and machine learning, which we have included in our literature review in section 3. The intention of the method is to handle data used for strategic decision making, typically big data found in a data-driven setting. Analyzing the behaviour and characteristics of such data, as well as establishing a thorough understanding of data quality and ML capabilities, has allowed us

to conceive these requirements.

While the literature study has been the main source of requirements, we have also followed a prototyping approach for eliciting and validating the requirements. We have done this by constructing the basic functionalities in a simple Python environment. This has allowed us to explore the practical feasibility of certain features and aspects of the method. For example, we have first applied a random forest algorithm to impute the missing values of a given dataset. Attempting this has given us insights in the average training- and imputation speed of the algorithm. During the literature study we have established that the processing speed of data is an important factor, due to the big data-characteristic of velocity. During the prototyping phase we have then established that a requirement for controlling the training time of the algorithm is required for accommodating this characteristic, as the training time of the ML algorithm would often span across several hours on datasets with tens of thousands of data instances.

Finally, the requirements have been validated by a field expert: a senior manager employed in the Data Management team at PwC NL.

4.1.2 Stakeholders

A variety of relevant stakeholders exists for our envisioned data quality enhancement method. A stakeholder can be defined as a person with interest in the use and and outcomes of this method. The main stakeholders are:

- **Decision Maker:** An individual with the authority to make strategic decisions within an organization can be referred to as a decision maker. In this context, the decisions are made based on data, or data plays a significant part within the decision making process. These decisions rely on high-quality data, which could require certain tooling to ensure the quality of data.
- **Data Analyst:** Data Analysts support the data-driven decision making process. Within an organization, they interpret the available data and provide valuable business insights. Data quality is an important factor to ensure high quality insights.
- **Data Scientist:** Data Scientists work with data on a regular basis, which are often responsible for ML tasks regarding the processing of data. As our envisioned method is specified for enhancing the quality of data used for strategic decision making, specifically for any form of prescriptive, descriptive or predictive modelling, a tool accomplishing this approach is able to provide significant support for the pre-processing phase of ML tasks.

- **Data Provider:** Data providers are individuals who are responsible for collecting, generating or supplying data for their specified purpose. Data providers ensure the data is of sufficient quality before its utilisation, typically including activities such as data cleaning, labelling and preprocessing. This role is otherwise known as a Data Engineer.

4.1.3 Functional Requirements

The definition of functional requirements as given by Robertson (2014) is *"Functional requirements specify what the product must do — the actions it must perform to satisfy the fundamental reasons for its existence."*[99] We have specified an array of functional requirements, categorized in the two main imagined functionalities, data quality enhancement and data quality assessment. These functional requirements will lay the basis for the general functionalities and qualities of what the method should achieve.

Data Quality Enhancement

The data which will be given as input should be enhanced in terms of data quality. It should detect the data quality issues mentioned in section 3.4.2 and handle them accordingly. Importance lies in the fact that ML techniques will be applied within our approach, and systematically improve on these data quality issues. The choice of ML requirements are based on the findings of section 3.4.

- The method should be able to enhance the data quality of data used for strategic decision making. The method should be tailored to accommodate such data.
- It should be able to detect and impute missing values in a dataset.
- It should be able to detect and remove or regularize outliers in a dataset.
- It should be able to detect and correct wrong data types and formats.
- It should be able to detect irrelevant attributes of a dataset.
- It should be able to reduce the dimensionality of a dataset.
- It should apply ML techniques for handling data quality issues.
- Models created during the training process of an ML task should be saved for future use.
- It should be able to produce output in the form of an improved dataset, after enhancing its data quality.

Data Quality Assessment

Besides actually enhancing a dataset on its quality issues, it is important

for stakeholders to be aware of the perceived quality of a dataset. This information should be presented towards both individuals on a strategy level and individuals working with the data on a technical level. The assessment of the quality of a dataset is mainly based on the findings of section 3.3.1, describing the dimensions of data quality.

- The dataset should be assessed on its perceived quality.
- A report detailing several aspects of the assessment and enhancement phases should be constructed as an artefact.
- The report should contain a numerical rating in the form of a grade attributed to its quality.
- The report should provide general information on the input data:
  - Amount of data instances.
  - Amount of attributes.
  - Names and data types of attributes.
  - Purpose of input data as strategic asset.
- The report should contain an assessment on the following dimensions of data quality:
  - Currency
  - Deduplication
  - Completeness
  - Relevancy
  - Consistency
- The report should provide technical details about the data quality enhancement process:
  - Accuracy of the data quality enhancement techniques (performance of ML algorithms).
  - Time taken for training the models.
  - Time taken for the entire data quality enhancement process.

4.1.4 Non-functional Requirements

Non-functional requirements are defined as *"properties, or qualities, that the product must have if it is to be acceptable to its owner and operator"*, according to Robertson (2014).[99] We have specified several non-functional requirements for our case, based on the aspects of software quality,[100] as this should lay

a sustainable foundation for developing software programs based on our envisioned method. These requirements are catered towards the handling of big data, described in section 3.1, as well as the characteristics of a data-driven environment, described in section 3.2.

- **Performance**
  - Training any ML model should not require more than 60 minutes of time. Due to the high velocity in which data is handled, training time should not exceed this time.
  - The method should be scalable in terms data volume. The amount of attributes and instances a prototype should maximally be able to handle depend on various considerations such as the nature of the dataset and available computational power. As a baseline, it should be able to handle sets with 10 million instances and 30 attributes.
  - The method should be able to handle data coming in real-time or near-real-time velocity.
- **Usability**
  - The method should follow a clear, linear approach. The user will have to specify what functionalities and features should be utilised, and the method will linearly execute the specified activities.
  - The approach should allow for a clear user interface if developed into a software tool. This way, the method ensures any prototype to be usable for a variety of user types.
- **Security**
  - All input and output files, models and reports should be processed and stored locally. This way, no third party will be in possession of sensitive datasets.
- **Maintainability**
  - The method should be designed in a modular way to enable maintenance and modification. As the field of ML, along with our understanding of data quality, is continually evolving, it ensures newer technologies to be integrated within the method.

4.1.5 Data Requirements

Finally, for adequate handling of input data, we have specified requirements to what types of data and characteristics the tool should be able to process. As well as the non-functional requirements described in section 4.1.4, these

requirements have been based on our analysis of literature regarding data-driven environments and aspects of Big Data, found respectively in section 3.2 and 3.1.

- The method should be able to be integrated within the data pipeline of an organization.
- It should be able to process standalone files of datasets.
- Input data should be presented in a structured, tabular format.
- The method should be able to process multiple file extensions. Examples are .csv, .json, .xml and .xlsx.
- It should be able to process and support multiple file locales and number formatting.
- It should be able to process various data types and formats.

4.2 Architecture

Based on the requirements described in section 4.1 we have proposed a general architecture for the method for enhancing data quality of data used for strategic decision making, as well as assessing its perceived quality. The architecture has been visualized using ArchiMate.[101] ArchiMate is a modelling language, developed by The Open Group, used as a framework for communicating and documenting architectural designs of IT systems. In Appendix C we have included a description of the ArchiMate notations, including its elements and relationship symbols. Figure 4.1 visualizes the architecture of our method. It portrays the method as a holistic approach rather than including processes that detail specific tasks and conditions. Yet, it takes in mind the characteristics and behaviour of data found in pipelines leading up to DDDM. As this type of data lies at the center of our focus during this research, we have accommodated it throughout the full design.

The initial step of the full process is reading the input data. This will be handled by the Data Prep Component. This can be executed in two ways. First, the method can be integrated within an active data pipeline of a business, where the input data enters the system via an API service that connects the business' data pipeline with our method. Secondly, a user can decide to manually upload their datasets via a user interface. Once the data is read in either of the two ways, it will enter a data preparation process. Data preparation entails standardizing the format of the input data, preparing it for the data enhancement and quality assessment procedures. This is essential, as input files often exist in a wide array of different formats. For example, datasets can differ in file extensions (.csv and .xlsx). Besides this,

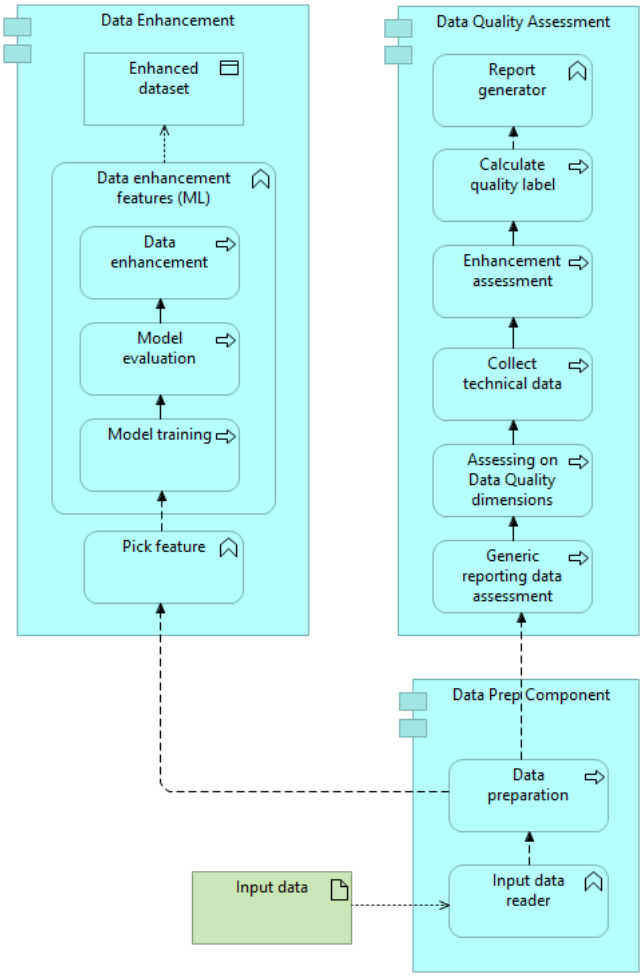


Figure 4.1: Architecture of the method made in ArchiMate

locale settings can influence the format of the data itself. For example, the US file locale often uses comma's to group thousands and points to denote decimals (e.g. 3,029,281.00), while European file locales use points to group thousands and comma's to denote decimals (e.g. 3.029.281,00). Making sure such differences are recognized and subsequently standardized is imperative for the main features to function properly. After processing the data, the method branches of into a Data Enhancement Component and a Data Quality Assessment Component.

To start off with the Data Enhancement Component, it starts by picking the enhancement features the user of the method has required. These features correspond to the data quality issues handled by the method (missing data, irrelevant data, outliers and wrong data types) as well as their subsequent

identification and enhancement. After a feature pick has been made, the corresponding ML task commences, as described in section 4.1.3. During this task, several objects will be produced: the ML models, which can be stored and used for future usage to avoid redundant training time, technical information about the training process (e.g. training time and model accuracy) and ultimately an enhanced dataset. After all required enhancement features have been completed, the user will be left with a definitively improved dataset.

The Data Quality Assessment Component fires off synchronously with the data enhancement component. Its initial step is collecting generic info on the input data (e.g. number of instances and names of attributes). Secondly, this process assesses the input data on the basis of the dimensions of Data Quality as described in section 3.3.1. Thirdly, all available technical information collected during processes within the Data Enhancement Component will be collected and solidified. Fourthly, the predictive power of the original dataset versus the enhanced dataset should be measured. By feeding both datasets to a simple ML algorithm used for predictive modelling should create output in which this difference can be measured using some performance metric. Fifthly, based on all previously measured and consolidated reporting data, a quality label is to be calculated to indicate the overall health in terms of data quality. This will be calculated for both the input dataset in its original state and the enhanced dataset. This way, any decision maker will be aware of the quality of the data used in their DDDM process and on what points it shows potential for improvement.

After completing this task, all available reporting information should be present. All generic reporting data, data quality assessment reporting data and technical reporting data will be appended to a report, which will be produced as output by the Data Quality Assessment Component.

4.2.1 Data Enhancement Component Design

The ArchiMate architecture in figure 4.1 from section 4.2 shows a generic overview of the Data Enhancement Component, as it illustrates the ML tasks as a template rather than a complete overview of the different enhancement features the method has to offer. In figure 4.2 we have exhibited a detailed overview of this component, using a UML class diagram.

The central class within the diagram is *data_enhancement*, which is created once the Data Enhancement Component has been used. Depending on how many features for enhancing the data are used, it creates a child class called *enhancement_feature*. This child class is responsible for training, evaluating

and exporting the ML models, performing the actual enhancements on the dataset as well as collecting technical reporting data used in the Data Quality Assessment Component. The *enhancement_feature* class also has child classes, depending on which data quality problem is handled: missing data, irrelevant data, outliers or wrong data types.

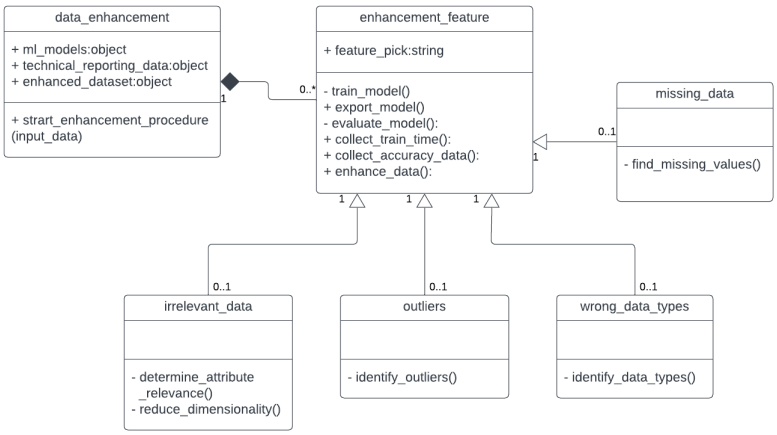


Figure 4.2: UML Class Diagram of the Data Enhancement Component

4.2.2 Data Quality Assessment Component Design

As for the Data Enhancement Component described in section 4.2.1, the Data Quality Assessment Component has been visualized as a generic component in the general architecture in figure 4.1 from section 4.2. Figure 4.3 provides a UML class diagram for a detailed vision of the Data Quality Assessment Component, including the different classes in the design. The diagram shows *reporting_data* as a central class in which all the reporting data is accumulated.

First off, we have the *reporting_data_technical* class, which processes the technical information required for the report. It mostly contains information on the accuracies of the ML models which have been trained in the process, along with their training times and the names of the algorithms. This class will exist regardless of whether a *reporting_data* class exists, as this data is only collected within the Data Enhancement Component if it has been used. It is entirely possible for a user to only make use of the Data Quality Assessment Component, and not the Data Enhancement Component. This class is the parent class of several child classes, depending on which features are used, if any are used at all. Of these child classes, the *technical_missing_data* class, which contains the technical reporting data for the missing data handler feature, has another child class. This child class includes reporting data for

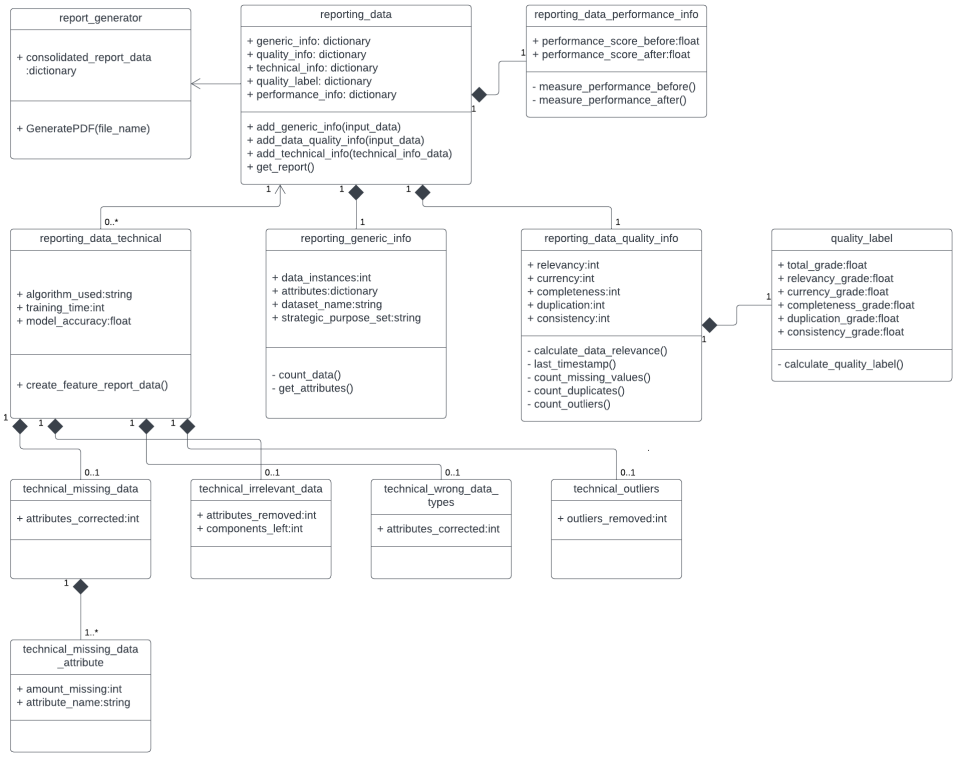


Figure 4.3: UML Class Diagram of the Data Enhancement Component

each attribute handles by this feature, as an ML model will be created for each attribute containing missing values.

The second child class of *reporting_data* is *reporting_generic_info*. This class contains all generic reporting data for the input data, such as the amount of data instances, an overview of all attributes, the name of the dataset and a description of its strategic purpose, which should be included in the final report.

The third child class of *reporting_data* is *reporting_data_quality_info*. This class contains and collects reporting data from the five primary dimensions of data quality, as described in section 3.3.1. This class has a child class *quality_label*, which will calculate a grade reflecting on the overall quality of the dataset based on these dimensions of data quality. This quality label will be calculated for the original input data as well as the enhanced dataset. The idea is to give the stakeholders an overview of what dimensions of data quality require additional attention and what dimensions have improved since the data enhancement.

Fourthly, we have the child class of *reporting_data_performance_info*, which will measure the predictive performance of both the original input dataset and the enhanced dataset. The purpose of this class is to give stakeholders clear metrics to how much the dataset has been improved in terms of potential for the DDDM process.

Finally, the UML diagram contains the *report_generator* class, which holds a collection of all reporting data created within the Data Quality Assessment Component-process. Subsequently, it generates a PDF with all the available information.

Chapter 5

Prototype

In order to validate the method designed in chapter 4 we have created a working prototype as a proof-of-concept. In this chapter we will introduce *HAQIM* (**H**olistic **A**ssessment of **Q**uality and **I**mprovement **M**ethod) as the realization of this method. As of completing the thesis, the prototype sits at version 1.0, containing a limited amount of functionalities in relation to the full design.

In this chapter we will first specify what elements of the design have been included within HAQIM. Secondly, we will describe the approach we have taken to implement the prototype and what technologies have been used in doing so. Lastly, this section contains a practical use case to illustrate the working of the prototype. This use case contains a step-by-step description of how the method has been used to enhance the data quality of an actual dataset by applying the prototype.

5.1 Functionalities

When creating HAQIM, not all functionalities and elements from the general architecture in figure 4.1 have been (fully) incorporated. In figure 5.1 we have visualized which elements of the general architecture have been included and excluded from HAQIM. Firstly, the prototype requires a manual input of datasets as it cannot be integrated within an actual datastream. The input data reader only support .csv files, containing structured, tabular data. Additionally, only Dutch locale settings are supported (using comma's to denote decimals and points to denote thousands). Secondly, within the Data Quality Assessment Component, during the assessment phase in which the dimensions of data quality of the input dataset are established, only four dimensions are included: currency, deduplication, relevancy and completeness. However, these quality measures are basic measurements of the degree to

which the dataset adheres to the fulfillment of these quality dimensions, so HAQIM takes the following metrics into account:

- **Currency:** The amount of days between the assessment of the dataset and last data instance in the input set.
- **Deduplication:** The amount of duplicate data instances.
- **Relevancy:** The relevance of each attribute in relation to the target attribute. We have calculated these relevancy scores by entering the attributes in a decision tree regressor, and taking the values used by the decision tree to determine their relevance to the target variable.
- **Completeness:** The total amount of missing data instances.

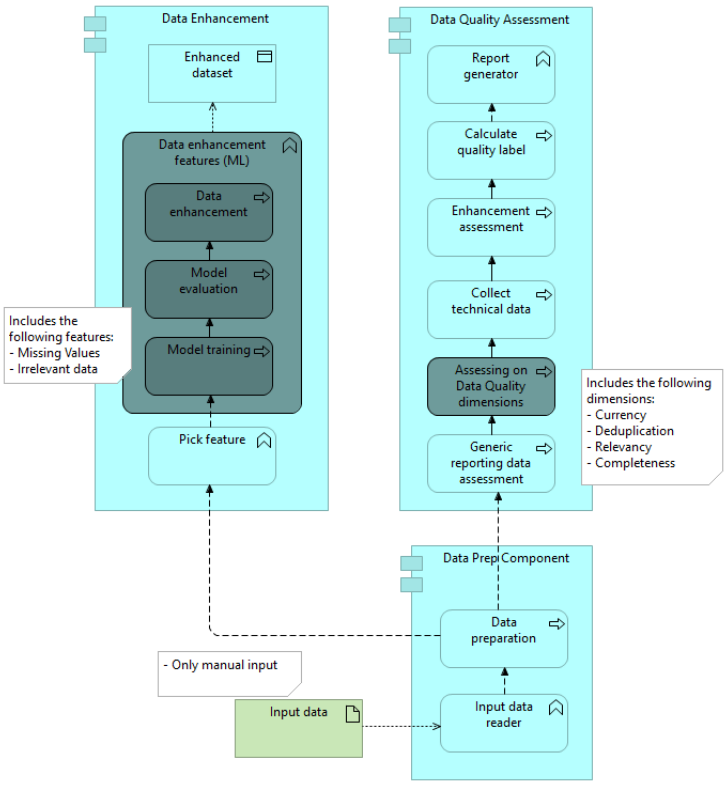


Figure 5.1: Visualization of inclusion and exclusion of features and elements in prototype creation

Thirdly, within the Data Enhancement Component, looking at the included features, only two out of four ML-based enhancement methods are included in HAQIM: handling missing values and irrelevant data. For handling missing values we have included a feature that applies a random forest algorithm to impute missing values. For handling irrelevant data, we have applied a decision tree algorithm to establish the importance of the attributes

within the dataset, and consequently used an AutoEncoder based on a Neural Network to reduce the dimensionality of the data.

In addition to the exclusion of several features and elements included in the general architecture, the prototype does not possess an elaborate user interface, while this has been the intention for any realization of the method.

As mentioned in section 1.3, this thesis research project has been subjected to a scope definition. Multiple constraints pertaining to time and resources have defined the scope of the entire research, including the development of HAQIM. The choices for the inclusion and exclusion of specific features and elements are based on these constraints.

5.2 Implementation

This section will describe the approach taken to implement HAQIM, including an overview of all the technologies which have been used in this process. When developing HAQIM, Python has been selected as the main programming language.[102] Python is a high-level programming language with an extensive selection of libraries and packages, which can be used for a variety of purposes. The language is convenient for data-centric tasks, including the handling of data and executing ML tasks. The Python packages and libraries that have been used to implement HAQIM are the following:

- Pandas[103]
- SciKit-Learn[104]
- NumPy[105]
- PyFPDF[106]

5.2.1 Algorithms

As mentioned in section 5.1, we have implemented the functionalities of handling missing data and irrelevant data within HAQIM. In this section we will describe how these algorithms have been implemented.

Missing data

For handling missing data, we have implemented the Random Forest algorithm. The choice for this algorithm has been based on our literature review. We have found the random forest algorithm to be the best performing algorithm according to the literature cited in section 3.4.3.

For the algorithm to train, we have first isolated data instances containing missing values. The remaining, full and inclusive part of the dataset will be used for training the Random Forest. For each attribute containing missing values we have created a different model. This way, all created models will be tailored to a specific attribute from that specific dataset.

For our model selection strategy we have made use of k-fold cross validation with k=3. During cross validation different models will be compared to pick the model with optimal performance. We have made use of a random search grid to optimize the hyperparameter settings of our models. Random search grids are commonly used for hyperparameter tuning. This method selects a few combinations of hyperparameter settings for a model at random, from which the best performing settings will be picked. In comparison with other methods such as a common grid search, which exhaustively searches all possible combinations, a random search grid is more effective and efficient in finding the optimal set of hyperparameter values.[107] In section 5.2.2 we will describe what hyperparameter settings we have used in HAQIM.

After training all models and selecting the models with optimal performance, they will be applied to the prior isolated part of the dataset containing the missing values. After using the models to impute the missing data, the isolated part will be merged with the dataset used for training, sorting data in the original order.

Irrelevant data

For handling irrelevant data, we have created a weighted Neural Network-based AutoEncoder. According to the literature we have cited in section 3.4.3. The most common ML technique to assess the relevance of data is through a decision tree, and to filter the data and reduce the dimensionality of the dataset the suggested ML algorithm is an AutoEncoder.

In our approach, we first employ a Decision Tree algorithm to evaluate the importance of each attribute in the input data relative to the target variable specified by the user. This assessment helps us determine the relevance of each attribute. The algorithm does this by assessing the importance of each attribute in the input dataset by analyzing their contribution to the predictive power of the target variable. This assessment is commonly achieved through techniques such as information gain. The decision tree algorithm constructs a tree-like structure, where each internal node represents a decision based on an attribute, and each leaf node corresponds to a predicted outcome or class label. During the construction process, the algorithm evaluates different

attributes and selects the most informative ones that lead to effective predictions.

To quantify the importance of attributes, the decision tree algorithm considers how attributes split the data into subsets that are more homogeneous in terms of the target variable. Attributes that result in significant information gain are deemed more important. By evaluating multiple attributes and their splits, the decision tree algorithm assigns relevance scores or rankings to each attribute, reflecting their relative importance in determining the target variable. In our methodology, we leverage these relevance scores obtained from the decision tree algorithm by assigning weights to the attributes in the input dataset, which will be scaled to a decimal value between 0 and 1. Higher values indicate a larger importance of the attribute.

Once the data has been scaled based on attribute importance, it is passed through a Neural Network. The Neural Network consists of multiple layers, including an input layer, in our case one hidden layer and an output layer. In a typical Neural Network, data flows from the input layer through the hidden layers and finally produces an output, often in the form of a prediction, at the output layer. However, with HAQIM, we employ a specific architecture, as shown in figure 5.2 to achieve dimensionality reduction and obtain a compressed representation of the data.

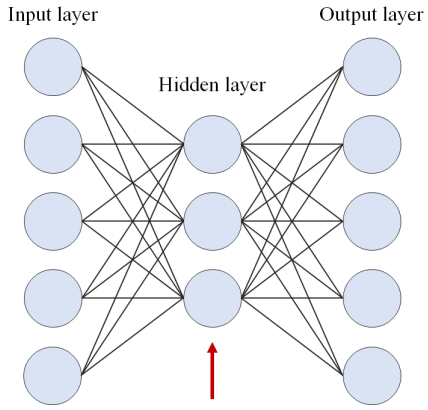


Figure 5.2: Location of the hidden layer on a Neural Network

The data, which has been scaled to importance, enters the input layer of the Neural Network as usual. It then progresses through the hidden layers, which contains a smaller number of neurons compared to the original input dimensionality. This reduction in neuron count essentially reduces the

dimensionality of the data. In our prototype, we have chosen to reduce the dimensionality to three attributes. This means that regardless of the initial number of attributes in the input dataset, HAQIM will transform the data to a compressed form represented by only three attributes in the hidden layer. By stopping the data flow at the hidden layer, HAQIM captures the essential features and patterns of the input data while discarding less significant details.

5.2.2 Hyperparameter settings

We have used multiple ML algorithms within the functionalities of HAQIM. Here we will describe what hyperparameter settings have been used.

For handling missing values, we have implemented a random forest to impute missing values. To find the best hyperparameter settings, we have made use of a random search grid. The ranges of the hyperparameter settings in the random search grid are noted in table 5.1.

Parameter	Tuning range	Description
n_estimators	{50, 60, 70 … 1990, 2000}	Number of trees in the random forest
max_features	{'auto', 'sqrt}	Number of features considered at every split
max_depth	{5, 10, 15 … 115, 120}	Maximum number of levels in a single tree
min_samples_split	{2, 5, 10, 15, 20}	Minimum number of samples required to split node
min_samples_leaf	{1, 2, 4, 5, 7}	Minimum number of samples required at each leaf node
bootstrap	{true, false}	Method for selecting samples when training a tree

Table 5.1: Search space of hyperparameter settings for random forest algorithm

As for our AutoEncoder, based on SciKit-Learn's MLPRegressor, we have also used a set of hyperparameter settings. Unlike the random forest algorithm, we have not used a random search space and a cross validation strategy to find the optimal hyperparameter settings. We have used a standardized set of values for each use of the algorithm.

Table 5.2 shows the values of the hyperparameter settings of our AutoEncoder. The hidden_layer_sizes hyperparameter denotes the architecture in terms of how many nodes each layer in the neural network consists of. The solver explains what method is selected for optimizing the capabilities of the network itself with the intent of minimizing loss. We have selected the 'adam' algorithm, which is an algorithm for first-order gradient-based optimizations, proposed by Kingma et al. (2014).[108] Additionally, for the activation function we have picked the 'relu' function. In a Neural Network, the role of the activation function is to transform the input weights into a value fed to the next layer or output. The 'relu' function, otherwise known as Rectified Linear Unit, which offers a simple linear computation to transform the values.

Parameter	Tuning range	Description
hidden_layer_sizes	10, 3 ,10	Number of neurons in each layer
solver	adam	Solver for weight optimization
activation	relu	Activation function for the hidden layer

Table 5.2: Hyperparameters used for our AutoEncoder

5.3 Use case

In this section we will provide a use case example of applying HAQIM to a real-life dataset. We will demonstrate step-by-step how HAQIM assesses the data quality of the set and collects all reporting data, as well as how it enhances its data quality.

Input data

For the use case demonstration, we have made use of the travel accessories webshop visitor and sales data, as described in section 2.4.1. For this use case example, we have artificially included missing values, as the original dataset contained full data completeness. In figure 5.3 we have included screenshot of the example data. In the figure we can notice missing values in the *tablet_visitors* and *desktop_visitor* attributes.

	date	referra	organic_search	direct_visitors	organic_socia	total_avg_time_spent	mobile_avg_time_spen	tablet_avg_time_spent	desktop_avg_time_spen	total_visitors	mobile_visitc	tablet_visitors	desktop_visitc	target
490	24-7-2022	1	11	1	0	23,58333333	22,7	0	28	12	10	0		1
491	25-7-2022	0	6	0	0	19,83333333	19,4	22	0	6	5	1		0
492	26-7-2022	0	10	3	3	33,625	12	0	55,25	16	8	0		0
493	27-7-2022	0	7	11	7	24,88	21,5	0	49,66666667	25	22	0		1
494	28-7-2022	0	14	14	4	23,1875	18,36	0	40,42857143	32	25	0		0
495	29-7-2022	0	6	15	7	32,10714286	32,23076923	0	30,5	28	26	0		2
496	30-7-2022	0	11	13	5	23,4137931	22,2	0	31	29	25	0		0
497	31-7-2022	0	8	7	3	14,36666869	10,33333333	0	31,66666667	18	15	0		0
498	1-8-2022	0	9	0	0	15	11,16666667	0	22,66666667	9	6	0		0
499	2-8-2022	1	6	1	0	24,875	18,5	0	44	8	6	0		0
500	3-8-2022	0	9	1	0	70	76,55555556	0	11	10	9	0		2
501	4-8-2022	0	9	0	0	17,33333333	19	0	4	9	8	0		0
502	5-8-2022	0	9	1	0	28	18	0	43	10	6	0		0
503	6-8-2022	0	8	1	0	34,55555556	34,55555556	0	0	3	3	0		0
504	7-8-2022	0	3	1	0	6,5	11,5	0	1,5	4	2	0		0
505	8-8-2022	0	11	2	0	36,46153846	40,1	0	24,33333333	13	10	0	3	0
506	9-8-2022	1	4	0	0	32,8	0	94	17,5	5	0	1	4	0
507	10-8-2022	0	9	1	1	14,81818182	14,7	0	16	11	10	0	1	0
508	11-8-2022	0	7	0	0	31,14285714	26,5	0	59	7	6	0	1	0
509	12-8-2022	0	4	1	0	22,2	23,75	0	16	5	4	0	1	0
510	13-8-2022	0	6	0	0	16,16666667	6	0	36,5	6	4	0	2	0
511	14-8-2022	0	7	0	0	24,85714286	18,4	0	41	7	5	0	2	0
512	15-8-2022	0	11	0	0	16,45454545	16,66666667	0	15,5	11	9	0	2	0
513	16-8-2022	0	13	1	0	36,85	41,15384615	0	28,85714286	20	13	0	7	1
514	17-8-2022	0	20	1	0	22,35238095	22,38461538	0	23,875	21	13	0	8	0
515	18-8-2022	0	8	3	0	10,63636364	12,6	0	9	11	5	0	6	0
516	19-8-2022	0	9	3	0	21,5	25,6	0	1	12	10	0	2	0
517	20-8-2022	0	9	1	0	44,8	46	0	40	10	8	0	2	0
518	21-8-2022	0	7	1	0	36,5	38	0	26	8	7		1	1
519	22-8-2022	0	8	1	0	33,22222222	42,2	0	22	9	5		4	0
520	23-8-2022	0	6	0	0	38,33333333	38,33333333	0	0	6	6		0	0
521	24-8-2022	1	6	3	0	92,1	84,83333333	0	103	10	6		4	1
522	25-8-2022	0	5	1	0	34,66666667	32,75	0	38,5	6	4		2	0
523	26-8-2022	0	5	0	2	26,57142857	25,5	0	33	7	6		1	0
524	27-8-2022	0	2	0	0	110,5	201	0	20	2	1		1	0

Figure 5.3: Screenshot of the example input data before applying the prototype

Applying the prototype

HAQIM requires a set of options to indicate which features the user wants to use, which is specified within the code. In figure 5.4 we can see that the options for handling missing values and irrelevant data are turned on, which are specified on line 69 and 70 in the example code. The target attribute of this dataset, containing the amount of sales, is called 'target'. The options also contain the name of the file of the dataset and the variable

referring to it within the code. These options will be passed on to the function *start_improvement()*, where the entire process will begin.

```
61  def main():
62
63      file = r"C:\Users\                                                    \example-use-case-data.csv"
64
65      df = pd.read_csv(file, sep=None, decimal=',', engine='python')
66      filename = os.path.splitext(os.path.basename(file))[0]
67
68      options = {
69      "missing_data": True,
70      "irrelevant_data": True,
71      "predictor": 'target',
72      "file": filename,
73      "input_data_set": df,
74      }
75
76      start_improvement(options)
77
78  if __name__ == '__main__':
79    main()
```

Figure 5.4: Screenshot of the option selection in HAQIM

Handling missing data

The first data quality issue to be touched by HAQIM is missing data. As it saves an enhanced dataset for each ML feature, we have included a screenshot of the resulting file in figure 5.5. The figure shows how the missing values from the *tablet_visitors* and *desktop_visitor* attributes have been handled using the random forest imputation algorithm.

	date	referral_visitor	organic_search_visitors	direct_visitors	organic_social_visitors	total_avg_time_spent	mobile_avg_time_spent	tablet_avg_time_spent	desktop_avg_time_spent	total_visitors	mobile_visitors	tablet_visitors	desktop_visitors	target
490	2022-07	1	11	1	0	23,58333333	22,7	0	28	12	10	0	3,315727595	1
491	2022-07	0	6	0	0	19,83333333	19,4	22	0	6	5	1	0,501902505	0
492	2022-07	0	10	3	3	33,625	12	0	55,25	16	8	0	4,776757041	0
493	2022-07	0	7	11	7	24,88	21,5	0	43,66666667	25	22	0	6,203418614	1
494	2022-07	0	14	14	4	23,1875	18,36	0	40,42857143	32	25	0	6,351754083	0
495	2022-07	0	6	15	7	32,10714286	32,23076923	0	30,5	28	26	0	7,472493091	2
496	2022-07	0	11	13	5	23,4137931	22,2	0	31	29	25	0	6,341087449	0
497	2022-07	0	8	7	3	14,38888889	10,93333333	0	31,66666667	18	15	0	3,399732961	0
498	2022-08	0	9	0	0	15	11,16666667	0	22,66666667	9	6	0	3,360295278	0
499	2022-08	1	6	1	0	24,875	18,5	0	44	8	6	0	2,35613412	0
500	2022-08	0	9	1	0	70	76,55555556	0	11	10	9	0	2,110810411	2
501	2022-08	0	9	0	0	17,33333333	19	0	4	9	8	0	2,355073658	0
502	2022-08	0	3	1	0	28	18	0	43	10	6	0	3,340861574	0
503	2022-08	0	8	1	0	34,55555556	34,55555556	0	0	9	9	0	1,442228631	0
504	2022-08	0	3	1	0	6,5	11,5	0	1,5	4	2	0	1,459177594	0
505	2022-08	0	11	2	0	36,46153846	40,1	0	24,33333333	13	10	0	3	0
506	2022-08	1	4	0	0	32,8	0	94	17,5	5	0	1	4	0
507	2022-08	0	9	1	1	14,81818182	14,7	0	16	11	10	0	1	0
508	2022-08	0	7	0	0	31,14285714	26,5	0	59	7	6	0	1	0
509	2022-08	0	4	1	0	22,2	23,75	0	16	5	4	0	1	0
510	2022-08	0	6	0	0	16,16666667	6	0	36,5	6	4	0	2	0
511	2022-08	0	7	0	0	24,85714286	18,4	0	41	7	5	0	2	0
512	2022-08	0	11	0	0	16,45454545	16,66666667	0	15,5	11	9	0	2	0
513	2022-08	0	13	1	0	36,85	41,15384615	0	28,85714286	20	13	0	7	1
514	2022-08	0	20	1	0	22,95238095	22,38461538	0	23,875	21	13	0	8	0
515	2022-08	0	8	3	0	10,63636364	12,6	0	9	11	5	0	6	0
516	2022-08	0	9	3	0	21,5	25,6	0	1	12	10	0	2	0
517	2022-08	0	9	1	0	44,8	46	0	40	10	8	0	2	0
518	2022-08	0	7	1	0	36,5	38	0	26	8	7	0	1	1
519	2022-08	0	8	1	0	33,22222222	42,2	0	22	9	5	0	4	0
520	2022-08	0	6	0	0	38,33333333	38,33333333	0	0	6	6	0	0	0
521	2022-08	1	6	3	0	32,1	84,83333333	0	103	10	6	0	4	1
522	2022-08	0	5	1	0	34,66666667	32,75	0	38,5	6	4	0	2	0
523	2022-08	0	5	0	2	26,57142857	25,5	0	33	7	6	0	1	0
524	2022-08	0	2	0	0	110,5	201	0	20	2	1	0	1	0

Figure 5.5: Screenshot of the example input data after imputing missing data

Handling irrelevant data

The second data quality issue to be touched by HAQIM is irrelevant data. It first uses a decision tree algorithm to establish the importance of each attribute relative to the target attribute in terms of predictive power. Then, it used an AutoEncoder based on a Neural Network to compress the dataset by reducing its dimensionality, weighted to the importance of the attributes in the set. The resulting dataset consists of a user-specified amount of attributes, which is 3 in the case of this example.

	encoded_feature_1	encoded_feature_2	encoded_feature_3	target
490	1,245653755	0,783328393	0,494739607	1
491	0,60786129	-0,102723709	-0,356499199	0
492	1,559992442	1,403134629	1,145036333	0
493	2,004810675	4,257640482	4,056327661	1
494	2,639838876	4,601036072	4,580122557	0
495	2,2122131	5,211265841	4,984579262	2
496	2,475554064	4,687473694	4,495087907	0
497	1,622220431	2,306705376	2,044209572	0
498	1,136727886	0,157952977	-0,072138573	0
499	0,785330582	0,264835581	-0,060474348	0
500	1,029644007	0,421666342	0,110742163	2
501	1,14542922	0,163266897	-0,06665167	0
502	1,251577346	0,349282866	0,066605157	0
503	1,050796365	0,25912624	-0,015806144	0
504	-0,011040823	-0,234180678	-0,341774981	0
505	1,3966435	0,750417412	0,520052816	0
506	0,688300431	-0,010805051	-0,003832621	0
507	1,184307228	0,531521497	0,236395758	0
508	0,712277489	-0,145662377	-0,316239875	0
509	0,315572043	-0,25841754	-0,425520479	0
510	0,641519362	-0,214394055	-0,434468163	0
511	0,795544335	-0,145772659	-0,290311688	0
512	1,298409088	0,314545782	0,036301803	0
513	1,796724493	1,468632713	1,155216634	1
514	1,82144398	1,545857495	1,222058573	0
515	1,432412076	0,578801918	0,302405636	0
516	1,296236875	0,712152437	0,430501108	0
517	1,062013406	0,376895459	0,070640817	0
518	0,83320467	0,05353813	-0,169852051	1
519	1,117240566	0,189339453	-0,080486675	0
520	0,52329772	-0,230562346	-0,428212219	0
521	0,737032971	0,787147541	0,478535684	1
522	0,690549058	-0,181710721	-0,35643903	0
523	0,7078646	-0,088408337	-0,134035064	0
524	-0,394458184	-0,060854203	-0,433363365	0

Figure 5.6: Screenshot of the example input data after handling irrelevant data

Chapter 6

Validation

In this chapter, we will describe the experiments that have been conducted as part of our method validation. We will first outline our experimental setup, followed by a description of comparison metrics used to quantify the results. Finally, we will discuss the limitations we have encountered during the experimentation phase of the research.

6.1 Experimental setup

The goal of the experiments is to validate the method we have designed in section 4, more specifically on the basis of effectiveness. As one of the primary goals of the method is to create a dataset enhanced in terms of data quality, to be used for analytical modelling, we will focus on the value our method creates for input data used for predictive modelling, which is what we have simulated during the experiments. In this section we will describe the key steps of our experiments in detail.

In figure 6.1 we have visualized the flow of activities of our experimental setup. It begins with our original datasets, of which six have been used in the experiments. Then, we have added artificially created data quality issues to these sets, using a structured approach to create irrelevant attributes and missing values. Then, three different scenarios were to be created for the comparison. First, we had the raw data with the data quality issues. Second, we had a dataset which has been enhanced using manual, primitive data improvement methods. Third, we had a dataset which has been enhanced by applying HAQIM. The datasets corresponding with the three different scenarios were subsequently used as input for a predictive modelling simulation, of which the predictive performances were compared.

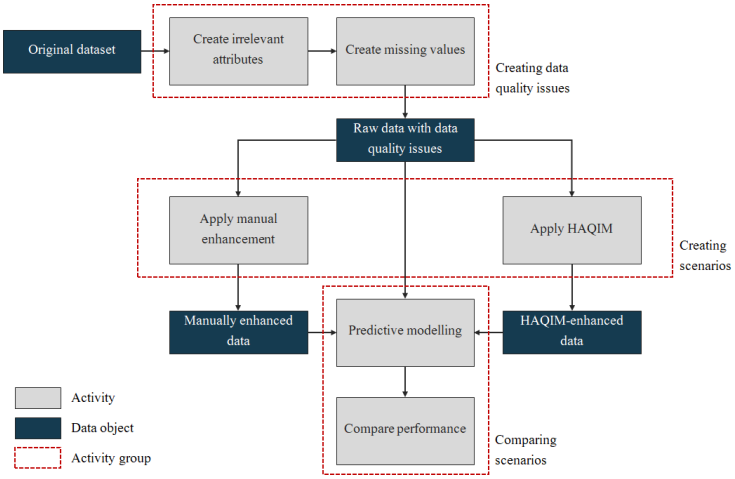


Figure 6.1: Flowchart visualizing the steps in the experimental setup

6.1.1 Datasets

For conducting the experiments, we have used a total of six different datasets. These datasets were either publicly available or retrieved from a private entity, who have consented to our use of their data for this research. The datasets differ in multiple aspects, such as amount of attributes, amount of data instances, type of organization to which the data belongs to and strategic purpose of the data. Full, elaborate descriptions of each dataset can be found in appendix A. In the appendix we have also included what the defined target variables are. A brief description of the acquired datasets are as follows:

1. **Network Operator Data:** This is a publicly retreived dataset of a Dutch energy network operator containing aggregated data per postal code. The dataset contains a multitude of attributes containing numerical and categorical data about information such as energy consumption as well as (the type of) electricity connections.

2. **Webshop Analytics Data:** This dataset has been retrieved from a business which manages a webshop specialized in travel accessories. The dataset contains both numerical attributes of data on the behaviour of website visitors and daily sales data.

3. **Dutch Fuel Prices:** This dataset has been retrieved from the public data repository of the government of the Netherlands. It contains three numerical attributes, each representing the average Dutch prices of a specific fuel type. The data is sorted by day.

4. **Bicycle Store Analytics Data:** This dataset as been retrieved from a business consisting of multiple physical store locations in the Nether-

lands. The dataset contains attributes detailing the user statistics of the website per day, corresponding to the amount of test drives requested at the bicycle stores.

5. **Avocado Prices:** This dataset has been retrieved from the public repository of Kaggle. This set contains the average avocado prices from different regions within the United States, as well as data containing the numbers of bags of avocados sold on a day.[109]

6. **Delivery Time Data:** This dataset has been retrieved from the Kaggle dataset repository as well. This set contains data instances representing different Indian restaurants. The attributes include information about food prices, customer ratings and food delivery times.[110]

6.1.2 Creating data quality issues

The six datasets we have used for experimentation have all been curated well and deterred from data quality issues prior to our retrieval of the data. No significant data quality issues which are relevant to the scope of the research have thus been detected. Hence, in order to simulate a situation in which data quality issues have not yet been handled or enhanced, we have added artificial data quality issues to the datasets in a structured manner. We have both added irrelevant data and missing values to the sets.

Irrelevant data

We have created a custom Python script to add irrelevant data to our datasets. We have first counted the amount of attributes in the set, and added an equal amount of synthetic attributes. These synthetic attributes contained randomly created values within the domain of the existing attributes. This way, we have provided a random, yet realistic range of attributes to our data.

Missing data

After having added synthetic irrelevant data attributes, we have added missing values to our data. For this activity, we have also programmed a customly made Python script. This script iterates through the data to remove values at random. However, to make the process not entirely random, we have added some checks and conditions to the script to ensure a balanced removal of data:

- A random number of attributes were allowed to contain missing values, with a minimum of 3 attributes and a maximum of 50% of the number of attributes.
- The maximum number of chained missing values cannot be greater than 2% of the total amount of data instances. Chained missing values are occurrences in which 2 or more values in a row are missing.

- For a maximum of 50% of the data instances there is one missing value.
- There cannot be more than 1 missing value in each data instance.

6.1.3 Creating scenarios

As mentioned in section 6.1, we have created three different scenarios to ensure a fair validation process of our method. These scenarios are as follows:

1. A raw dataset containing the data quality issues added as described in section 6.1.2. This dataset will enter the predictive modelling process without any data cleaning or preprocessing.
2. A dataset being enhanced through manual, primitive methods. We have created a custom Python script that enhances the data quality of each individual set based on its needs. We have imputed the missing values using a simple strategy: filling in the mean of the values in the related attribute. Subsequently, we have selected the top 5 attributes correlating to the target attribute.
3. A dataset which has been enhanced by applying HAQIM. The raw dataset, with our artificially created data quality issues, has been used as input for the tool. A resulting dataset, with missing values imputed using a random forest algorithm, and irrelevant data handled using a decision tree regressor and NN-based AutoEncoder, was set as the third scenario to be compared.

6.1.4 Comparing scenarios

In order to simulate an environment which fits the purpose of this research - enhancing the data quality of data used for strategic decision making - we have used the data for each scenario to be tested in the same predictive modelling process. We have created a simple linear regression, provided by SciKit-Learn with the default settings, to test the scenarios. The decision to use the default settings of the linear regression is based on simplicity and consistency. For these linear regression models we make predictions on the defined target attributes. These target attributes are mentioned per dataset in appendix A.

Using linear regression as a simulation technique in this context serves multiple purposes. Firstly, linear regression is a well-established and commonly applied predictive modeling technique that is often employed in real-world scenarios. By utilizing it within our research, we aim to create a realistic and relatable environment that mirrors actual decision-making processes.

Secondly, incorporating a predictive modeling step, such as linear regression, allows us to evaluate the impact of data quality issues on the performance of the model. In real-world scenarios, decision-making processes often rely on predictive models to make informed choices. By comparing the scenarios we will gain insights in how data quality issues can affect the quality of the final models. Additionally, by comparing the outcomes of linear regression models under different scenarios, we can quantify the influence of data quality improvements on the predictive performance.

6.2 Comparison metrics

This section will describe the different performance metrics which have been used to evaluate the effectiveness of each scenario in ensuring predictive strength. We have selected different performance metrics to use for our comparison (MAE, MSE, RMSE, r^2). The choice of performance metrics has been based on prior research which have included a comparison of ML models.[111,112,113] Below are the descriptions of all performance metrics along with a formula. The following symbols are used in these formulas:

- y = Real target variable
- $\hat{y}$ = Predicted target variable
- n = Amount of data instances

MAE

MAE (Mean Absolute Error) is calculated by taking the absolute difference between each predicted value and their corresponding true value. Then, this number is divided by the total amount of predicted instances. Its formula is noted as follows:

$$MAE = (\frac{1}{n}) \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

MSE

MSE (Mean Squared Error) is similar to the MAE. It takes the sum of all squared differences between the predicted values and their corresponding true values. Then, this number is divided by the total amount of predictions. Its formula is noted as follows:

$$MSE = (\frac{1}{n}) \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

RMSE

RMSE (Root Mean Squared Error) is similar to the MSE. It is calculated by taking the square root of MSE. Its formula is noted as follows:

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{n}}$$

r^2

The coefficient of determination, otherwise known as r^2, is used to indicate the proportion of variance in the dependent variable which can be explained by the independent variables. It is a value typically between 0 and 1, calculated by dividing the sum of squares of residuals with the total sum of squares. Its formula is noted as follows:

$$r^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Chapter 7

Results

In this chapter we will present the results obtained from our research. We will do this by firstly answering the research subquestions, which are facilitated by our literature review, method design and prototype development, found in section 3, 4 and 5 respectively. Then, we will display the results used to answer the main research question.

7.1 Research Subquestions

Here we will answer all research subquestions. We will reiterate the questions first, followed by an summation of the corresponding results.

7.1.1 SQ1

What are the characteristics of data typically used for analytical modelling, specifically for predictive purposes within a data-driven setting?

To answer this question, we must first describe what a data-driven setting is. In section 3.2 we have described the notion of data-driven organizations. A precise and agreed-upon definition in literature explaining data-driven organizations does not exist, as it is a multi-interpretable concept. We have included various definitions given by scientific articles in table 3.2, which are supplemented by the elements listed by Hupperz et al. (2021) which they claim are the building blocks of a data-driven organization. This material has allowed us to interpret the concept of data-driven businesses.

We interpret data-driven businesses as using data as a key resource within multiple aspects of the business. This refers to the existing processes and its business model. It must be supported by undergoing, or already having undergone, a digital transformation, embracing technology in key aspects of

the business. Data science is also integrated within the business, leveraging data to gain a competitive advantage. Additionally, a data-driven business leverages data for strategic decision making, also known as data-driven decision making (DDDM), which in turn enables the business to be characterized as being 'insights-driven' rather than merely being 'data-driven'. DDDM makes extensive use of analytical capabilities, such as predictive, prescriptive or descriptive modelling.

The type of data typically used for analytical modelling within this context is classified as 'Big Data'. A concise and agreed-upon definition of Big Data does not exist within scientific literature, as there is a multitude of aspects to describe this concept. Through a literature study we have identified the most frequently occurring aspects used to describe Big Data. These are volume, velocity, variety, value and veracity. Volume refers to the vast amount of data generated. Velocity refers to the rapid generation of data. Variety refers to the diversity and heterogeneity of data. Value refers to the usefulness of data, which can also be described as the value to the business. Veracity described the general quality of the data.

To answer the question, a data-driven setting can be described as we have done so above, specifically within the domain of DDDM when talking about analytical modelling. The data which appears within this described environment can be characterized using the aspects of big data. So, we have an environment where data is used as a key strategic resource for decision making. This data, used in a process for analytical modelling, can be described as having a high level of volume, velocity and variety, with a additional attention to its quality aspects and its value to the business.

7.1.2 SQ2

What common data quality issues exist within the context of data being used for analytical, and specifically predictive purposes within a data-driven setting?

In order to recognize what data quality issues commonly exist within this context, we must first understand data quality as a concept. The concept of data quality is subjective, and its definition can vary widely depending on the context in which the data is used for and from what point of view it is being assessed. A general and holistic definition of data quality is "Fitness for use", which implies its subjectiveness. However, data quality can be described by a set of dimensions through which it can be assessed. During our literature review we have examined 12 papers which have mentioned 49 unique dimensions of data quality. From this collection of data quality dimensions, the 6 most-frequently cited were currency, consistency, complete-

ness, accuracy, relevancy and accessibility.

In terms of data quality issues, many exist within a range of contexts, also depending on how the data in question is assessed. For example, frequently found data quality issues vary in data which is used in an ML or an ETL (Extract Transform Load) context. We have examined a total of 13 articles, mentioning 73 unique data quality issues. During our literature study we have merged some mentioned issues due to a difference in wording or issues being described similarly. In a general context, the 8 most frequently mentioned data quality issues were:

- **Missing data:** Refers to all forms in which data is missing, on multiple levels. This ranges from some values in a table not being present to entire databases being unavailable. However, this most commonly refers to data entries not being present.
- **Duplicate data:** This entails to full data entries appearing more than a single occasion, either in the same data object or across different ones.
- **Wrong data type:** This issue occurs when values are assigned to incompatible or incorrect data types. For example numerical values being stored as a string.
- **Uniqueness violation:** This issue refers to situations where a data instance is expected to be unique, yet possess duplicate values.
- **Syntax violation:** This issue refers to situations where a data entry does not conform to the specified formatting rules or syntax specification.
- **Misspelling:** Misspellings describe errors raised as a result of the inaccurate spelling of certain values.
- **Irrelevant data:** This refers to the presence of unnecessary or inadequate meaningful data.
- **Incorrect data:** This issue is an umbrella term for data quality issues that do not represent the real world values in terms of factual correctness.

A full overview of the data quality issues we have collected from our examined papers can be found in appendix B.3.

7.1.3 SQ3

What machine learning techniques can be effectively applied to tackle data quality issues commonly found within the described context?

Before being able to answer this question, we must first understand which data quality issues are ML-solvable within this context. We have divided our

collection of data quality issues into two main categories: those which are solvable with context knowledge, and those which are solvable without context knowledge. For data quality issues solvable without context knowledge, we have distinguished issues which are ML-solvable and non-ML-solvable. The category of non-ML-solvable issues contain both issues which cannot be solved using ML algorithms at all, and those that can but where the usage of ML is deemed significantly inefficient. This categorization has been visualized in figure 3.5. For our research, we have focused on those that are ML-solvable within the category of data quality issues solvable without context knowledge: missing values, wrong data types, irrelevant data and outliers.

We have examined a total of 11 papers to assess the use of ML algorithms to handle the aforementioned data quality issues. Retrieved from these papers, the top picks of algorithms to address these specific data quality issues using ML-based techniques were:

- **Missing data:** Random Forest
- **Irrelevant data:** Neural Network-based AutoEncoder
- **Wrong data types:** Neural Network
- **Outliers:** KNN

The full list of papers we have analyzed and their corresponding picks of algorithms can be found in appendix B.4.

7.1.4 SQ4

How can we apply Machine Learning algorithms to address specific data quality issues in a holistic manner?

We have answered this question by designing the method as described in section 4. The method's architecture has been visualized in figure 4.1. We have created this architecture based on a set of requirements described in section 4.1. The main method for eliciting these requirements has been through reviewing literature, as well as prototyping sessions. These requirements have subsequently been verified by a field expert.

The architecture describes a method which enables holistically tackling data quality issues within the context of data used as a strategic asset in a data-driven environment. This method includes two main components: the Data Enhancement Component and Data Quality Assessment Component.

The Data Enhancement Component possesses several features corresponding to specific data quality issues to be solved. Each of these create several outputs. Firstly, after an algorithm has concluded its training phase, an ML model is stored for future usage. Secondly, an enhanced dataset is created per attribute. Finally, technical reporting data which includes information such as training time and model accuracy is collected during the runtime of a feature. After this component has been concluded, a fully enhanced dataset is generated.

The function of the Data Quality Assessment Component is to evaluate and quantify data quality of the input. It encompasses a set of functions to assess various dimensions of data quality, as we have established in the literature review. Within our method, the dimensions include completeness, consistency, timeliness, relevancy and deduplication. In specific, the Data Quality Assessment Component performs a series of analyses and checks on the input dataset to identify the overall health in terms of data quality. It assesses, besides a quantification of the dimensions of data quality, a before-and-after situation of the usefulness of the data in terms of analytical performance. Additionally, it also collects technical data such as training time during the enhancement phase and subsequently generates a quality label based on all gathered information.

The component generates a report, which is its key output product. It presents insights to various stakeholders, such as (strategic) decision-makers, data engineers and -scientists to gain a comprehensive understanding of the strengths and limitations of their datasets.

7.1.5 SQ5

How can a prototype be constructed which implements this approach?

In section 5 we have described HAQIM, which is the prototype we have developed to serve as a proof-of-concept for our envisioned method. HAQIM is an acronym for **H**olistic **A**ssessment of **Q**uality and **I**mprovement **M**ethod. HAQIM is able to read large datasets as input and subsequently assess its perceived data quality based on the dimensions of data quality as described in section 3.3.1 and to enhance the dataset by imputing missing values and handling irrelevant data, both with the use of ML techniques.

The development of HAQIM has been enabled by several factors. First, the prototype has been developed using the Python programming language. Python is a programming language with a comprehensive standard library

and offers solutions for ML-based methods and adequately handles datasets. Within Python, we have made use of multiple packages which have enabled us to develop HAQIM. First, we have used Pandas, using its native DataFrame feature to operate on large datasets. Secondly, SciKit-Learn has been used to implement ML algorithms and subsequently measure their predictive performances using SciKit-Learn's performance metrics functionalities. Thirdly, we have used NumPy to assist in data operations. Finally, PyFPDF has been used to create a PDF, using Python code, which includes all reporting data as described in section 4.2.2.

7.2 Research Question

How can we develop a novel method based on machine learning techniques to effectively improve multiple data quality issues at once, for data found in a data-driven context mainly used for analytical purposes?

To answer the main research question, we have conducted several experiments to measure the effectiveness of the model through our proof-of-concept (HAQIM), as we have described in section 2.4. For each dataset available to our research we have measured the performance metrics in terms of predictive power in all scenarios. The scenarios included the following types of datasets:

- **Scenario I:** The original dataset without any data quality enhancement.
- **Scenario II:** A dataset which has been enhanced using manual, simplistic methods for data quality improvement.
- **Scenario III:** A dataset which has been enhanced using HAQIM.

A full overview of all performance metrics we have measured (MAE, MSE, RMSE and r^2) can be found in Appendix E. In figures 7.1 to 7.6 we have included graphs showing the difference in RMSE scorings per dataset used in our validation process. These graphs illustrate the predictive power in terms of accuracy for the models created. A lower RMSE is an indication of a higher accuracy. We have noticed varying results, depending on the dataset. Most datasets have a lower RMSE in scenario III than scenario I, with an exception of the Dutch Fuel Prices Data and Bicycle Store Data. Similarly, most datasets have a lower RMSE in scenario III than scenario II, with an exception of the Webshop Analytics Data and the Bicycle Store Data. However, this difference is minimal. We can see an average decrease of 3.4% in RMSE scorings across all datasets when comparing the predictive performances of scenario I and II. When comparing scenario I and III, the decrease is 9.1%.

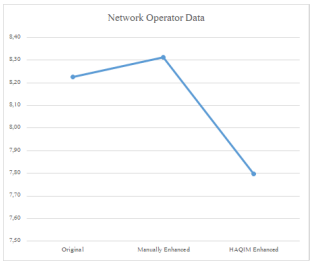


Figure 7.1: RMSE difference for validation scenarios (Network Operator Data)

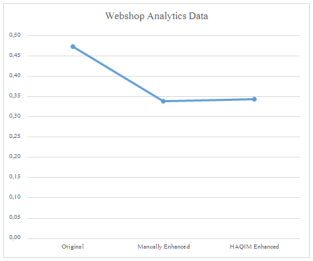


Figure 7.2: RMSE difference for validation scenarios (Webshop Analytics Data)

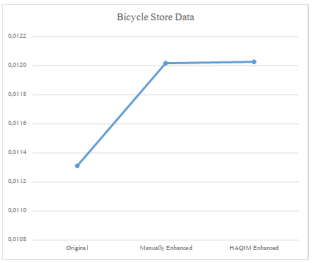


Figure 7.3: RMSE difference for validation scenarios (Bicycle Store Data)

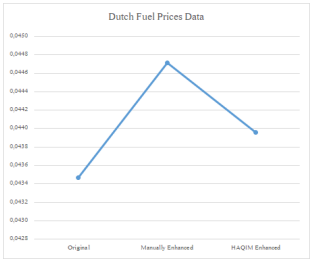


Figure 7.4: RMSE difference for validation scenarios (Dutch Fuel Prices Data)

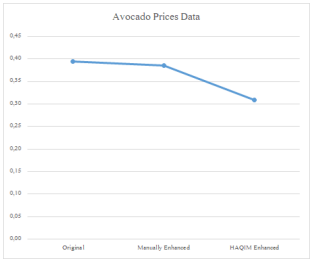


Figure 7.5: RMSE difference for validation scenarios (Avocado Price Data)

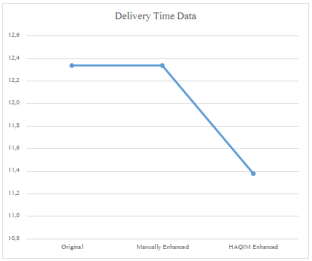


Figure 7.6: RMSE difference for validation scenarios (Delivery Time Data)

While MAE and MSE can be used to explain the accuracy of the models' prediction, r^2 explains the proportion of the variance of the attributes in the dataset that explains the target attribute. In figure 7.7 we have visualised the differences in r^2 scorings per scenario for all datasets we have used in our validation process. We can see that for most datasets, the r^2 scorings linger around the same values, with slight increases for scenario III. The exceptions are the Webshop Analytics Data, which shows a slight decrease in r^2 scorings, and the Avocado Prices Data, which shows an enormous increase of 583%.

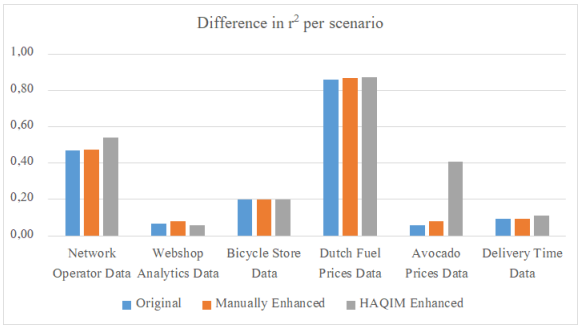


Figure 7.7: Differences in r^2 scorings per scenario per dataset

We have also investigated the relation between the sizes of datasets and the change in percentage of their r^2 scores comparing scenario I and III. We have visualised this relation in a scatterplot in figure 7.8. We can see four datasets grouped at the bottom left corner, with extreme outliers in the top left and bottom right corners. Including the outliers, using the pearson r to calculate the relation between change in r^2 scores and amount of data instances, we get a correlation coefficient of -0.065. Excluding the outliers, the correlation coefficient is 0.929. The outliers have been marked with a red circle in figure 7.8

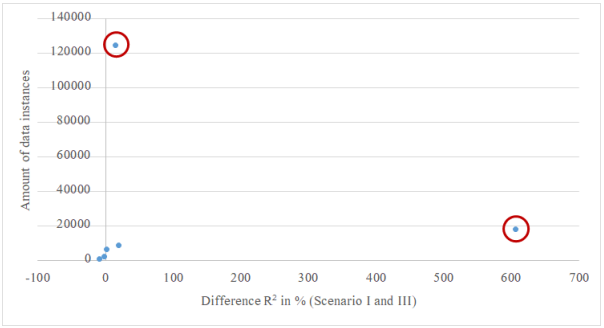


Figure 7.8: Change in r^2 scorings and the size of the datasets used

Chapter 8

Discussion

In this chapter we will discuss the results from our research, which have been mentioned in chapter 7. For this research we have investigated the use of ML-based techniques to improve data quality, specifically for data used for strategic decision making, typically found within data-driven environments. During the first element of the research process we have conducted an extensive literature study on the topics of data-driven businesses, big data, data quality and machine learning. During this literature study we have reviewed a multitude of scientific, and to an extent non-scientific sources to provide an in-depth examination of these topics, and how they ultimately relate to designing a methodology for enhancing data quality within the given context.

We have first provided a description of data-driven businesses with an overview of its characteristics. We have described being data-driven as using data as a key resource within multiple aspects of a business. Applying Data-Driven Decision Making (DDDM) will typify a business to be rather 'insights-driven' than merely being 'data-driven'. We have found that big data is a significant component herein. The concept of big data lacks a concise definition, however can be described using its aspects of volume, velocity, variety, value and veracity.

We have also researched the concept of data quality. Though lacking a concise definition as well, we have used the description of 'Fitness for use' throughout this thesis. We have researched a multitude of scientific papers researching data quality, and consolidated a list of most frequently mentioned dimensions of data quality. Additionally, we have reviewed scientific literature to consolidate a list of data quality issues being analysed in literature. The final stage of the literature review process involved a study of machine learning. It first included a description of the concept of ML in itself, followed by a study to which data quality issues were to be handled within the scope of

this research. We have made a distinction between data quality issues which could be solved contextless with the use of machine learning and those that are not. Subsequently we have matched these data quality issues to ML algorithms optimal for solving these.

After we had laid a foundation of knowledge based on literature, we have designed a method suitable for holistically enhancing data quality based on these issues. We have used the data from our literature review, prototyping and validation from domain experts to consolidate a list of requirements for our design. Based on these requirements we have designed the architecture of this method. Based on this architecture we have created a proof-of-concept in the form of a working prototype, named HAQIM (**H**olistic **A**ssessment of **Q**uality and **I**mprovement **M**ethod). This prototype has been programmed using Python, as well as a multitude of publicly available packages. However limited due to time and resource constraints, HAQIM has incorporated several key elements from the method architecture.

The final element of this thesis research has been evaluating the method on its effectiveness, through empirical experimentation on our proof-of-concept prototype HAQIM. We have gathered six different datasets, from either public or private repositories. For each dataset, we have validated the method using three different scenarios. One using the original datasets with data quality issues, one using a dataset enhanced through manual, simplistic data quality improvement techniques and one using HAQIM to enhance their data quality. The approach we have taken for validation of our method has been to apply a simple method for predictive modelling. We have used all scenarios for each dataset in a linear regression model. We have chosen this validation strategy for reasons of simplicity and consistency. The results varied strongly, however overall positive. For most datasets, the predictive performance of datasets being enhanced by HAQIM have been proven to harness more accurate results than being enhanced manually or not at all. However, these results have illustrated that the performance for predictive modelling has not been altered drastically. Also, we have suspected an association between a positive change in predictive power and the sizes of datasets being used in our validation process. The larger the dataset, the higher the change in r^2 is, meaning an increase in predictive power. However, as we have only made use of 6 datasets during our validation phase, it is too early to reach a conclusion.

Additionally, due to the nature and amount of limitations we have encountered during the research, we have found an uncertainty in the question if the overall results would have been similar would other choices have been made. Examples are the choice of features of our data quality enhancement method

to incorporate into the prototype, HAQIM, or the approach we have taken to manually add data quality issues to our collection of datasets. However, all results should be reproducible with the datasets we have used in our validation phase. Due to the random nature of adding data quality issues there is a possibility of variation.

8.1 Limitations

During this research we have been confronted by various limitations. The primary origin of these limitations stem from general time and resource constraints. This research has been part of a thesis project of a masters' study: ICT in Business at Leiden University. This implies that the research had to be completed within a semester, and in terms of resources no budget was available. This, in turn, caused some elements of this thesis to not be as elaborate as it could have been would the time and overall resource amount be increased. In this section we will detail the specific limitations under which the research was performed.

8.1.1 Literature Study

The literature study has been subjected to several limitations. We have reviewed a number of papers for a variety of subjects. For studying the subjects of big data and data-driven businesses we have reviewed 20 papers, for data quality and its dimensions we have studied 16 papers, for data quality issues we have studied 13 papers and for ML and its data quality applications we have studied 21. While the total number of reviewed papers, at 70, is on the higher side of a thesis literature study, broken down in the different subjects we could have included more scientific papers to review. This would have, in turn, resulted in a more comprehensive collection, which could have enabled us to develop more in-depth analyses of the subjects.

Another major limitation we have faced during the literature study has been the verification of the data we have collected from all scientific journals. Theoretical knowledge retrieved from scientific journals do not necessarily represent practical experience of domain experts. A method to dispel this gap between practical and theoretical knowledge is by conducting interviews or validate the findings of the literature study.

8.1.2 Method & Prototype

The phases of the research in which our data quality enhancement-method has been designed and the development of our prototype have been subjected to a multitude of limitations as well. First of all, in this research we have

not followed the requirements engineering-process as rigorously. The main methods of requirement elicitation included prototyping and collecting data from our literature study. A more comprehensive form of requirement elicitation would have included workshops, interviews or surveys with domain experts or other individuals with practical experience. We have only verified the requirements once, yet we believe a more elaborate validation protocol would have allowed us a superior list of consolidated requirements.

Most limitations have been encountered during the development of our prototype, HAQIM. We have envisioned a holistic method for enhancing multiple data quality issues, based on machine learning techniques. The following list provides an overview of aspects in we which had to limit the workings of the prototype:

- **Data input:** Ideally the method would allow for integration within any data stream leading up to the decision making process. This would include the handling a variety of data (file) types and formats to be used. In the prototype we have only included a method to read .csv files containing data presented in a tabular format.
- **Handling data quality issues:** The prototype only is able to handle the data quality issues of missing values and irrelevant data. In the method proposed by our research we have also included contextless ML-solvable data quality issues such as wrong data formats and outliers.
- **Dimensions of data quality assessment:** We have included the dimensions of currency, deduplication, relevancy and completeness within the Data Quality Assessment Component of our prototype. In the method design we have included consistency as well. Since assessing the input data on this data quality dimension required more complex functionalities to implement, we have excluded it from the prototype.
- **User interface:** Ideally any program based on this method would have included an accessible user interface. As we have developed the prototype primarily for the purpose of validating the method, this feature has not been implemented.

8.1.3 Validation

The validation phase of our research process has been somewhat limited. To validate our proposed method by conducting experiments, we have tested three different scenarios for six available datasets. The first limitation of this phase has been not having more datasets available for experimentation. Each dataset has been delicately investigated for suiting the purpose of our research, containing an adequate amount of data and containing data types suitable for validation. This is a relatively time-intensive process. Being

able to append more available datasets to the validation process would have required more time than what was feasible. However, a larger number of datasets would have allowed us to create more in-depth analyses of relations corresponding to properties of the datasets. One example is the relation between dataset size and changes in predictive performance after enhancing. Currently, we have suspected a relation between these variables, however with the current data it can only support premature conclusions.

As all datasets have been carefully curated before retrieval, we have added artificially created data quality issues in an organized, yet random manner. The approach we have taken to add these quality issues to the datasets does not necessarily reflect truly realistic scenarios. This study had thus been limited by the fact that we have not acquired datasets with real-world data quality issues. We have secondly limited our validation by not including multiple versions of the same dataset with different variations or types of data quality issues.

Additionally, we have not investigated the statistical properties of the data beforehand, such as patterns within the distribution of the data. Similarly, the artificially created datasets do not possess such patterns as well.

8.2 Future Research

In this section we will describe our recommendations for future research. Firstly, handling the limitations we have encountered during this research as described in section 8.1 could provide a good base for future research. While we propose a method, we have not been able to prove its effectiveness in its entirety. The proof-of-concept we have developed in the form of a working prototype only includes a limited number of functionalities. By validating the effectiveness of this prototype to enhance data quality with machine learning techniques, we have proven the method to be effective to a certain extent. Adding more features to the prototype is a valuable lead for future research.

Secondly, including more types of data quality issues to be handled would make an interesting research. In section 3.4.2 we have made a distinction between data quality issues which can only be solved with specific context knowledge and those which can be solved regardless of context. Within the latter, we have categorized data quality issues that cannot be solved with ML (or applying machine learning would be significantly inefficient) and those that can be solved efficiently with ML. In our proposed method, we have only dealt with the latter category. Our recommendation is to explore other types of data quality issues, such as those that require context knowledge.

By integrating the method with data sources that would explain elements of a specific business' context, this could be realized.

Thirdly, as we have suspected an association between the sizes of datasets and the change in predictive power when they have been processed by HAQIM, it is an interesting topic for future research. As our validation phase only included a small number of different datasets, namely 6, it is too early to form a conclusion on this. A collection including a larger number of datasets would allow for a more profound research.

Our fourth and final recommendation for future research is exploring the use of AutoML (Automated Machine Learning) for creating a method to enhance data quality. AutoML is a concept in which parts of the machine learning process are covered automatically by the algorithm. Based on the (type of) data and the problem at hand, an AutoML implementation covers the steps of data preprocessing, model selection, hyperparameter tuning, training and model validation without the need of human intervention. It has shown to be more effective than traditional ways of implementing machine learning, as proven by Cesar de Sá et al. (2022).[82]

Chapter 9

Conclusion

For this thesis research project, we have studied the use of machine learning to enhance data quality, specifically for handling data quality issues appearing within data used for strategic decision making. We have conducted an extensive research process involving two main elements: a literature study and the development of a methodology. During the literature study we have reviewed a multitude of scientific, and to an extent non-scientific sources to provide an in-dept overview of data-driven businesses, big data, data quality and machine learning. Our literature study has allowed us to offer a reflection of the current state of these topics in scientific literature.

We have culminated the theoretical knowledge attained from our literature study into a method design which has been realised through a proof-of-concept. This proof-of-concept has been accomplished by developing HAQIM, our working prototype. By following a validation process, we have proven our method to be effective in dealing with data quality issues using machine learning. This has allowed us to succeed in completing the primary aim of our research: exploring the possibilities of machine learning techniques to enhance data quality.

Bibliography

[1] G. K. Tayi and D. P. Ballou, "Examining data quality," *Communications of the ACM*, vol. 41, pp. 54–57, 2 Feb. 1998. DOI: 10.1145/269012.269021.

[2] L. L. Pipino, Y. W. Lee, R. Y. Wang, and R. Y. Yang, "Data quality assessment," *Communications of the ACM*, vol. 45, pp. 211–218, 4ve 2002.

[3] K. F. Luchtenberg and Q. V. Vu, "The 2008 financial crisis: Stock market contagion and its determinants," *Research in International Business and Finance*, vol. 33, pp. 178–203, 2015, ISSN: 02755319. DOI: 10.1016/j.ribaf.2014.09.007.

[4] V. Acharya, T. Philippon, M. Richardson, and N. Roubini, "The financial crisis of 2007-2009: Causes and remedies," *Financial Markets, Institutions and Instruments*, vol. 18, pp. 89–137, 2 May 2009, ISSN: 09638008. DOI: 10.1111/j.1468-0416.2009.00147_2.x.

[5] L. Francis and V. R. Prevosto, "Data and disaster: The role of data in the financial crisis," *Casualty Actuarial Society E-Forum, Spring 2010*, vol. 62, 2010.

[6] M. Janssen, H. van der Voort, and A. Wahyudi, "Factors influencing big data decision-making quality," *Journal of Business Research*, vol. 70, pp. 338–345, Jan. 2017, ISSN: 01482963. DOI: 10.1016/j.jbusres.2016.08.007.

[7] C. E. Morr and H. Ali-Hassan, *Descriptive, predictive, and prescriptive analytics*, 2019. DOI: 10.1007/978-3-030-04506-7_3.

[8] R. Schutt and C. O'Neil, *Doing Data Science: Straight Talk from the Frontline*. O'Reilly Media, Inc., 2013.

[9] P. M. Hartmann, M. Zaki, N. Feldmann, and A. Neely, "Capturing value from big data – a taxonomy of data-driven business models used by start-up firms," *International Journal of Operations and Production Management*, vol. 36, pp. 1382–1406, 10 2016, ISSN: 17586593. DOI: 10.1108/IJOPM-02-2014-0098.

[10] V. Grover, R. H. Chiang, T. P. Liang, and D. Zhang, "Creating strategic business value from big data analytics: A research framework," *Journal of Management Information Systems*, vol. 35, pp. 388–423, 2 Apr. 2018, ISSN: 1557928X. DOI: 10.1080/07421222.2018.1451951.

[11] B. T. Hazen, C. A. Boone, J. D. Ezell, and L. A. Jones-Farmer, "Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications," *International Journal of Production Economics*, vol. 154, pp. 72–80, 2014, ISSN: 09255273. DOI: 10.1016/j.ijpe.2014.04.018.

[12] K. L. Keller and R. Staelin, "Effects of quality and quantity of information on decision effectiveness," *Journal of Consumer Research*, vol. 14, pp. 200–213, 2 Sep. 1987.

[13] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Computing Surveys*, vol. 41, 3 Jul. 2009, ISSN: 03600300. DOI: 10.1145/1541880.1541883.

[14] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," vol. 14, Committee on Data for Science and Technology, 2015. DOI: 10.5334/dsj-2015-002.

[15] V. N. Gudivada, J. Ding, and A. Apon, "Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations," *International Journal on Advances in Software*, vol. 10, pp. 1–20, 1 2017. [Online]. Available: https://www.researchgate.net/publication/318432363.

[16] W. Shi *et al.*, "Improving power grid monitoring data quality: An efficient machine learning framework for missing data prediction," IEEE, Aug. 2015, pp. 417–422, ISBN: 978-1-4799-8937-9. DOI: 10.1109/HPCC-CSS-ICESS.2015.16. [Online]. Available: https://ieeexplore.ieee.org/document/7336197/.

[17] D. Reinsel, J. Gantz, and J. Rydning, "The digitization of the world from edge to core," IDC, 2018.

[18] A. Engelbrecht, J. Gerlach, and T. Widjaja, "Understanding the anatomy of data-driven business models - towards an empirical taxonomy," 2016. [Online]. Available: http://aisel.aisnet.org/ecis2016_rphttp://aisel.aisnet.org/ecis2016_rp/128.

[19] A. Gourévitch, L. Faeste, E. Baltassis, and J. Marx, "Data-driven transformation - accelerate at scale now," Boston Consulting Group, May 2017. [Online]. Available: https://www.bcg.com/publications/2017/digital-transformation-transformation-data-driven-transformation.

[20] M. Favaretto, E. de Clercq, C. O. Schneble, and B. S. Elger, "What is your definition of big data? researchers' understanding of the phenomenon of the decade," *PLoS ONE*, vol. 15, 2 2020, ISSN: 19326203. DOI: 10.1371/journal.pone.0228987.

[21] A. D. Mauro, M. Greco, and M. Grimaldi, "A formal definition of big data based on its essential features," *Library Review*, vol. 65, pp. 122–135, 3 Apr. 2016, ISSN: 00242535. DOI: 10.1108/LR-06-2015-0061.

[22] G. Manogaran, C. Thota, and D. Lopez, *Human-computer interaction with big data analytics*, 2022. DOI: 10.4018/978-1-6684-3662-2.ch076.

[23] R. Bean, *Has progress on data, analytics, and ai stalled at your company*, 2023. [Online]. Available: https://hbr.org/2023/01/has-progress-on-data-analytics-and-ai-stalled-at-your-company.

[24] P. Mikalef, M. Boura, G. Lekakos, and J. Krogstie, "Big data analytics and firm performance: Findings from a mixed-method approach," *Journal of Business Research*, vol. 98, pp. 261–276, May 2019, ISSN: 01482963. DOI: 10.1016/j.jbusres.2019.01.044.

[25] H. Lee, E. Kweon, M. Kim, and S. Chai, "Does implementation of big data analytics improve firms' market value? investors' reaction in stock market," *Sustainability (Switzerland)*, vol. 9, 6 2017, ISSN: 20711050. DOI: 10.3390/su9060978.

[26] T. Redman, "The impact of data quality on the typical enterprise," *Communications of the ACM*, vol. 41, pp. 79–82, 2 Feb. 1998. DOI: 10.1145/269012.269025.

[27] D. Dey and S. Kumar, "Reassessing data quality for information products," *Management Science*, vol. 56, pp. 2316–2322, 12 Dec. 2010, ISSN: 00251909. DOI: 10.1287/mnsc.1100.1261.

[28] T. Redman, *Bad data costs the u.s. $3 trillion per year*, 2016. [Online]. Available: https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year.

[29] V. Charnock, "Electronic healthcare records and data quality," *Health Information and Libraries Journal*, vol. 36, pp. 91–95, 1 Mar. 2019, ISSN: 14711842. DOI: 10.1111/hir.12249.

[30] R. Y. Wang, "A product perspective on total data quality management," *Communications of the ACM*, vol. 41, pp. 58–65, 2 Feb. 1998. DOI: https://doi.org/10.1145/269012.269022.

[31] D. Hayn *et al.*, "Predictive analytics for data driven decision support in health and care," *it - Information Technology*, vol. 60, pp. 183–194, 4 Aug. 2018, ISSN: 2196-7032. DOI: 10.1515/itit-2018-0004.

[32] B.-C. Ubaldi, C. V. Ooijen, and B. Welby, "A data-driven public sector: Enabling the strategic use of data for productive, inclusive and trustworthy governance," Organisation for Economic Co-operation and Development, 2019. DOI: 10.1787/09ab162c-en. [Online]. Available: https://dx.doi.org/10.1787/09ab162c-en.

[33] J. vom Brocke, A. Hevner, and A. Maedche, *Introduction to design science research*, 2020. DOI: 10.1007/978-3-030-46781-4_1.

[34] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of Management Information Systems*, vol. 24, pp. 45–77, 3 Dec. 2007, ISSN: 07421222. DOI: 10.2753/MIS0742-1222240302.

[35] T. Hall, S. Beecham, and A. Rainer, "Requirements problems in twelve software companies: An empirical analysis," vol. 149, Oct. 2002, pp. 153–160. DOI: 10.1049/ip-sen:20020694.

[36] A. Aurum and C. Wohlin, *Requirements engineering: Setting the context*, 2005. DOI: 10.1007/3-540-28244-0_1.

[37] I. Sommerville, "Integrated requirements engineering: A tutorial," *IEEE Software*, vol. 22, pp. 16–23, 1 2005. DOI: 10.1109/MS.2005.13.

[38] E. Mohammadi and A. Karami, "Exploring research trends in big data across disciplines: A text mining analysis," *Journal of Information Science*, vol. 48, pp. 44–56, 1 Feb. 2022, ISSN: 17416485. DOI: 10.1177/0165551520932855.

[39] O. Ylijoki and J. Porras, "Perspectives to definition of big data: A mapping study and discussion," *Journal of Innovation Management Ylijoki*, vol. 4, pp. 69–91, 2016, ISSN: 2183-0606.

[40] M. Bulger, G. Taylor, and R. Schroeder, "Data-driven business models: Challenges and opportunities of big data," Oxford Internet Institute, Sep. 2014.

[41] Gartner, *Big data*. [Online]. Available: from%20https://www.gartner.com/en/information-technology/glossary/big-data.

[42] R. Hilbig, S. Hecht, and B. Etsiwah, "Berlin start-ups-the rise of data-driven business models," Dec. 2018. [Online]. Available: https://www.researchgate.net/publication/329529109_Berlin_Start-ups_-_The_Rise_of_Data-Driven_Business_Models.

[43] H. Hannila, R. Silvola, J. Harkonen, and H. Haapasalo, "Data-driven begins with data; potential of data assets," *Journal of Computer Information Systems*, vol. 52, pp. 29–38, 1 2022. DOI: https://doi.org/10.1080/08874417.2019.1683782.

[44] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, pp. 137–144, 2 2015, ISSN: 02684012. DOI: 10.1016/j.ijinfomgt.2014.10.007.

[45] M. Al-Mekhlal and A. A. Khwaja, "A synthesis of big data definition and characteristics," Institute of Electrical and Electronics Engineers Inc., Aug. 2019, pp. 314–322, ISBN: 9781728116631. DOI: 10.1109/CSE/EUC.2019.00067.

[46] C. Cartledge, "How many vs are there in big data?" 2016. [Online]. Available: http://www.clc-ent.com/TBDE/Docs/vs.pdf.

[47] H. Boinepelli, *Applications of big data*, 2015. DOI: 10.1007/978-81-322-2494-5_7.

[48] T. Rabl and H.-A. Jacobsen, "Big data generation," 2012, pp. 20–27. DOI: 10.1007/978-3-642-53974-9_3. [Online]. Available: http://msrg.org.

[49] C. L. Jurkiewicz, "Big data, big concerns: Ethics in the digital age," *Public Integrity*, 2018, ISSN: 15580989. DOI: 10.1080/10999922.2018.1448218.

[50] R. Hillard, "The information drive enterprise," *Telecommunications Journal of Australia*, vol. 61, pp. 48.1–48.7, 3 2011.

[51] M. Hupperz, I. Gür, F. Möller, and B. Otto, "What is a data-driven organization?," 2021. [Online]. Available: https://www.researchgate.net/publication/351282206.

[52] M. J. Diván, "Data-driven decision making," IEEE, 2017, pp. 50–56, ISBN: 9781538605141. DOI: 10.1109/ICTUS.2017.8285973.

[53] E. Brynjolfsson, L. M. Hitt, and H. H. Kim, "Strength in numbers: How does data-driven decisionmaking affect firm performance?" 2011. DOI: https://dx.doi.org/10.2139/ssrn.1819486.

[54] A. Bousdekis, K. Lepenioti, D. Apostolou, and G. Mentzas, "A review of data-driven decision-making methods for industry 4.0 maintenance applications," 2021. DOI: 10.3390/electronics. [Online]. Available: https://doi.org/10.3390/electronics.

[55] F. Provost and T. Fawcett, "Data science and its relationship to big data and data-driven decision making," *Big Data*, vol. 1, pp. 51–59, 1 Mar. 2013, ISSN: 2167-6461. DOI: 10.1089/big.2013.1508.

[56] V. M. Bhimavarapu, R. S. Gautam, and V. M. Bhimavarapu, "Data driven decision making: Application in finance," *IRE Journals*, vol. 5, pp. 52–56, 12 2022, ISSN: 2456-8880. [Online]. Available: https://www.researchgate.net/publication/361814510.

[57] P. Oliveira, P. R. Henriques, and F. Rodrigues, "A formal definition of data quality problems," 2005. [Online]. Available: https://www.researchgate.net/publication/220918803.

[58] H. Galhardas, D. Florescu, D. Shasha, E. Simon, and C.-A. Saita, "Declarative data cleaning: Language, model and algorithms," 2001.

[59] W. Fan, "Data quality: From theory to practice," *SIGMOD Record*, vol. 44, pp. 7–18, 3 2015.

[60] M. Scannapieco, P. Missier, and C. Batini, "Data quality at a glance," Datenbank-Spektrum, 2015. [Online]. Available: https://www.researchgate.net/publication/220102773_Data_Quality_at_a_Glance.

[61] N. Abdullah, S. A. Ismail, S. Sophiayati, and S. M. Sam, "Data quality in big data: A review," *Int. J. Advance Soft Compu. Appl*, vol. 7, 3 2015, ISSN: 2074-8523.

[62] C. Fox, A. Levitin, and T. Redman, "The notion of data and its quality dimensions," *Information Processing Management*, vol. 30, pp. 9–19, I 1994. DOI: 10.1016/0306-4573(94)90020-5.

[63] F. Sidi, P. H. S. Panahy, L. S. Affendey, M. A. Jabar, H. Ibrahim, and A. Mustapha, "Data quality: A survey of data quality dimensions," *2012 International Conference on Information Retrieval Knowledge Management*, 2012. [Online]. Available: https://doi.org/10.1109/InfRKM.2012.6204995.

[64] S. W. Tee, P. L. Bowen, P. Doyle, and F. H. Rohde, "Factors influencing organizations to improve data quality in their information systems," *Accounting and Finance*, vol. 47, pp. 335–355, 2 Jun. 2007, ISSN: 08105391. DOI: 10.1111/j.1467-629X.2006.00205.x.

[65] W. W. Eckerson, "Data quality and the bottom line," 101communications LLC, 2002. [Online]. Available: http://download.101com.com/pub/tdwi/Files/DQReport.pdf.

[66] M. D. Wilkinson *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, Mar. 2016, ISSN: 20524463. DOI: 10.1038/sdata.2016.18.

[67] S. Samonas and D. Coss, "The cia strickes back: Redefining confidentiality, integrity and availability in security," *Journal of Information Systems Security*, vol. 10, pp. 21–45, 3 2014, ISSN: 1551-0123. [Online]. Available: www.jissec.org.

[68] M. Souibgui, F. Atigui, S. Zammali, S. Cherfi, and S. B. Yahia, "Data quality in etl process: A preliminary study," vol. 159, Elsevier B.V., 2019, pp. 676–687. DOI: 10.1016/j.procs.2019.09.223.

[69] W. Kim, B.-J. Choi, S.-K. Kim, and D. Lee, "A taxonomy of dirty data," *Data Mining and Knowledge Discovery*, vol. 7, pp. 81–99, 2003.

[70] X. Chu, I. F. Ilyas, S. Krishnan, and J. Wang, "Data cleaning: Overview and emerging challenges," vol. 26-June-2016, Association for Computing Machinery, Jun. 2016, pp. 2201–2206, ISBN: 9781450335317. DOI: 10.1145/2882903.2912574.

[71] L. Berti-Équille, *Measuring and modelling data quality for quality-awareness in data mining*, 2007. DOI: 10.1007/978-3-540-44918-8_5.

[72] V. Kellen, "Business performance measurement," 2003, pp. 1–36. [Online]. Available: http://www.depaul.edu.

[73] N. Laranjeiro, S. N. Soydemir, and J. Bernardino, "A survey on data quality: Classifying poor data," Institute of Electrical and Electronics Engineers Inc., Jan. 2016, pp. 179–188, ISBN: 9781467393768. DOI: 10.1109/PRDC.2015.41.

[74] I. Taleb, H. T. E. Kassabi, M. A. Serhani, R. Dssouli, and C. Bouhaddioui, "Big data quality: A quality dimensions evaluation," Jul. 2016. DOI: http://dx.doi.org/10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.0122.

[75] J. Maślankowski, *Data quality issues concerning statistical data gathering supported by big data technology*, 2014. DOI: 10.1007/978-3-319-06932-6_10.

[76] D. C. Corrales, J. C. Corrales, and A. Ledezma, "How to address the data quality issues in regression models: A guided process for data cleaning," *Symmetry*, vol. 10, 4 Apr. 2018, ISSN: 20738994. DOI: 10.3390/sym10040099.

[77] P. Vassiliadis, T. Sellis, and A. Simitsis, "Optimizing etl processes in data warehouses," 2005, pp. 564–575, ISBN: 0769522858. DOI: 10.1109/ICDE.2005.103.

[78] S. Shalev-Schwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014, ISBN: 978-1-107-05713-5. [Online]. Available: http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning.

[79] J. Han, M. Kamber, and J. Pei, *Data Mining. Concepts and Techniques*, 3rd ed. Morgan Kaufman, 2011.

[80] Y. Baştanlar and M. Özuysal, "Introduction to machine learning," *Methods in Molecular Biology*, vol. 1107, pp. 105–128, 2014, ISSN: 10643745. DOI: 10.1007/978-1-62703-748-8_7.

[81] D. Kreuzberger, N. Kühl, and S. Hirschl, "Machine learning operations (mlops): Overview, definition, and architecture," May 2022, ISSN: 21693536. DOI: 10.1109/ACCESS.2023.3262138. [Online]. Available: http://arxiv.org/abs/2205.02302.

[82] N. C. de Sá, M. Baratchi, V. Buitenhuis, P. Cornelissen, and P. M. van Bodegom, "Automl for estimating grass height from etm+/oli data from field measurements at a nature reserve," *GIScience and Remote Sensing*, vol. 59, pp. 2164–2183, 1 2022, ISSN: 15481603. DOI: 10.1080/15481603.2022.2152304.

[83] Z. Lv, R. Lou, H. Feng, D. Chen, and H. Lv, "Novel machine learning for big data analytics in intelligent support information management systems," *ACM Transactions on Management Information Systems*, vol. 13, pp. 1–21, 1 Mar. 2022, ISSN: 2158-656X. DOI: 10.1145/3469890.

[84] A. Y. Sun, B. R. Scanlon, H. Save, and A. Rateb, "Reconstruction of grace total water storage through automated machine learning," *Water Resources Research*, vol. 57, 2 Feb. 2021, ISSN: 19447973. DOI: 10.1029/2020WR028666.

[85] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," *Journal of Big Data*, vol. 8, 1 Dec. 2021, ISSN: 21961115. DOI: 10.1186/s40537-021-00516-9.

[86] M. Alabadla *et al.*, "Systematic review of using machine learning in imputing missing values," *IEEE Access*, vol. 10, pp. 44 483–44 502, 2022, ISSN: 21693536. DOI: 10.1109/ACCESS.2022.3160841.

[87] S. Jäger, A. Allhorn, and F. Bießmann, "A benchmark for data imputation methods," *Frontiers in Big Data*, vol. 4, Jul. 2021, ISSN: 2624909X. DOI: 10.3389/fdata.2021.693674.

[88] C. Velasco-Gallego and I. Lazakis, "Real-time data-driven missing data imputation for short-term sensor data of marine systems. a comparative study," *Ocean Engineering*, vol. 218, Dec. 2020, ISSN: 00298018. DOI: 10.1016/j.oceaneng.2020.108261.

[89] M. Hulsebos *et al.*, "Sherlock: A deep learning approach to semantic data type detection," Association for Computing Machinery, Jul. 2019, pp. 1500–1508, ISBN: 9781450362016. DOI: 10.1145/3292500.3330993.

[90] M. Bahri, F. Salutari, A. Putina, M. Sozio, and M. S. AutoML, "Automl: State of the art with a focus on anomaly detection, challenges, and research directions," *International Journal of Data Science and Analytics*, vol. 14, pp. 113–126, Feb. 2022. DOI: 10.1007/s41060-022-00309-0. [Online]. Available: https://doi.org/10.1007/s41060-022-00309-0.

[91] S. Omar, A. Ngadi, and H. H. Jebur, "Machine learning techniques for anomaly detection: An overview," *International Journal of Computer Applications*, vol. 79, pp. 33–41, 2 Oct. 2013. DOI: http://dx.doi.org/10.5120/13715-1478.

[92] H. Bourlard and S. H. Kabil, "Autoencoders reloaded," *Biological Cybernetics*, vol. 116, pp. 389–406, 4 Aug. 2022, ISSN: 14320770. DOI: 10.1007/s00422-022-00937-6.

[93] M. B. Kursa and W. R. Rudnicki, "The all relevant feature selection using random forest," Jun. 2011. [Online]. Available: http://arxiv.org/abs/1106.5112.

[94] J. Wang, H. Zhang, J. Wang, Y. Pu, and N. R. Pal, "Feature selection using a neural network with group lasso regularization and controlled redundancy," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, pp. 1110–1123, 3 Mar. 2021, ISSN: 21622388. DOI: 10.1109/TNNLS.2020.2980383.

[95] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[96] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, p. 175, 3 Aug. 1992, ISSN: 00031305. DOI: 10.2307/2685209.

[97] K.-l. Hsu, H. V. Gupta, and S. Sorooshian, "Artificial neural network modeling of the rainfall-runoff process," *Water Resources Research*, vol. 31, pp. 2517–2530, 10 Oct. 1995, ISSN: 00431397. DOI: 10.1029/95WR01955.

[98] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoders," Mar. 2020. [Online]. Available: http://arxiv.org/abs/2003.05991.

[99] S. Robertson and J. Robertson, *Mastering the Requirements Process*, 3rd ed. Addison-Wesley, 2014, ISBN: 978-0321815743.

[100] I. Castillo, F. Losavio, A. Matteo, and J. Boegh, "Requirements, aspects and software quality: The reasq model," *Journal of Object Technology*, 2010. DOI: http://dx.doi.org/10.5381/jot.2010.9.4.a4. [Online]. Available: http://www.jot.fm..

[101] T. O. Group, *ArchiMate® 3.2 Specification*.

[102] G. van Rossum and P. D. Team, "Python tutorial release 3.8.1 guido van rossum and the python development team," Python Software Foundation, 2020.

[103] P. D. Team, *Pandas-dev/pandas: Pandas*, Feb. 2020. DOI: 10.5281/zenodo.3509134. [Online]. Available: https://doi.org/10.5281/zenodo.3509134.

[104] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[105] C. R. Harris *et al.*, "Array programming with numpy," *Nature*, vol. 585, pp. 357–362, 7825 Sep. 2020. DOI: 10.1038/s41586-020-2649-2. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2.

[106] O. Plathey, *Pyfpdf*, 2023. [Online]. Available: https://pypi.org/project/fpdf2/.

[107] J. Bergstra, J. B. Ca, and Y. B. Ca, "Random search for hyper-parameter optimization yoshua bengio," 2012, pp. 281–305. [Online]. Available: http://scikit-learn.sourceforge.net..

[108] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014. [Online]. Available: http://arxiv.org/abs/1412.6980.

[109] J. Kiggins, *Avocado prices*, 2018. [Online]. Available: https://www.kaggle.com/datasets/neuromusic/avocado-prices.

[110] G. Dutta, *Online food delivery time prediction*, 2022. [Online]. Available: https://www.kaggle.com/datasets/gauravduttakiit/online-food-delivery-time-prediction.

[111] G. Forkuor, O. K. Hounkpatin, G. Welp, and M. Thiel, "High resolution mapping of soil properties using remote sensing variables in south-western burkina faso: A comparison of machine learning and multiple linear regression models," *PLoS ONE*, vol. 12, 1 Jan. 2017, ISSN: 19326203. DOI: 10.1371/journal.pone.0170478.

[112] R. M. Adnan, X. Yuan, O. Kisi, and Y. Yuan, "Streamflow forecasting using artificial neural network and support vector machine models," *American Scientific Research Journal for Engineering*, 2017, ISSN: 2313-4402. [Online]. Available: http://asrjetsjournal.org/.

[113] J. Chen *et al.*, "A comparison of linear regression, regularization, and machine learning algorithms to develop europe-wide spatial models of fine particles and nitrogen dioxide," *Environment International*, vol. 130, Sep. 2019, ISSN: 18736750. DOI: 10.1016/j.envint.2019.104934.

Appendix A

Datasets

A.1 Network Operator Data

Dataset name: Network Operator Data

Organization type: Dutch Energy Network Operator

Public or private dataset: Public

Sources: Data aggregation of electricity connection for consumers

Date range of data: 2022

Amount of data instances/rows: 124,784

Target attribute: SLIMME_METER_PERC: Percentage of homes having a smart electricity meter installed

Attributes:

- NETBEHEERDER: EAN-Code of regional electricity network operator
- NETGEBIED: Indication of the electricity network region
- STRAATNAME: Name of the street on which the electricity connections are
- POSTCODE_VAN: Postal code (begin of area)
- POSTCODE_TOT: Postal code (end of area)
- WOONPLAATS: Name of the municipality
- LANDCODE: Country code (Netherlands)
- PRODUCTSOORT: Name of product (electricity or gas), in this case electricity.
- VERBRUIKSSEGMENT: Segmentation of energy usage
- AANSLUITINGEN_AANTAL: Amount of electricity connections in the specific place
- LEVERINGSRICHTING_PERC: Percentage of net electricity usage
- FYSIEKE_STATUS_PERC: Percentage of connections in use
- SOORT_AANSLUITING_PERC: Percentage of most common type of electricity connection
- SOORT_AANSLUITING: The type of electricity connection (Amount of fuses x Amount of ampère)
- SJA_GEMIDDELD: Average yearly consumption of electricity in kWh
- SJA_LAAG_TARIEF_PERC: Percentage of connections with a double tariff
- SLIMME_METER_PERC: Percentage of smart electricity meters

Notes:

The original dataset contained data of both electricity and gas connections for consumers. For this research, we have singled out the data instances with electricity connections only

A.2 Webshop Analytics Data

Dataset name: Webshop Analytics Data

Organization type: Webshop for travel accessoires

Public or private dataset: Private

Data sources: Google Analytics (GA4) + WooCommerce

Date range of data: 23-3-2021 - 9-4-2023

Amount of data instances/rows: 748

Target attribute: Sales: Amount of unique sales per day

Attributes:

- Date: day of data measurement
- Referral_visitors: Amount of website visitors through referrals (clicking a link on another website)
- Organic_search_visitors: Amount of website visitors through an organic search (For example finding the website through a search engine such as Google)
- Direct_visitors: The amount of website visitors finding the website by directly entering the domain name in the URL bar of their web browser
- Organic_social_visitors: The amount of website visitors finding the website through a social media platform (Such as Instagram or Facebook)
- Total_avg_time_spent: The average amount of time spent in seconds on the website by users across all device types (Desktop + Tablet + Mobile)
- Mobie_avg_time_spent: The average amount of time spent in seconds on the website by users who accessed it through a mobile device in seconds
- Tablet_avg_time_spent: The average amount of time spent in seconds on the website by users who accessed it via a tablet
- Desktop_avg_time_spent: The average amount of time spent in seconds on the website by users who accessed it via a desktop or laptop computer
- Total_visitors: Total amount of website visitors on that given day
- Mobile_visitors: Amount of website visitors using a mobile device
- Tablet_visitors: Amount of website visitors using a tablet
- Desktop_visitors: Amount of website visitors using a desktop or laptop computer
- Sales: Amount of unique sales (Customers making a purchase on the website)

A.3 Dutch Fuel Prices

Dataset name: Dutch Fuel Prices

Organization type: Dutch governmental organization

Public or private dataset: Public

Sources: Government data

Date range of data: 1-1-2006 - 3-4-2023

Amount of data instances/rows: 6,302

Target attribute: Lpg_3: Average daily price for the LPG fuel type

Attributes

- Date: Date of data measurement
- BenzineEuro95_1: Average daily price of Euro 95 in the Netherlands
- Diesel_2: Average daily price of Diesel in the Netherlands
- Lpg_3: Average daily price of LPG in the Netherlands

A.4 Bicycle Store Analytics Data

Dataset name: Bicycle Store Analytics data

Organization type: Chain of Dutch bicycle stores

Public or private dataset: Private

Sources: Google Analytics (GA4)

Date range of data: 8-2-2021 - 9-4-2023

Amount of data instances/rows: 2,373

Target attribute: Test_drive: Percentage of web user requesting a bicycle test drive

Attributes:

- Date: day of data measurement
- Segment: The type of website user in terms of device usage. The categories are as follows:
  - "Alle gebruikers": All types of devices
  - "Mobiel verkeer": Mobile devices
  - "Tablet- en desktopverkeer": Tablet, desktops and laptop computers
- users: Total amount of website visitors
- Avg_session_duration: Average duration in seconds of a website user session
- Sessions_per_user: Amount of website sessions per user on a given day
- Sessions: Amount of sessions on a given day
- Bounce: Bounce percentage, which is the percentage of users leaving the website with no interaction (scrolling, clicking links etc)
- Pageviews: Amount of pages viewed on a given day
- Test_drive: Conversion rate of website users, the percentage of users who have requested a test drive on a given day
- Unique_pageviews: The amount of unique pageviews on a given day

A.5 Avocado Prices

Dataset name: Avocado Prices

Organization type: Avocado Supplier

Public or private dataset: Public

Sources: United States Avocado Sales

Date range of data: 2015 - 2018

Amount of data instances/rows: 18,249

Target attribute: AveragePrice: The average price of the avocados

Attributes:

- Date: Date of data measurement
- AveragePrice: The average price of a single avocado on a given day
- Type: Type of avocado, either conventional or organic
- Year: Year of data measurement
- Region: The region in which the avocados were sold
- 4046: Number of avocados sold with PLU 4046
- 4225: Number of avocados sold with PLU 4225
- 4770: Number of avocados sold with PLU 4770

Notes:

We have retrieved this data from Kaggle, which in turn has used publicly available data from the Hass Avocado Board.

A.6 Delivery Time Data

Dataset name: Delivery Time Data

Organization type: Aggregation of Indian Restaurant Delivery Data

Public or private dataset: Public

Sources: Synthetic

Date range of data: None specified

Amount of data instances/rows: 8,782

Attributes:

- Date: Date of data measurement
- Salesperson: Name of the individual making the sale on a car
- Customer Name: Name of individual purchasing the car
- Car Make: Brand of the car
- Car Model: Model of the car
- Car Year: Year of the production of the car
- Sale Price: The price for which the car was sold
- Commission Rate: Commission rate of the salesperson
- Commission Earned: The amount of commission earned by the salesperson

Appendix B

Literature review

B.1 Big Data Aspects

Source	**Characteristics**
Bulger et al. (2014)	Volume
	Variety
	Velocity
	Veracity
Gartner.	Volume
	Velocity
	Variety
Hilbig et al. (2018)	Volume
	Velocity
	Variety
	Value
	Veracity
de Mauro et al. (2016)	Volume
	Velocity
	Variety
	Technology
	Analytical Methods
	Value
Hannila et al. (2022)	Volume
	Velocity
	Variety
	Veracity
	Variability
	Value
Gandomi & Haider (2015)	Variety

Source	**Characteristics**
	Volume
	Velocity
	Variability
	Complexity
	Value
	Veracity
	Venue
	Vocabulary
	Vagueness
	Exhaustive
	Fine-grained
	Validity
	Visualization
	Vulnerability
	Volatility
Al-Mekhal & Ali Khwaja (2019)	Volume
	Variety
	Velocity
	Veracity
	Variability
	Value
Cartledge (2016)	Variety
	Velocity
	Volume
	Validity
	Value
	Variability
	Veracity
	Viability
	Virility
	Viscosity
	Visibility
	Visualization
	Volatility
	Vagueness
	Venue
	Vocabulary
	Vincularity
	Visible
	Vitality

Table B.1: List of Big Data aspects found in literature

B.2 Dimensions of Data Quality

Source	Dimensions
Fan (2015)	Consistency
	Duplication
	Completeness
	Currency
	Accuracy
Hazen et al. (2014)	Accuracy
	Currency
	Consistency
	Completeness
Abdullah et al. (2015)	Accuracy
	Integrity
	Consistency
	Completeness
	Validity
	Currency
	Accessibility
Fox et al. (1994)	Accuracy
	Currency
	Completeness
	Consistency
Sidi et al.(2012)	Currency
	Consistency
	Accuracy
	Completeness
	Accessibility
	Duplication
	Data specification
	Presentation quality
	Reputation
	Safety
	Security
	Believability
	Understandability
	Objectivity
	Relevancy
	Effectiveness
	Interpretability
	Ease of manipulation
	Free-of-Error

Source	**Dimensions**
	Maintainability
	Useability
	Reliability
	Amount of data
	Freshness
	Value added
	Learnability
	Data decay
	Concise
	Consistency
	Integrity
	Navigation
	Usefulness
	Efficiency
	Availability
	Data coverage
	Transactability
Wang, (1998)	Accuracy
	Objectivity
	Believability
	Reputation
	Accessibility
	Security
	Relevancy
	Value added
	Currency
	Completeness
	Amount of data
	Interpretability
	Understandability
	Concise
	Consistency
Tee et al. (2007)	Accuracy
	Reliability
	Relevancy
	Consistency
	Precision
	Currency
	Understandability
	Concise
	Usefulness

Source	**Dimensions**
Eckerson (2002)	Accuracy
	Integrity
	Consistency
	Completeness
	Validity
	Currency
	Accessibility
Gudivada et al.(2017)	Data Governance
	Data Specification
	Integrity
	Consistency
	Currency
	Duplication
	Completeness
	Data provenance
	Data heterogenity
	Streaming data
	Outliers
	Dimensionality reduction
	Feature selection
	Feature extraction
	Business rules
	Accuracy
	Gender bias
	Security
	Availability
Pipino et al. (2002)	Accessibility
	Amount of data
	Believability
	Completeness
	Concise
	Ease of manipulation
	Free-of-Error
	Interpretability
	Objectivity
	Relevancy
	Reputation
	Security
	Currency
	Understandability
	Value added

Source	**Dimensions**
Scannapieco et al. (2015)	Accuracy
	Completeness
	Currency
	Volatility
	Consistency
Cai & Zhu (2015)	Availability
	Usability
	Reliability
	Relevancy
	Presentation quality

Table B.2: Dimensions describing the concept of data quality

B.3 Data Quality Issues

Source	Context DQ issues	Issues
Souibgui et al. (2019)	ETL Process	Uniqueness violation
		Poor schema design
		Embedded values
		Duplicate data
		Missing data
		Wrong data type
		Naming conflicts
		Syntax violation
		Mapping of data
Berti-Équille (2007)	ETL Process	Wrong data type
		Data formats
		Duplicate data
		Approximations
		Measurement errors
		Hardware/software constraints
		Human errors
		Computational constraints
Kellen (2003)	General	Lack of validation routines in data entry systems
		Syntax violation
		Data formats
		Code structures
		Data conversion errors
		Changes in systems
		Complexity of system integrations
		Poor system design
Oliveira et al. (2005)	General	Missing data
		Syntax violation
		Incorrect data
		Domain violation
		Violation of business rules
		Invalid substring
		Misspelling
		Imprecise value
		Uniqueness violation
		Synonyms
		Semi-empty tuple
		Violation of functional dependency
		Approximate duplicate tuples
		Referential integrity violation
		Incorrect reference
		Syntax violation
		Circularity among tuples in a self-relationship
		Data heterogenity
		Homonyms
Sidi et al. (2012)	General	Poor schema design
		Uniqueness violation
		Referential integrity violation
		Data entry errors

Source	Context DQ issues	Issues
		Misspelling
		Irrelevant data
		Contradictory values
		Heterogeneous data models + schema design
		Naming conflicts
		Inconsistent data
Laranjeiro et al. (2016)	General	Missing data
		Incorrect data
		Misspelling
		Ambiguous data
		Extraneous data
		Outdated data
		Misfielded values
		Incorrect reference
		Duplicate data
		Domain violation
		Violation of functional dependency
		Wrong data type
		Referential integrity violation
		Uniqueness violation
		Structural conflicts
		Different word orderings
		Different aggregation levels
		Temporal mismatch
		Different units
		Different representations
		Synonyms
		Homonyms
		Use of special characters
		Data formats
Taleb et al. (2016)	General	Missing data
		Incorrect data
		Data entry errors
		Irrelevant data
		Outdated data
		Misfielded values
		Uniqueness violation
		Functional dependency violation
		Wrong data type
		Poor schema design
		Lack of integrity constraints
Maślankowski (2014)	ETL Process	Wrong data
		Noisy data
		Irrelevant data
		Incomplete data
		Redundant data
Corrales et al. (2018)	Machine Learning	Missing data
		Irrelevant data
		High dimensionality
		Duplicate data
Kim et al. (2003)	General	Missing data

Source	Context DQ issues	Issues
		Wrong data type
		Dangling data
		Duplicate data
		Lost update
		Dirty read
		Unrepeatable read
		Lost transaction
		Wrong categorical data
		Outdated data
		Data entry errors
		Misspelling
		Extraneous data
		Misfielded values
		Wrong derived data
		Inconsistent data
		Abbreviation error
		Incomplete data
		Alias
		Encoding format error
		Different representations
		Different units
		Use of special characters
		Different orderings
Eckerson (2002)	General	Data entry errors
		Violation of business rules
		Duplicate data
		Missing data
		Incorrect data
		Syntax violation
		Changes in systems
		Spiderweb of interfaces
		Lack of referential integrity checks
		Poor system design
		Data conversion errors
Chu et al. (2016)	General	Missing data
		Misspelling
		Data formats
		Replicated entries
		Violation of business rules
Gudivada et al. (2017)	Machine Learning	Missing data
		Duplicate data
		Data heterogenity
		Irrelevant data
		Inconsistent data
		Incomplete data

Table B.3: Collection of data quality issues found per reference per context

B.4 ML Algorithms used to tackle Data Quality issues

Reference	DQ Issue	Algorithm proposed
Sun et al. (2021)	Missing data	AutoML
Emmanuel et al. (2021)	Missing data	MissForest (Based on RF)
		KNN
Alabadla et al. (2022)	Missing data	xGBoost
		Neural Network
		KNN
Jäger et al. (2021)	Missing data	Random forest
		KNN
Velasco-Gallego & Lazakis (2020)	Missing data	Vector AutoRegression
Hulsebos et al. (2019)	Wrong data type	Neural Network
Bahri et al. (2022)	Outliers	AutoML
Omar et al. (2013)	Outliers/anomalies	KNN
		Bayesian Network
		Supervised Neural Network
		Decision Tree
		Support Vector Machine
		Clustering Techniques
		Unsupervised Neural Network
		k-Means
		Fuzzy C-Means
		Unsupervised Niche Clustering
		Expectation-Maximization Meta Algorithm
		One-Class Support Vector Machine
Bourlard & Kabil (2022)	Irrelevant data	AutoEncoder
Kursa & Rudnicki (2011)	Irrelevant data	Random Forest
Wang et al. (2021)	Irrelevant data	Neural Network

Table B.4: Non-exhaustive collection of ML algorithms used to deal with certain data quality issues

Appendix C

Design

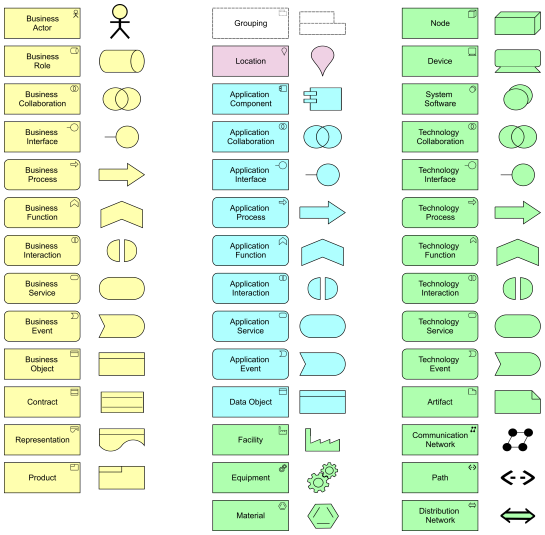


Figure C.1: Core elements in the ArchiMate notation[101]

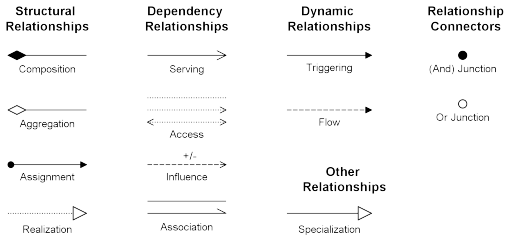
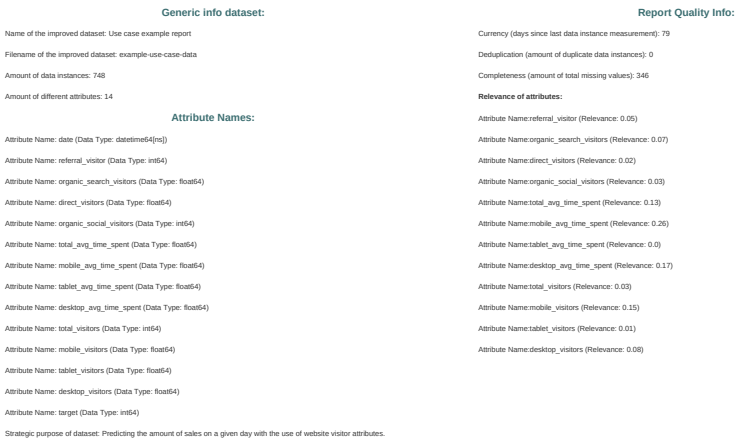


Figure C.2: Relationships in the ArchiMate notation[101]

Appendix D

Prototype



Generic info dataset:

Name of the improved dataset: Use case example report

Filename of the improved dataset: example-use-case-data

Amount of data instances: 748

Amount of different attributes: 14

Attribute Names:

Attribute Name: date (Data Type: datetime64[ns])

Attribute Name: referral_visitor (Data Type: int64)

Attribute Name: organic_search_visitors (Data Type: float64)

Attribute Name: direct_visitors (Data Type: float64)

Attribute Name: organic_social_visitors (Data Type: int64)

Attribute Name: total_avg_time_spent (Data Type: float64)

Attribute Name: mobile_avg_time_spent (Data Type: float64)

Attribute Name: tablet_avg_time_spent (Data Type: float64)

Attribute Name: desktop_avg_time_spent (Data Type: float64)

Attribute Name: total_visitors (Data Type: int64)

Attribute Name: mobile_visitors (Data Type: float64)

Attribute Name: tablet_visitors (Data Type: float64)

Attribute Name: desktop_visitors (Data Type: float64)

Attribute Name: target (Data Type: int64)

Strategic purpose of dataset: Predicting the amount of sales on a given day with the use of website visitor attributes.

Report Quality Info:

Currency (days since last data instance measurement): 79

Deduplication (amount of duplicate data instances): 0

Completeness (amount of total missing values): 346

Relevance of attributes:

Attribute Name:referral_visitor (Relevance: 0.05)

Attribute Name:organic_search_visitors (Relevance: 0.07)

Attribute Name:direct_visitors (Relevance: 0.02)

Attribute Name:organic_social_visitors (Relevance: 0.03)

Attribute Name:total_avg_time_spent (Relevance: 0.13)

Attribute Name:mobile_avg_time_spent (Relevance: 0.26)

Attribute Name:tablet_avg_time_spent (Relevance: 0.0)

Attribute Name:desktop_avg_time_spent (Relevance: 0.17)

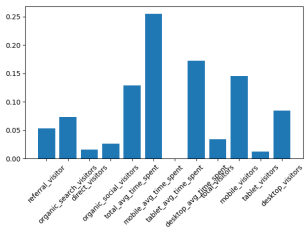
Attribute Name:total_visitors (Relevance: 0.03)

Attribute Name:mobile_visitors (Relevance: 0.15)

Attribute Name:tablet_visitors (Relevance: 0.01)

Attribute Name:desktop_visitors (Relevance: 0.08)

Figure D.1: Report produced by the HAQIM Prototype (1)



Performance comparison:

We have measured the R2 score of your dataset before and after enhancing its dataquality.

This score corresponds to how strongly a model based on this data can explain your target variable.

Predictive power before: 0.10295149824444938

Predictive power after: 0.1587682365514127

This is an increase of 54.22%

Cleaning your dataset from data quality problems will lead to better decision making.

Quality label of input data:

Total Data Quality Grade: 3.0 / 5

Completeness Grade: 53.7%

Deduplication Grade: 100.0%

Timeliness Grade: 35.0%

Relevancy Grade: 11.1%

Quality label of enhanced data:

Total Data Quality Grade: 3.46 / 5

Completeness Grade: 100.0%

Deduplication Grade: 100.0%

Timeliness Grade: 35.0%

Relevancy Grade: 11.1%

Technical information:

Class Name: organic_search_visitors

Algorithm Used: Random Forrest

Model File Name: [redacted]

Amount Missing: 66

MAE: 0.49752807045791664

RMSE: 0.8225175868796303

Accuracy: False

Class Name: direct_visitors

Algorithm Used: Random Forrest

Model File Name: [redacted]

Amount Missing: 66

MAE: 0.467374453911220 2

RMSE: 0.8032475644873267

Accuracy: False

Class Name: total_avg_time_spent

Algorithm Used: Random Forrest

Model File Name: [redacted]

Amount Missing: 54

MAE: 10.189626486035953

RMSE: 28.03549535717949

Accuracy: False

Class Name: mobile_visitors

Algorithm Used: Random Forrest

Model File Name: [redacted]

Amount Missing: 66

MAE: 0.37858786026487895

RMSE: 0.7342653065184302

Accuracy: False

Class Name: tablet_visitors

Algorithm Used: Random Forrest

Model File Name: [redacted]

Amount Missing: 41

MAE: 0.02546296296306

RMSE: 0.1001840709522344 7

Accuracy: False

Class Name: desktop_visitors

Algorithm Used: Random Forrest

Model File Name: [redacted]

Amount Missing: 53

MAE: 0.4722193395316 54

RMSE: 0.9304589424041237

Accuracy: False

Figure D.2: Report produced by the HAQIM Prototype (2)

Appendix E

Results

	Original data	Manually enhanced	HAQIM enhanced
MAE	6.11	6.177642	5.749942
mse	67.655865	69.106323	60.780740
rmse	8.225319	8.313021	7.796200
r2	0.471160	0.475918	0.539373

Table E.1: Performance scores when using the Network Operator Data

	Original data	Manually enhanced	HAQIM enhanced
MAE	0.229770	0.171950	0.172833
mse	0.224374	0.115367	0.117963
rmse	0.473681	0.339657	0.343457
r2	-1.657120	0.080246	0.059550

Table E.2: Performance scores when using the Webshop Analytics Data

	Original data	Manually enhanced	HAQIM enhanced
MAE	0.007954	0.008384	0.008512
mse	0.000128	0.000144	0.000145
rmse	0.011313	0.012020	0.012029
r2	0.201220	0.198963	0.197745

Table E.3: Performance scores when using the Bicycle Store Data

	Original data	Manually enhanced	HAQIM enhanced
MAE	0.033656	0.035230	0.034390
mse	0.001889	0.002000	0.001933
rmse	0.043465	0.044717	0.043963
r2	0.858265	0.866264	0.870738

Table E.4: Performance scores when using the Dutch Fuel Prices Data

	Original data	Manually enhanced	HAQIM enhanced
MAE	0.318074	0.307355	0.238338
mse	0.155107	0.148273	0.095496
rmse	0.393837	0.385062	0.309025
r2	0.057787	0.080980	0.408100

Table E.5: Performance scores when using the Avocado Prices Data

	Original data	Manually enhanced	HAQIM enhanced
MAE	9.277533	9.297251	8.904536
mse	152.189722	152.212941	129.547557
rmse	12.336520	12.337461	11.381896
r2	0.092684	0.092545	0.110103

Table E.6: Performance scores when using the Delivery Time Data