Data Mining in Norwegian Level-of-Living Survey Data

Author: Magnus Lyngseth Vestby

Supervisor: Ankica Babic

Department of Information Science and Media Studies

University of Bergen

2020

Acknowledgements

First and foremost I would like to thank my supervisor, Professsor Ankica Babic. She has been a great support throughout the process from start till finish. She handed me "Factfulness" in the early days of 2019, the book which later became the source of inspiration for this thesis.

I would also like to thank special advisor Rolf E.S. Halse at the Norwegian Centre for Research Data for his advice and counselling early in the research process. His input led me to the level-of-living studies and its very useful survey data.

Another source of support has been my supervisor group with my lovely fellow students who all struggled from their home offices during the pandemic, but helped each other nonetheless.

Also I would like to thank the support from the lively and fun reading room 635, which in the final weeks before deadline was constant stream of positive energy and motivation.

Magnus Lyngseth Vestby
Bergen, 21.07.2020

"Educate and inform the whole mass of the people(...). They are the only sure reliance for the preservation of our liberty."

Thomas Jefferson, 31 December 1787.[14]

Abstract

The thesis analyses how level-of-living survey data can be explored using data mining techniques and how well the resulting patterns can be visualized to inform non-experts.

The project utilized the design science research framework for the project structure and methodology, and the knowledge discovery in databases (KDD) methodology for developing the models and visualizations.

To answer the research questions several machine learning methods were tested on a data set with selected variables describing education, disability, health, age, and marital status over a period of 50 years (1973-2017). Scikit-learn was used to employ the machine learning models. Ridge regression was found to be optimal model for the goals of this thesis based on the patterns produced and the accuracy of the predictions. The patterns found by the Ridge regression were visualized in graphs and bar charts. The visualizations were then evaluated using semi-structured interviews, tasks, and a visualizations usability scale.

The results show that visualizations based on the patterns found during data mining of the level-of-living surveys, were informative and interesting to the participants in the evaluation. The visualizations scored highly on the visualizations usability scale, with an average score of 87.5. This meant that the group had little to no problems interpreting the graphs and figures. The participants were surprised by some of discovered patterns regarding inequalities related to gender and level of education. It shows that interesting patterns in the Norwegian level-of-living surveys can be found with the use data mining techniques. It also shows that these patterns can be visualized so that non-experts can retrieve information.

This thesis represents a proof by construction. It shows that patterns in the Norwegian level-of-living surveys can be found with the use of data mining techniques. The model developed here can be reused for similar projects and data mining tasks, but future developers need to pay attention to all steps of the KDD-process including the data cleaning. Furthermore, a user interface should be designed to enable a different kind of user groups to discover their patterns of interest in the level of living data. For future work, a user interface should be designed to enable user groups to find their own patterns of interest in the level-of-living data.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

After the second world war, politicians wanted to know more about the state of the public, especially for the disadvantaged. One of the answers were the level-of-living studies, created by a Swedish team of Statisticians in 1968 [15]. It's purpose was to create a collection of expertly curated data that could describe the status of the nation's citizens. Starting initially with a focus on the vulnerable groups of society, it has in later years grown into a tool for insight into the prosperity and living conditions of the general public. The data is a collection of variables divided into important indicators of level-of-living such as health and access to care, employment and working conditions; economic resources, educational resources; family and social integration; housing and neighbourhood facilities, security of life and property, recreation and culture and political life. This data has been collected for 50 years with varying main themes. This work has given statistician, sociologist and social anthropologists basis for reports, articles and other publications that can showcase different opportunities, problems or vulnerabilities that exist in the Norwegian society. The idea of using statistics to enlighten the public and improve the decision making of stakeholders has been present in Norway since the "Statistiske Centralbureau"(Central Bureau for Statistics) was established with 15 employees in 1876 [36].

"Factfulness" by Hans Rosling et al, is one of the best examples of using statistical data for informative visualizations. In this book, the authors present how data can be put into great use to explain global trends of living standard, health and general progress. Often, we seem to be wrong about the facts. We see things more negatively than the data is, especially if we observe the progress over time. The authors have developed clear and engaging ways of presenting visualizations which inform the public. For academics, a book like Factfulness can trigger self-reflexivity in our research and publishing practices [18]. The book relies on the visualisations that help bring the facts to life.

The motivation behind this research is to explore the possibilities for knowledge extraction from level-of-living surveys data by automatic methods to decrease the need for complex manual expert processing. It also is encompassed in the Thomas Jefferson quote referenced at the start of this thesis and in the words of Hans Rosling underneath.

> *This book is my very last battle in my lifelong mission to fight devastating global ignorance (...) and redirect their energies into constructive activities"[33].*

The idea of making complex information available to the public in a easy to learn and easy to read fashion as one of cornerstones of any democracy. A democracy is dependent on an informed public which can make decisions based on clear, unbiased, and correct information.

Information that is available to all, and easy to interpret. Only by these measures can we combat misunderstandings and a misallocation of resources.

1.1 Research Questions

RQ1: Can machine learning regression models be used to data mine Norwegian level-of-living survey data?

RQ2: Can visualizations present the results from *RQ1* and make the information interesting for non-experts?

1.2 Thesis outline

Here follows the outline of the research project:

Chapter 2: Theory and Data set is a short introduction to machine learning and the data used in this project.

Chapter 3: Literature Review is a summary of the literature and related works on this project.

Chapter 4: Methodology and methods describes the methodologies used during the research and the applied methods of the research.

Chapter 5: Development is a summary of the development iterations and the requirements to the artefact produced.

Chapter 6: Results displays the main results of the research.

Chapter 7: Evaluation summarizes the evaluation done after the final iteration.

Chapter 8: Discussion presents and discusses the main methodologies, methods and development process used in the research. This chapter also answers the research questions.

Chapter 9: Conclusion and future work is the final chapter of this thesis and present a summary and recommended future work.

1.3 Norwegian Centre for Research Data

This thesis has been approved and supported by the Norwegian Centre for Research Data (NSD). NSD is a national archive which ensures open and easy access to research data. NSD sees research data as a public good which should be available for researchers to improve empirical research without economic, jurisdictional, or practical barriers. It also provides several information and support services for both national and international researchers. NSD enables researchers to save both time and increase their capacity for active research instead of data collection. NSD has been instrumental in providing the data set and helping with any privacy or research-related questions.

Chapter 2

Theory and Data set

2.1 Machine Learning

Machine learning is described by A. Muller and S. Guido as the process of extracting knowledge from data [24]. It's an intersection of statistics, artificial intelligence and computer science. The general definition of machine learning, attributed to Arthur Samuel is that,

> [Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed [34]. (Arthur Samuel, 1959)

A more technical definition came from T. Mitchell in 1997 [21]. It describes what the learning entails in the machine learning perspective.

> A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. (Tony Mitchell, 1997)

Together these definitions explain the core of machine learning. To make models that make use of historical data to perform a certain task with an explicitly stated performance measure so that the machine can, with good and usable data, improve with experience and increase its performance without explicitly being programmed. Good machine learning models give the programmer the ability to solve a multitude of problems with differentiating data without having to rewrite the program for every possible case.

2.2 The data set

The data spans a period of 50 years from 1968 to 2018. The first health survey conducted by Statistics Norway in 1968 was the starting point of the nationwide health surveys and was repeated in 1975, 1985 and 1995. The purpose of the health surveys was to collect information on the prevalence of diseases and injury and assess the impact this has the individuals use of health care services and physical activity. From 1973 an additional survey on level-of-living was initiated by the Norwegian government [31]. The survey as part of a science and research project named Level-of-Living study (Levekårsundersøkelsen). This goal of this study was to find out the levels of living in the population with special emphasis on lower income groups and other groups which were assumed to be living under special or problematic conditions. This survey had an interval of every 3-4 years. A third related study was initiated called the Norwegian survey of Housing. The purpose of this survey was to provide a broad overview of living conditions in Norway based on the size and standard of housing compared to household size, composition and attitudes. The survey of housing was conducted in 1967(not present in this data set), 1973, 1981, 1988 and 1995. All three surveys was conducted by Statistics Norway. After 1995 the specialized surveys were coordinated in a single, annual survey on living conditions with rotating themes. Finally in 2011 the EU-SILC (European Union-Survey on income and living conditions) format was adopted by Statistics Norway to streamline and conform the data to a European standard. This work is coordinated by Eurostat (European Statistical Office), which is a Directorate-General of the EU. The EU-SILC covers a wide variety of levels of living-variables with additional themes for each year on a three year cycle. As a whole the EU-SILC covers themes such as economy, housing conditions, recreational activities, social networks, health, education, working conditions and level of worry for crime. The health survey of 1968 differentiates from the rest and is difficult to compare directly. It has therefore not been included in this research. The survey was structured with household as a sampling unit and were intended to include all permanent residents in the household. Adults over 18 where interviewed as a separate interview device, while interview objects(IOs) where interviewed as one interview unit with spouse and children under 18.

Chapter 3

Literature Review

In this chapter I will go through literature which showcase methods for informing the public on important issues. The chapter also includes a paper on the value of visualization.

3.1 Knowledge through visualizations

Paper I: Factfulness: Ten reasons we're wrong about the world - And why things are better than you think [32]

Rosling, H. and Rönnlund, A.R. and Rosling, O. (2018), book, Sceptre

The book Factfulness, released in 2018 by the Roslings, gives an interesting argument for the misinformation present in most western countries. Rosling discovered that most people are wrong about the general status of the world. This was present in all kinds of subjects, be it poverty, education or health. To debunk these myths, Rosling applies a number of techniques. One of the most prominent tools are the extensive use of visualizations. Information is contextualised in novel and interesting ways which gives the reader an immediate insight. With every point that Rosling makes he backs it up with a graph, figure or other types of visuals. Rosling also writes of the importance of looking at long-term trends and not focusing on the more dramatic short-term increases or decreases. A graph with a limited time span can be a poor representation of real life developments, but often is more dramatic and interesting. Rosling argues that especially the media are more interested in short-term negative developments that the longer overview. Rosling strives for a truer representation of the state of the world. The book gives ten reasons to why the public has a skewed world view, many of which are misconception created both by human tendencies, such as our tendencies to look at things binary, i.e. there is a poor-rich divide, good and bad, while in reality there are a lot of nuances. But also there is a lack in general knowledge which has not been countered by neither education or media. And throughout this book it is clear that much of that information is easy-to-read and easy-to-understand, by just using the right visualizations. Rosling also explains the problems of having a mismatch between the publics perception and the true state of the world. A skewed world view restricts decision makers and the public in their discourse. For example, the politics of humanitarian aid and philanthropy is hindered by a lack of knowledge, which might negatively affect the policies and create a misallocation of resources. To combat this misinformation, Rosling uses a lot of new and interesting visualizations techniques. These new visualizations portray information in new and exciting ways, which makes the information easier to understand and contextualizes the different states of the world. With this Rosling is able to make the reader knowledgeable about the world in a novel way.

Paper II: **Social inequalities in health [28]**

Norwegian Institute for Public Health

In 2018 the NIPH released their newest report on the social inequalities in health. This is an expert curated report which gives important and valuable insight which informs both the public and decision makers. The report uses research from almost 50 different sources into one singular report about the state of health in Norway. The report uses key points and visualizations to summarize the results. This report is important for politicians for multiple reasons. The give import insight into what is the current negative and positive trends. This gives decision makers actionability to employ policies that are effective and target the right areas of society. Since politicians are not necessarily experts on the domain they work in, they are dependent on reports that provide knowledge that is easy to read and easy to understand. The reports are the result of research over a long period of time and are backed by many hours of manual and expert labour. An example of findings from the report is that citizens with high education and high income have a higher life expectancy than citizens with low income and low education. And these differences are increasing, especially among women, the differences are observed on the country, county and municipal level. In effect this gives women and men with the highest education and income on average 5-6 years longer to live and with better health than those with lowest education and income. This information is visualized through graphs which shows the different trends for the different groups.

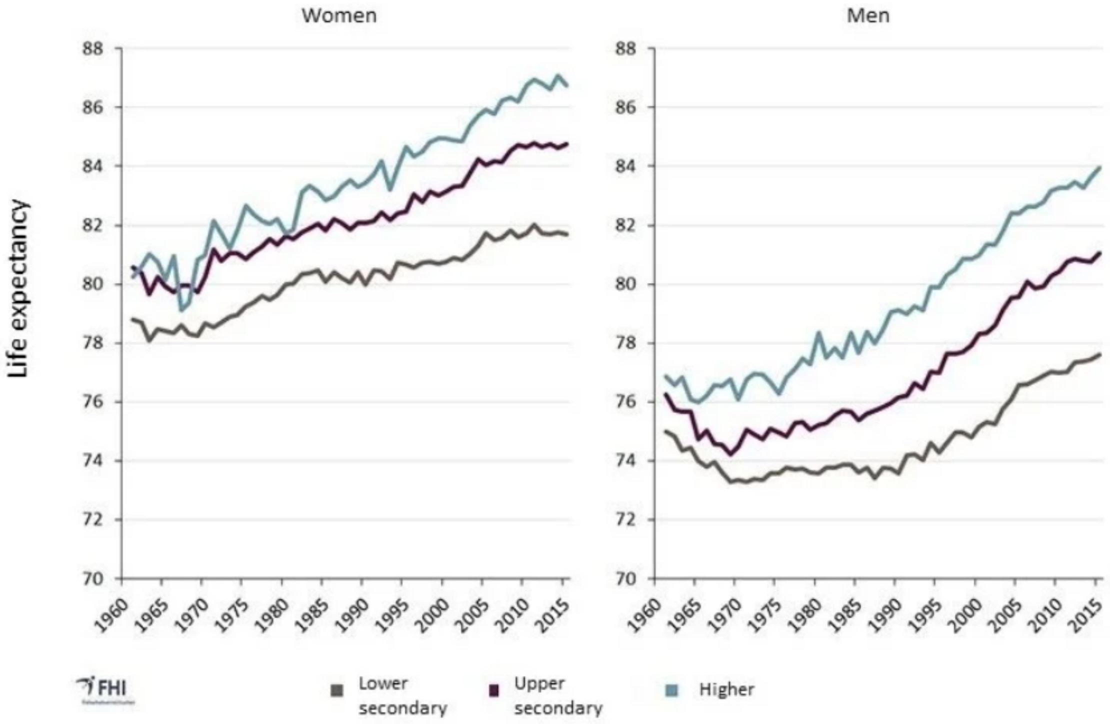


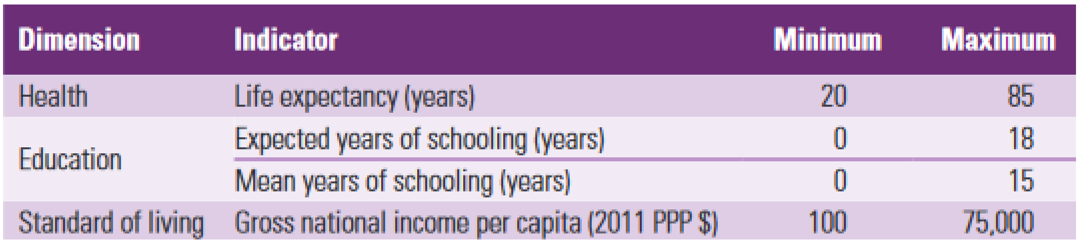
Figure 1: Life expectancy for women and men aged 35 in Norway, 19612015, grouped by education level [37]

.

Paper III: **Human development index**

United Nations

Another way of conveying information of development is the Human Development Index. HDI is a United Nations initiative that uses a statistical index to describe the general well-being of a population. It uses three measurements, life expectancy, education and income in gross national income(GNI) per capita. The index goes from 0(worst) to 1(best). In the last recorded year, 2019, Norway had the highest score, with 0.954 and Niger at the bottom with a score of 0.377. This provides an easy, but powerful tool to compare countries without having to understand the science and immense statistical work behind the numbers. There are three main goals of the HDI. It is to measure people, opportunities and choice. To measure people is about focusing on improving the lives of people, instead of the general economy. An assumption often made about GNI growth is that it benefits the population at large, but this is not always a true assumption. GNI per capita is very crude measurement which can miss the inequalities which might hinder human development. With literacy and life expectancy in addition it is a greater measurement of development than GNI per capita alone. It also indicates that GNI per capita growth is as a tool to increase development, but it is not a goal in itself. As the United Nations Development Programme states in their Human Development Reports there are three foundations for human development: to live a long, healthy and creative life, to be knowledgeable and have possibilities to use them. Choice in this context is also is also an important part of human development. Choice is about providing people with alternatives. While human happiness cannot be guaranteed, society should give people the option to make the right or wrong choice. It is an important part, both personally and collectively, of fulfilling ones potential both creatively and productively. In order to calculate HDI there are set three dimensions with related indicators.



Dimension	Indicator	Minimum	Maximum
Health	Life expectancy (years)	20	85
Education	Expected years of schooling (years)	0	18
	Mean years of schooling (years)	0	15
Standard of living	Gross national income per capita (2011 PPP $)	100	75,000

Table 1: The three main dimension with related indicators for 2017.

1 shows the range of indices set for 2017. To achieve a score of 1 a nation must perform as well or better than all the maximum limits. Under the minimum gives a score of 0. Each dimension is calculated using the simple equation:

$$Dimension\,index = \frac{actual\,value - minimum\,value}{maximum\,value - minimum\,value} \quad (1)$$

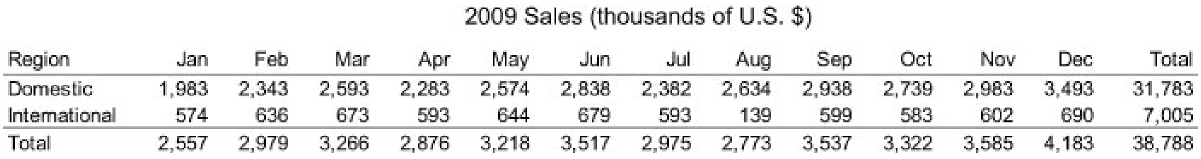
Equation 1 shows the calculation of a dimension index. For the education score the equation is done for the two separate indices and then the mean of the two is calculated and used as the final index score. The three final scores are then multiplied together and then squared by a factor of three.

In 2010 a new dimension was proposed to the HDI. The Inequality-adjusted HDI(IHDI) considers the distribution of growth in each of the three indices (health, income, education). In a country with no inequality, the HDI and IHDI are the same. The IHDI is a representation of the loss of development that inequality can lead to. The results of this are for example that the USA lose 12 % of its score when adjusted for inequalities and drops 13 places down in rank. With just a single number, the HDI-report manages to give an powerful tool of comparison.

Paper IV: Data Visualization for human perception

Stephen Few

In the encyclopedia for human interaction, Stephen Few[8] writes about the necessity of visualization in order to analyze data and communicate. Information is abstract and visualization is to give form to information. This means that it is crucial in order to detect trends and communicate the results in a effective, meaningful way. Stephen Few gives to example to illustrate quite clearly the point of visualization.



2009 Sales (thousands of U.S. $)

Region	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
Domestic	1,983	2,343	2,593	2,283	2,574	2,838	2,382	2,634	2,938	2,739	2,983	3,493	31,783
International	574	636	673	593	644	679	593	139	599	583	602	690	7,005
Total	2,557	2,979	3,266	2,876	3,218	3,517	2,975	2,773	3,537	3,322	3,585	4,183	38,788

Figure 2: Example of sales table

This table is very informative and easy when, for example, looking for the specific sales for a specific time. But it is hard to intuitively explore the data any further. Therefore Few presents the data in a graph.

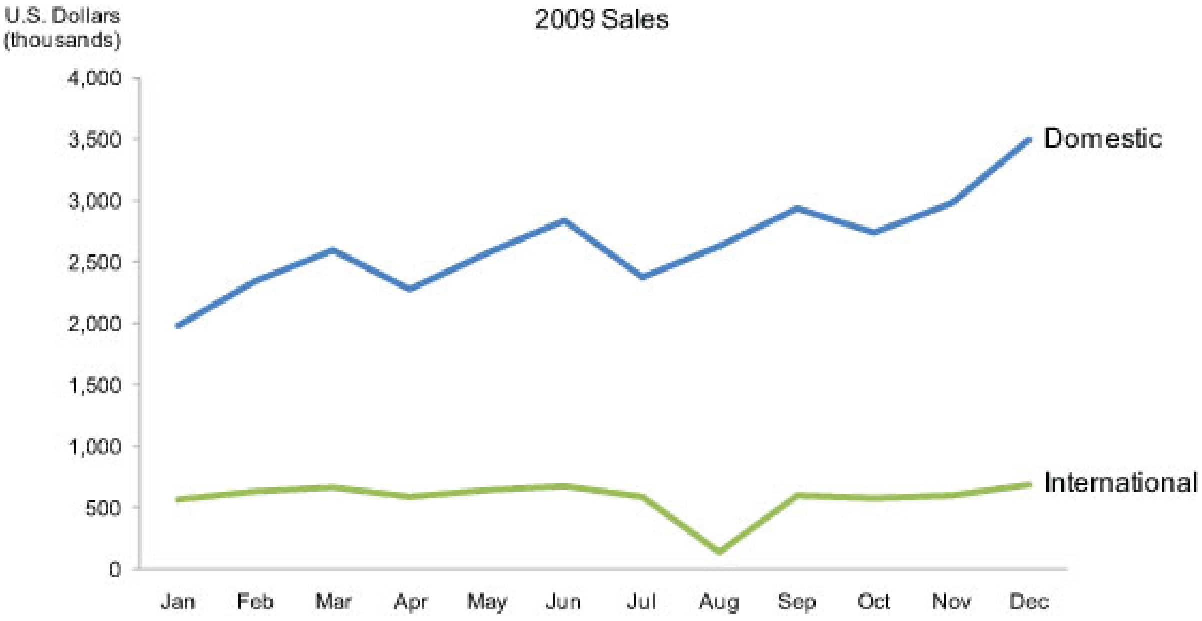


Figure 3: Visualization of the sales table

The graphs communicate to the reader very differently from the tables. It enables the reader to intuitively understand the data, with the results of knowledge discovery and insights. Insights such as the trending upwards domestic sales, the relative flat trends in international sales and a easy way to see the difference compared to each other in scale, physically showing the difference in sales domestically and internationally. Stephen Few also discusses the aim of a visualization. If the visualization cannot be deciphered by the brain and the eyes it is not working optimally. Stephen Few has written a list of criteria that an informative visualizations should incorporate:

- The relation between the different values should be clear, either as a comparison or as a whole.
- Represent any value accurately.
- Make it easy to compare values.
- Make it easy to see the ranked order of values, for example be quickly able to see which variable is the biggest factor for income.
- Make it intuitively what the visualization is trying to convey. The usefulness of the visualizations should be clear.

3.2 Related Works

Microdata.no[9] is a resource provided by NSD and Statistics Norway. Microdata is a project which aims at giving researchers an infrastructure for easy access to high quality statistical data. It also manages statistical confidentiality and protecting the data and privacy of research subjects. The information system is web-based and has the ability to extract data from a number of different databases by using query language. Microdata also offer a handful of visualization techniques to produce visualizations. The program has the function to create histograms, box plots, bar charts, pie charts, hexbin plots and several machine learning methods. The available data sets does not include level-of-living surveys. It incorporates tools for data exploration and analysis. It is a great starting point for a researcher interested in the variables available. It does however yet have all the full functionality of libraries found for machine learning on python, such as scikit-learn, described in section 5.3.1. The framework is a specialized framework for researchers and does not offer an interface for a general public. The information system demands an amount of familiarity or time invested in order to produce adequate results. As of May 2020 only verified researchers have access to these tools.

Chapter 4

Methodology

4.1 Design Science Research

According to Hevner et al, there are two paradigms that characterize the research in information science [11]. The behavioural science and design science. The behavioural science is rooted in natural science research. To goal of behavioural research is to explore theories that explain or predict human behaviour on the human and organizational level when designing, using, implementing and managing information systems. The resulting theories give practitioners and researchers the knowledge of how humans, technology and organizations interact and how to optimize and improve this interaction. The results are used improve the effectiveness and efficiency of information systems.

The other paradigm of information science research is rooted in engineering and in the science of the artificial, namely the design science. Its mainly concerned about problem-solving and is a study of the innovative. It incorporates the technology, ideas and products that improves the analysis, design, management, implementation and use of information systems. The distinction between the paradigms does not mean that innovations in design science is not dependent of the behavioural and natural laws of the domain. The fact is that the creation of new ideas and products depend on the researchers existing knowledge of behaviour. The knowledge is needed in order to apply, test and modify the novel innovations in way that complements and improves the information systems and its interactions with human behaviour.

Hevner et al. propose the design science research method to structure and create a framework for researchers to better understand, evaluate and measure the quality of design science research.

The underlying foundations for design science research are at any time open for technological revolutions. The definition and use-cases of an information system can at any time change. One example is the introduction of the world wide web which changed the way researchers looked at the design, use and implementation of an information system. It is therefore important to have a framework which is adaptive and more concerned with the process than any type of artefact.

This list of seven guidelines proposed by Hevner is the result of this work and it enables researchers and practitioners to understand the requirements of good design science research.

Guideline	Description
Guideline 1: Design as an Artefact	Design-science research must produce a viable artefact in form of a construct, a model, a method, or an instantiation.
Guideline 2: Problem Relevance	The objective of design-science research is to develop technology-based solutions to important and relevant business problems.
Guideline 3: Design Evaluation	The utility, quality, and efficacy of a design artefact must be rigorously demonstrated via well-executed evaluation methods
Guideline 4: Research Contributions	Effective design-science research must provide clear and verifiable contributions in the areas of the design artefact, design foundations, and/or design methodologies.
Guideline 5: Research Rigor	Design-science research relies upon the application of rigours methods in both the construction and evaluation of the design artefact.
Guideline 6: Design as a Search Process	The search for an effective artefact requires utilizing available means to reach desired ends while satisfying laws in the problem environment.
Guideline 7: Communication of Research	Design-science research must be presented effectively both to technology-oriented as well as management-oriented audiences.

Table 2: The seven guidelines of design science research.

These seven guidelines frame the work of the researcher. The guidelines might be tweaked and customized to the different domains in design science.

Design as an artefact
The result of the design science research process is an IT-artefact. Hevner defines this as a purposeful, innovative artefact in specific problem domain. This includes models, algorithms, methods, constructs and instantiations are as valuable as information systems. The results of a design science are not normally a full information system, but rather the ideas, models, practices and products which enable efficient and effective analysis, design, use and implementation of information systems. The artefact can be a proof by construction. The idea is often to expand the use of information technology to new areas and to solve problems in domains not earlier thought to be a domain for information technology.

Problem relevance
Research in information systems attempt to enable development of solutions to unsolved problems using information technology. It aims to create an artefact that takes a domain from a current state of a system to a goal state of the system. This can be formulated in business theory as maximizing utility or profits. The relevance of a problem can be stated in its relation to the community who will utilize it. A problem must therefore be relevant for environment of practitioners or the researchers in the knowledge base.

Design evaluation
There are a wide variety of evaluation metrics in design science and in the construction of artefacts in general. The evaluation must match the goal of the established requirements. Depending on the requirements the metrics can be from a wide variety of methods as shown in Table 3.

Categories	Description
Observational	Case Study: Study artefact in depth in business environment. Field study: Monitor use of artefact in multiple projects.
Analytical	Static Analysis: Examine structure of artefact for static qualities (e.g., complexity). Architecture Analysis: Study fit of artefact into technical IS architecture. Optimization: Demonstrate inherent optimal properties of artefact or provide optimally bounds on artefact behaviour. Dynamic Analysis: Study artefact in use for dynamic qualities (e.g., performance).
Experimental	Controlled Experiment: Study artefact in controlled environment for qualities (e.g., usability). Simulation: Execute artefact with artificial data.
Testing	Functional (Black Box) Testing: Execute artefact interfaces to discover failures and identify defects. Structural (White Box) Testing: Perform coverage testing of some metric (e.g., execution paths) in the artefact implementation.
Descriptive	Informed Argument: Use information from the knowledge base (e.g., relevant research) to build a convincing argument for the artefacts utility. Scenarios: Construct detailed scenarios around the artefact to demonstrate its utility.

Table 3: An overview of the main categories and types of design evaluation methods.

The nature of design means that it encompasses characteristics that cannot be analytically evaluated. The aesthetics of the artefact is at the designers discretion. There must be a degree of freedom for the designer to style the artefact. Human perception and taste is hard to evaluate instrumentally by any research methods.

Research contributions

A design science research process must result in a contribution to the knowledge base. The artefact is often the contribution itself. The artefact must provide a solution to a relevant problem. It can extend the current knowledge base or use previous knowledge in a novel way. The artefact in this way contributes to the environment of the IS community. An example of this type of artefact is system development methods. The second way that an artefact can contribute is by extending or improving the foundation of existing design knowledge. This can be novel methods, constructs or instantiations. This is the return value from the rigor cycle where knowledge is given back into the research community and its knowledge base. An example of this type of artefact is design algorithms. The last way that an artefact can contribute is through methodologies. Novel ways of evaluating, for example in the form of new analysis metrics are important for design science research. An example of this type of artefact is the Technology Acceptance Model (TAM) which gives researchers a metric to understand why organizations can be hesitant in using the researched artefacts. The artefact produced during the design cycle must contribute in one of these three ways.

Research rigor

Rigor in design science focuses on the effective use of the available knowledge base. The success of a project is dependent on the researchers effective use of appropriate methodology and techniques and the correct means to justify it. The design science researcher must constantly reassess their techniques and evaluation methods to understand why or why not the artefact works. The goal of design science is to figure out how well a solution works, not to theorize or prove that it could. It is more practical in that sense, and it is therefore important that design science is not weighed down by excessive focus on formalism and rigor. The demands of research rigor must therefore be balanced with the natural process of design science.

Design as a Search Process
To find a solution to a problem in design science is not a straightforward task. One method would be to assess the problem space and find every possible solution within the current laws of the environment. Laws of the environment in this context refers to the boundaries of the current technology and methodology and environment meaning the context of the problem space. This method would result in that the researcher would need to formalize every possible infrastructure, evaluate their utility and constrain and finally specify the cost and benefit of the possible solutions. This would be an enormous task. It is impossible to formalize the way to the optimal design. Design is therefore seen a search process that tries to result in a satisfactory solution. A solution which most importantly works, without necessarily knowing all the whys and hows of the solution. This enables the researcher to build on previous work and improving it or addressing its shortcomings as good design science research. This also means that a problem can be derived into its subproblems and be tackled in iterations, searching for the most satisfactory cumulative solution.

Communication of Research
Design science research should be well communicated to stakeholders on both the academic level, but also the practitioners in the environment. If a proposed design science research solution can appeal on the management level and technical level it is possible for practitioners to use the artefact in their work, but also enables researchers to build upon and extend the knowledge learned from the artefact. Business-oriented organization audiences are interested in the knowledge required to implement the solution and what the effectiveness and novelty of the solution proposed in the artefact.

4.1.1 The three-cycle view

Another helpful view on the design science research methodology is the three-cycle perspective. The three research cycles contextualize the design science research activities in the environment and knowledge base. There are three cycles, the relevance, rigor and design cycle. In the relevance cycle the researcher wants to find the current environment of an IS, the people, technical system and/or organization which will be improved with an artefact. Here the ultimate goals and requirements for the artefact is set. This cycle is also where the output of the research is evaluated. The output needs to be put back into the environment in order to find out if it in fact improves the environment and if so, by how much. Field testing is what proves if the right artefact or process was created. The rigor cycle is the bridge between the design science research and knowledge base which is based on. All research in design science build upon the vast knowledge of previous research and uses it as its foundation for further work. This ensures that the research is innovative and contributes back to the knowledge base.

> It is the rigor of constructing IT artefacts that distinguishes Information Systems as design science from the practice of building IT artefacts [13]. (Juhani Iivari, 2007)

The artefacts that contribute to the knowledge base can also include not only the artefact itself, but the new design products and process (meta-artefacts) and extensions made to existing theory and methods. One way of looking at the relevance cycle versus the rigor cycle is that the rigor cycle is elemental in providing the research credibility of the project while the relevance cycle is crucial in being able to pass the artefact on to practitioners. They have two different audiences, but both are important in creating good design science research. The last cycle is the internal design cycle. This is the heart and soul of a design science research project and encompasses the iteration of building the artefact, evaluating, and retrieving feedback. In this cycle one of the most important parts is to balance the efforts used in building versus evaluating and feedback. Both activities need to find a foundation in the environment and knowledge base. It is in the design science cycle that the hard work and research is done. For while the other cycles are important, the most effort must be put in the actual research.

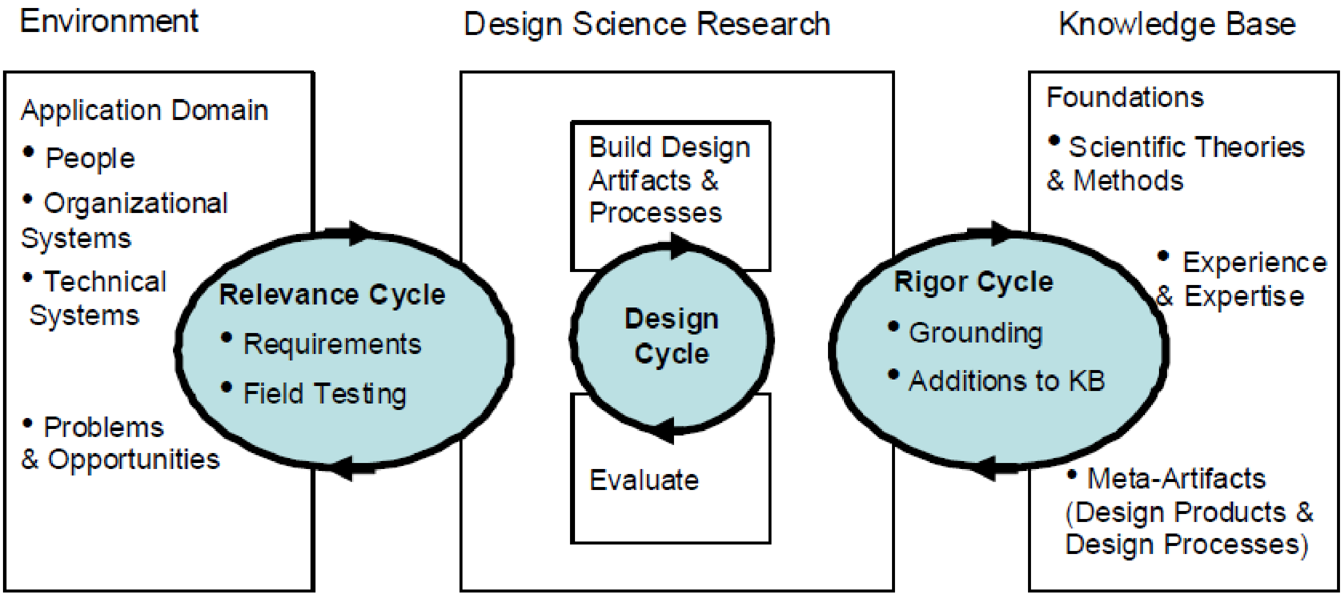


Figure 4: The three-cycle view of design science [11]

.

4.1.2 Knowledge Discovery in Databases

Knowledge discovery in databases (KDD) is a process which takes in data and tries to extract meaningful patterns, trends and information through a step by step process. This type of methodology has increased in importance. As the size of data increases so the need for new and more effective methods has increased. Traditional methods of KDD involve specialists who manually analyse and interpret a database and writes a report. This report is then used by decision makers to adjust accordingly. An example is the NIPH report on social inequality in health in the literature overview. This type of KDD is slow and expensive. With the ever-increasing size of data these old techniques have become impractical. Fayyad, Piatetsky-Shapiro & Smyth (1996) described the KDD process as:

> *The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.*

The process starts with a large digitized data set ready for processing. The first step is to learn the application domain using prior knowledge and formulating goals for the application. These goals are used to select the relevant data. The next step is data cleaning and pre-processing.

Deciding on strategies on missing values, removing outliers and noise, resolving any DBSM issues such as data types and schema are all operations which happens in this step. The processed data are then transformed in the next step. Transforming the data in this context refers to practices such as finding the most useful features to represent the data, reducing dimensionality of the data to remove redundant or unnecessary complexity for the goal of the project. Another option is to find invariant representations of the data. The transformed data is then data mined using algorithms that extract the most useful data according to the set goals. The functions can for example be clustering the data, summarizing, classifying or using regression. Data mining also includes the choice of which methods to use. What models and parameters that are best achieves the goal of the process? The mined data results in patterns which are then used in the final step, interpretation/evaluation. This step is used to evaluate and interpret the resulting patterns, returning to previous step if necessary. The interpretation also encompasses any visualizations needed to convey the patterns and making the patterns useful for the user. After the final step the final state of the process is knowledge. This part signifies the use of the knowledge, by improving the performance of a system, taking actions based on the knowledge gained or documenting and reporting or presenting it for interested parties. This also includes looking at any conflict between the new knowledge and any prior knowledge.

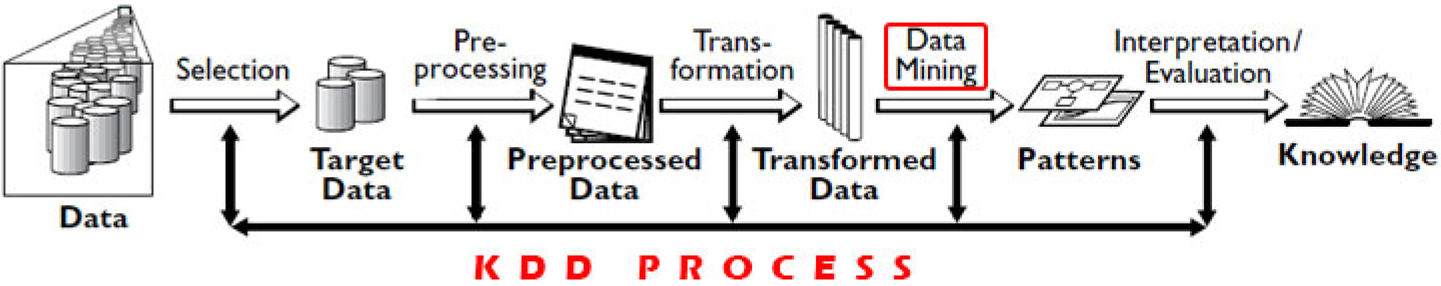


Figure 5: The KDD process as described by Fayyad, Piatesky-Shapiro and Smith 1996 [6]

.

The data mining step of the KDD-process has two distinct classes. The knowledge discovery goals that are used in data mining is either *verification* or *discovery* [7]. *Verification* is the where the system is limited to verifying the hypothesis set by the researcher. *Discovery* on the other hand is an attempt at explore the data set where the system is autonomously finding patterns. For this project the discovery was the goal. Data mining tasks under the discovery goals are mainly split into two groups, predictive and descriptive. Predictive data mining means building a model that can be used to predict future behaviour or values on a given feature based on historic data. This includes techniques such as classification and regression. Descriptive data-mining tasks on the other hand focuses on describing the data effectively, efficiently and understandable. Examples of techniques in this group are data characterization, which attempts to generalize the data in some way that conveys the characteristics or features of the target data. Another example is data discrimination, which compares general characteristics of a data set against a contrasting class. One of the challenges of the KDD-process is how to evaluate the interestingness of patterns found after the data-mining step. A problem that is present with all data mining who is automated is that the process often produces a lot of patterns that are of no interest to the user. KDD researchers has worked on defining measures of interestingness of patterns to combat this problem.

A paper by Klemettinen [16] observed that objective measurements are not sufficient in most data mining processes, for even with many strict rules of objective interestingness there were many objectively interesting patterns that were of no use for the user. Klemettinen suggested a template solution where the rules which decide of a pattern is interesting were not defined by the attributes of the data, but by a user by a user-specified vocabulary. The user-defined vocabulary would be defined in terms of the data attributes, so that a pattern is deemed interesting if it matches a restrictive template. Klemettinen suggestion lets the user into the process but does not address the issue of what is subjectively interesting and how it can be used to specify the templates.

Avi Silberschatz and Alexander Tuzhilin writes in their paper, What Makes Patterns Interesting in Knowledge Discovery Systems from 1996 [35], that there are two different ways of measuring interestingness. They describe the two measurements as objective and subjective. Objective measurements rely solely on the structure of the pattern and the underlying data used in the discovery process. Subjective measurements depend on the class of individuals who examine the results, meaning what is interesting depends mostly on who sees the pattern. What is interesting for group A might be of no interest to group B. Silberschatz and Tuzhilin propose a classification of interestingness which measures two factors, the unexpectedness and actionability of a pattern [35]. The unexpectedness means that the more surprising a pattern is, the more interesting it is while actionability means if the user can act to his or her advantage on the pattern. These two are combined in a classification, with emphasis on the first measure. A pattern is unexpected if it contradicts a belief we have. These beliefs need to be logically stated in a formula. The paper makes a distinction between two types of belief in this method. Hard beliefs and soft beliefs. Hard beliefs are constraints which cannot be changed regardless of evidence. A pattern contradicting this belief is regarded as an error. An example would be that life expectancy cant be lower than 0. Soft beliefs are beliefs that can be modified with new evidence. An example is the belief that sugar is not carcinogenic. Soft beliefs all have a degree of how strong we believe in it, and this degree can best be logically expressed using Bayesian probability. In this method the belief x is defined as a conditional probability of P(x|y), the probability of that x holds, given evidence y. An assumption is made for the belief held, and new evidence calculates the difference between held belief x0 compared to updated belief x1. The larger the difference is between the two, the more interesting a pattern is. If new data inserted into a belief changes the belief probability over a certain threshold value, it will indicate an interesting pattern.

4.2 Machine Learning methods

4.2.1 Pre-processing

Pre-processing describes the methods and techniques needed to prepare the data for analysis and use. Data in raw form are often problematic in scale, noise and inconsistencies. This is especially true for the data set used in this study, which is a combination of data sets from studies with several revisions over the years resulting in variables being hard to compare and difficult to use.

Feature selection
Feature selection is the process of choosing which variables are most interesting in terms of the prediction or output you want. There are automatic methods for feature selection, but for this project this was mostly done on a manual basis. The basis for the features selected were not for optimal prediction value, but to look at interesting variables that can convey certain knowledge. The important part was therefore to choose features that were interesting in solving the research

questions.

Feature engineering
Feature engineering is the task were new features are introduced or existing features are altered to fit the objective of the project. This is done to better define the structures and relations in the data. Creating new features can be difficult, because it requires a good understanding of the data and how features work together and impact one another. In this project feature engineering was made to reduce noise or data which unchanged would result in error. In effect this improves accuracy and makes the models more efficient. Another part of the engineering performed was the transformation of categorical variables. Machine learning dictates that all input must be numerical. Take a variable such as car maker. So first process is to label each category with a number that represents the corresponding car maker, ie making ford equal to label 1, Volkswagen equal to 2 and so forth. But the relative difference between 1 being the ford, and 10 being for example Hyundai is not bigger than the difference between two any other car makers. They are all distinct car makers, so the numerical labeling creates a difference that does not exist. So to avoid this issue, one employs binary encoding. Every car maker label is then divided to it's own column, ie carmaker_ford, with a corresponding boolean value of 1 or 0. So a categorical variable with 10 possible categories is divided into 10 boolean variables.

Numerical values are easier to work with directly, but do require some engineering for effective machine learning. Feature scaling is the method used to normalize the range of the features in the data set. This is done to avoid some features being vastly over-emphasized because of their large numerical ranges of possible values. This assumption might often not be the case, but more the nature of the raw data. The importance is more in the relative difference between value X and Y. To combat the false assumption that machine learning does on large numbers, one can scale the data. An example might be to set all values as a number between -1 and 1. This way, the difference between 10000 and 11000 is equal to the difference in 10 and 11.

4.2.2 Supervised learning

Supervised learning is a method for learning which adapts its model using data sets of correctly labeled input/output pairs. The supervised learning method uses this data to generalize and formulate a algorithm that can be inputted with new unseen data and label the output. Supervised learning often require manually made data sets to conduct the training, but after training it is able to process new, never-seen-before data. The type of supervised learning used in this study is regression. In regression the prediction is expressed in real numbers, known as a floating point number in programming. This is used, for example, when predicting income as a real number. This is opposed to classification models where a class label is the desired output, for example predicting if a flower is an iris or a rose. In other words if the outputted predicted values are continuous, the problem is one of regression, while if the predicted values are categorical a classification model is more appropriate.

4.3 Linear Regression models

Linear models are widely used and have been studied and applied for decades. As their name implies they perform predictions using linear functions of the input features.

The general prediction formula for a linear model is:

$$y = w[0] * x[0] + w[1] * x[1] + ... + w[p] * x[p] + b \tag{2}$$

For equation 2 x is the variable and w is the weight given a specific variable. The weight is the learned factor to adjust the model for better prediction. Every variable has their own weight and are iteratively modified to increase performance score. The b is the offset on the y-axis. y is the predicted value.

4.3.1 Generalization

When training a model on the data we want the model to be good at generalizing from a training set to a test set. The model should be as accurate as possible on the unseen test set, given the learned weights from the training set. The more complex a model is the easier it will adapt the model specifically to the training set, adapting its function to fit tightly and give great score on the training set. If the training set and test set are very similar this will work well and give excellent results. But for real data the difference between the training set and test might often be quite different. So that for a model to perform with maximum accuracy it needs to solve *the fitting issue*. The fitting issue can manifest in two states, underfitting and overfitting. A model which is underfitted is too general in its implementation. To illustrate this point, we look at one of the first major applications of machine learning. The labelling of spam mail. Spam mail is unwanted, mass-generated and mass-sent email, often used for advertising or malicious intent. A classification model which is underfitted would mean a model which is too general in its implementation and wrongly labels real emails as spam. The opposite of this would be an overfitted model which is too "strict" in its classification. Emails that are spam will not be caught in the filter and the model will not be able to catch new types of spam, because of the overfitted, narrow, type of filter implemented. The trade-off between the two forces of over- and underfitting is expressed in the bias/variance trade-off. The generalization error of a model can be expressed by the sum of three errors; bias, variance and irreducible error. Bias describes the simplifying assumptions that a model does to ease learning. An example would be that linear models assume that the data is linear in structure while it might be quadratic. A model with high bias(many assumptions about the data) is likely to underfit, resulting in a trained model that is too general. Variance is the amount of change that happens when a model is trained using different training sets. An ideal model will not change much from one training set to another. If the variance is low that means that the models has successfully found the underlying, input/output patterns and data structures which results in high prediction scores. A high variance often means that the model has an overfitting issue. The last part of generalization errors are the irreducible errors that describe the errors found in the data set itself. These errors cause noise in the learning process. This noise can only be reduced by preprocessing the data correctly, removing outliers and incorrect data. As a general rule, a complex model will often have a low bias and a high variance, while reducing the complexity of a model will increase bias and lower variance.

4.3.2 Regularization

One method of increasing generalization power of linear regression is to impose regularization on the model. The regularization manipulates the weights of the learned model in order to decrease over-fitting to the training data. In the process of learning, a linear model attempts to minimize error. The most common and classical method for this is the ordinary least squares. The model finds the weights and bias which results in the lowest mean squared error between the predictions and true values. To avoid the weights and biases being over-fitted to that exact training set, a regularization component must be implemented. An example of this is ridge regression which is one of the most commonly used alternatives of regularization. It minimizes the weights of every parameter to be as small as possible while maintaining a high prediction score. In other words it wants every weight to be as close to zero as possible to avoid having a few variables having all of the impact on the final predictions. An alternative to Ridge is Lasso. Lasso is similar to Ridge in that it attempts to minimize the weights to as low a value as possible. The difference is that Lasso can lower the weights to zero, meaning that the feature has no impact on the prediction value. This results in an automatic feature selection where only the features with weight over a certain threshold are kept.

4.3.3 Optimization

For optimization, one of the best examples are gradient descent. During optimization the method attempts to minimize a cost function. To do this a loss function is used to create a loss curve. It then follows an iterative process, where the learning rate determines how large the "step" down the loss function curve is. The goal is to find the global minimum of the loss curve. To avoid being stuck in a local minimum on the loss curve, a momentum is added to the descent. This controls for the "steepness" of the descent. It can be seen as a ball rolling down a hill.

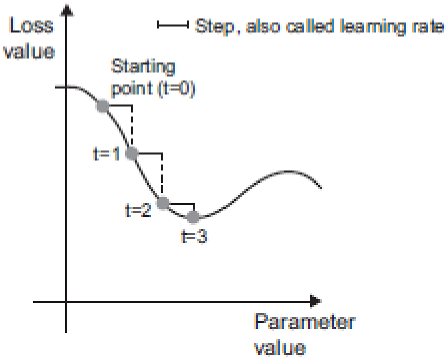


Figure 6: Illustration of the steps of the gradient descent optimization [4].

4.3.4 Tree regression

Tree regression models create a tree-structure which can be seen as a hierarchy of if/else questions, with the last node of the tree giving a predicted value. Another way of explaining it is to view at as a game of "Guess Who?". "Guess Who?" is a two player guessing game, where you have several characters and the opposing player can only ask yes or no questions to guess which character you are. Questions such as "does your character have black hair?" lead you down the tree structure and. For regression the questions are the same, but with numerical thresholds instead of yes or no answers, and the final answer is a continuous value.

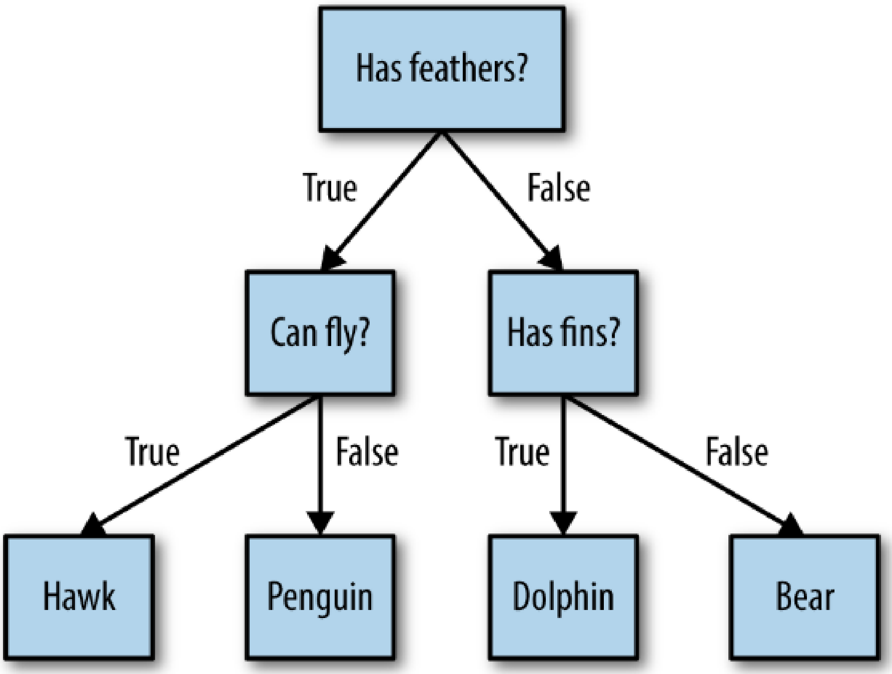


Figure 7: A decision tree to distinguish between animals [24].

Decision trees has a tendency to overfit to the training data if the trees are created too early in the learning process. In order to generalize better, a decision tree is pruned. This means halting the creation of the tree, and making it slower to establish the structure. A variant of the decision tree is the random forest regression. Instead of using the full data set to build one tree, random forest models choose random rows of the data set and construct multiple decision tree. The idea is based on that a single decision tree overfits to training data, but given a diverse amounts of decision tree which slightly differ form each other, it creates a better average score. The results of a new input is then sent all the trees and outputs the value that fits best according to the collection of trees. This has been shown to increase predictive power.

4.4 Evaluation

This section describes the purpose and methods of evaluation.

4.5 Purpose of Evaluation

The purpose of this evaluation is to find out if the artefacts produced during the KDD-process within the design science research framework answer the research questions defined in Section 1.1 and fulfill the requirements of the artefacts set in section 5.2. To evaluate the artefacts several types of evaluation methods were applied.

The methodologies used for evaluation were:

- A semi-structured interview
- Task-based testing
- Questionnaire - Visualizations Usability Scale
- Critical friend.

4.5.1 Semi-structured interview

This is a central part of the evaluation of the artefacts. It allows the participants of the evaluation to speak freely about the artefact, tasks and questions asked. This makes it easier to gather the thoughts and experience of the user, without the constraints of more structured forms of evaluation. The main advantage of the semi-structured interview form is that it allows for follow-up questions or feedback from conversation outside of the preplanned structure. This allows for a wider collection of interesting and valuable data. The questions are prepared and are asked after the tasks are performed as then the participants will have been familiarized with the artefacts. The first questions was based on the definition of interestingness by Silberschatz and Tuzhilin described in Section 4.1.2 in the part on interestingness. The next questions is more open-ended and lets the participants interpret interesting as they see it themselves. The semi-structured interview had the following questions as the main line of enquiry:

1. **Did the visualizations confirm or contradict any beliefs you have?**
2. **Did you find the result interesting? Why or why not?**

4.5.2 Task-based

To evaluate how well the graphs were understood, three tasks were designed to challenge the participants. By completing these tasks they need to familiarize themselves with the graph. This is also makes the interview afterwards better as it ensures that the participants have some basic understanding of what information is being shown. The three questions was related to information visualized on three different types visualizations. The first being a simple graph, the second a bar chart overview over a specific year and third being a more advanced bar chart with both time and different income groups. The participants had several graphs to choose from and where tasked to find out on their own which graph easiest answered the question and then to give the answer they believed to be true. The three questions were as follows:

1. **What is the trend of being female for cumulative income?**
2. **In 2017, what were two biggest factors that lead to lower income? And by how much?**
3. **In which year income group was the only time where being female was not negative for income?**

4.5.3 Questionnaire

A questionnaire allows for easy comparison between different participants and a general scoring scale for both usability and aesthetic. It has an inherent context for what constitutes good and bad. Meaning that an expression of a certain sentiment in the semi-structured interview form can be hard to know the exact context of. For example, a response on a question of "how do you like X", with responses such as "it's good" or "it looks bad", is hard to scale, because the context is subjective. The system usability scale(SUS), created by John Brooke in 1996[3], has ten questions with five response options on a Likert-scale[19], from strongly agree to strongly disagree. The ten questions are formulated in such a way that even-numbered questions expressing negative attitudes while the odd-numbered questions express a positive attitude. To calculate the score it therefore has a unique scoring system which results in a score from 0-100. The score has corresponding tables so the evaluator can find out what a good score is. Visual Aesthetics of Websites Inventory(VisAWI)[23][22] is a tool created by Mosagen & Thielsch for assessing the design of websites. It asks a series of questions in four different design categories. The categories are: simplicity, diversity, colorfulness and craftmanship. Each category has 4 to 5 questions and are divided between positively-keyed questions and negatively-keyed questions. Nielsen heuristics[26] is one of the most well known evaluation techniques for information systems, mainly websitets. The heuristics is a list of ten principles for user interface design. They are not as specific as the SUS or VisAWI, but are more general broad rules of thumb. The questionnaire constructed for this thesis are based on a combination of Nielsen heuristics, VisAWI and the SUS. By using these three evaluation methods, I have created a questionnaire focused on visualization. The quick and dirty visualization usability scale (VUS) is described underneath.

1. I think that the layouts are pleasantly varied.
2. I found the design uninteresting.
3. The visualizations are easy to understand.
4. I think that I would need the help of an expert to understand the visualizations.
5. The designs were generally well structured.
6. I thought the design was inconsistent.
7. I found the colour composition appealing.
8. I would imagine that most people would find the layout too complex.
9. I felt confident in my understanding of the information presented in the visualization.
10. I needed a lot of prior knowledge to understand the visualizations.

These ten question will be used together with the five option Likert-scale as a questionnaire after the semi-structured interview and presentation of the visualizations created in this project. The resulting score is difficult to contextualize, because this is the first implementation of it. But the research done by Bangor, A. and Kortum, P. and Miller, J. [1] give an indication for what adjectives describe a certain score, see Figure 8. The score of the evaluation will be based on this research, and while it's not perfectly applicable, it gives an certain indication.

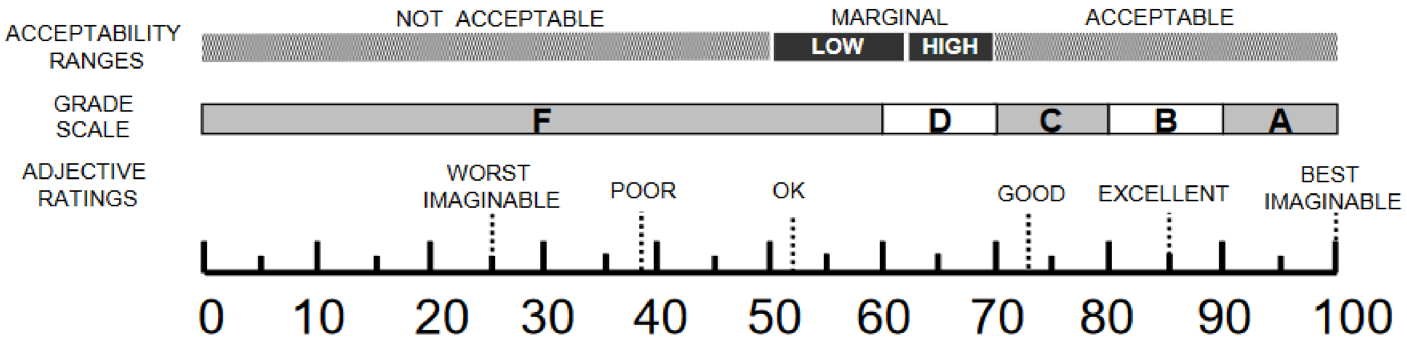


Figure 8: The different metrics to which a SUS-score can be judged.

4.5.4 Critical friend

A critical friend is a method of evaluation which has its origin from educational research. It has been best defined by Costa and Kallick in the educational leadership journal [5]. They describe the critical friend in this quote:

> *A critical friend can be defined as a trusted person who asks provocative questions, provides data to be examined through another lens, and offers critiques of a persons work as a friend. A critical friend takes the time to fully understand the context of the work presented and the outcomes that the person or group is working toward. The friend is an advocate for the success of that work [5].*

The critical friend(s) in this project are the other master students that either work with me in the same supervisor-group or has a similar type of machine learning project, in addition to the supervisor.

Chapter 5

Development

This chapter describes the requirements set for this project, the tools used and the development iterations.

5.1 Ethical Considerations

This project has been approved by the Norwegian Centre for Research Data(NSD). All participants in this study were informed and consenting. The approval from NSD can be found in Appendix A and the consent form can be found in appendix B. All participants were informed of the use and storage of their personal information. Participants have also been informed of their right to be removed from the study at any time. No sensitive information have been collected during this research.

5.2 Establishing Requirements

A requirement is according to Preece et al. [30] a statement of how the product will perform or do. Requirements are identified and captured through the establishing activity, the first part of the process of These requirements are split into to main categories, functional and non-functional. Functional requirements describes what the product will do, while the non-functional describe the characteristics, also known as constraints, of the product. The listed requirements have been modified since the start of the project and are not the exact requirements at project start, but the finished set of requirements.

5.2.1 Functional Requirements

- Process comma separated values (CSV) files.
- The data must be preprocessed
- Make use of and test multiple machine learning(ML) models.
- Use the best ML method to mine patterns from the data.
- Produce visualizations from any of the variables.
- Produce different types of visualizations.

5.2.2 Non-functional Requirements

- The visualizations must be intuitive to understand.
- Users must be able to extract knowledge from visualizations.
- The visualizations should be appealing.
- The visualizations should have good design and layout.

5.3 Iterations

5.3.1 Development tools

Python

Python is a high-level object oriented programming language created in 1991 by Guido van Rossum[38]. It is open-sourced and has many of the best libraries in machine learning. This was the biggest reason for choosing python for this particular project. Libraries such as pandas and scikit-learn makes it easy and efficient to work with data sets and machine learning. In addition, it is a personal preference for my personal familiarity and experience.

Matplotlib

Matplotlib is a plotting library made for Python. It is a tool for making or creating static, animated, and interactive visualizations in Python [12].

Pandas

Pandas is a powerful tool for working with data. Pandas offers numerous way of reading, writing and manipulating data objects both fast and efficiently [20]. It offers a lot of out-of-the-box solutions that work very well with other libraries, such as sci-kit.

Sci-kit learn

Sci-kit learn is a machine learning library started as a Google summer of Code project in 2007 [29]. It is today one of the most popular machine learning libraries and has been used for both pre-processing and modelling. It has a rich variety of methods and models which are pre-built and configurable through tweaking of parameters.

GitHub

Github is a tool for working with version controls during software development. Github makes it easy to maintain and branch out different solutions for the software and is crucial in task management and keeping an overview of the project.

Pycharm

Pycharm was the chosen integrated development environment (IDE) for this project. Pycharm has many easy to use and powerful functions which ease the running and bug fixing during development.

Trello

Trello is a planning tool in kanban-style. Kanban is a development method created for organizing the development tasks [17]. Trello is a web based board that can be stickied with notes. The board has different columns with the different steps in the process of development, such as Backlog, To Do, Doing and Done. It provides an easy interface and is a great planning tool.

5.3.2 First iteration

Selection of Data Features

The first task in the development of the artefact was to choose the appropriate data to conduct this research. The level-of-living-studies have several variables which has been altered throughout the revisions. The challenged was to find variables who were consistent throughout the decades. The first iteration started the work of looking at the different data sets and comparing them. For this work a number of factors for selection were looked at. The first iteration was also the start of ML models, but then with the easier data sets after the EUSILC standardization of the data sets. The focus for the first iteration can be summarized in three points.

1. Look at what features of the different data sets that could be used for the final data set.

2. Start to work on ML methods on the newer EUSILC surveys which are uniform and therefore easier to implement and test ML techniques on.

3. Create the first visualizations based on the first ML models.

The data needed a lot of cleaning and pre-processing in order to be useful for data analysis. This meant that the use of all level-of-living surveys would be time-consuming and impractical. There is a total number of 38 survey data sets with hundreds of questions for every year, with a variance in questions asked and formulation of the questions. With a non-uniform data it was important to decide on which years would be included for this project and scope. The years were selected for both the usability of the data and the goal of this thesis. For the purpose of this thesis a representative selection of data sets were chosen. This was the data sets of the years 1973, 1983, 1995, 2005, 2013, 2017. These data sets were chosen on their likeness to each other and for the time span they covered.

With the chosen data sets established the next focus was to explore each individual data set. Each question asked in the surveys was defined as a variable. The first tasks was finding the best variables to use for this project. All variables available in the data sets are of interest, since the surveys are constructed by experts to produce interesting or important insights on general welfare and standards of living. To decide which variables that should be used in the KDD-process several factors were considered.

Long-term data
One of the biggest factors in choosing variables were the similarities of the variable throughout the data set. A variable unique to a single data set would be of little use for the purpose of this thesis. And while there are some variance in the type of questions asked or how they are asked throughout the data sets, there are some variables that are comparable despite the changing of the surveys. These were especially important since the time-aspect, looking at long-term trends, is important to produce visualizations that reflect the true state of society.

Inconsistent data
Throughout the data set there are a large number of inconsistencies. An example is the variable that describes the number of children that the interviewee has. In some data sets this is described in two separate variables where the age is divided on young children between 0-5 and a second variable with older children 5-16. Some data sets also has a separate question for children moved out of the household. To solve issues such as this variables were constructed to be useful for the combined data set. Meaning that the separate children variables was included in a single numberOfChildren0to16 variable. For every variable this is a weighing of the usability and precision of the data. More precision means more complex findings and results, but it can be harder to find comparative data from other data sets. The inconsistent data is the principle factor for the variable complexity and usability. With data sets from a changing collection methodology and questions from decades of revisions this posed a big challenge in terms of time and finding the best possible results.

Missing data
For some data sets there are missing data. In the example of "antbarn"(number of children)-variable there were a number of years with missing data. Here some of the years have had that variable anonymized with every interviewee being assigned with the value 0. This means that any comparison on this specific variable is impossible for those data sets. Since the data can be missing for some years it demands that the methods and machine learning that is applied must not be reliant on every data point being present for every calculation.

Overly-specific data
This points to the applicability of the variables. Some of the variables are more niche in their formulation and might not be that interesting for the analysis. An example might be the earliest health surveys which covered the distance from the household to different public services, such as nearest clinic, pharmacy and type of transportation to each. These variables were too specific for scope of this project.

Year	No of variables
1973	408
1983	667
1995	614
2005	2346
2013	1232
2017	1263

Table 4: Overview over the number of variables from each survey.

The high number of variables on the later surveys comes from repeated questions for every person in the household, for example several questions for each child or other person other than the interview object (IO) living in the household. These have later been anonymised and only the questions asked the IO are present in the data sets. The variable to be examined in this study was the cumulative income of the interview object. This represented all income to the person, meaning all benefits, salary, capital income, pension and more. This was used a base for all data mining, to find out what factors lead to lower or higher income and by how much. In all literature reviewed in Chapter 3, income stood out as a very important factor. The final variable-list was defined by the rules stated in selection, and for its potential interestingness.

Variable	Description
Aargang	The year the survey was conducted
Alder	Age of IO at time of interview
arb1	Did the interview object complete 1 hour of work last week
kode218	Did the interview object receive disability benefit last year
saminnt	Cumulative income of last year
helskomb	Serious injury and/or illness last year
utdnivaa	Level of education
sivstat	Marital status
kjonn	Gender

Table 5: Variables with descriptions

Underneath is the variables and what engineering has been done from the original data set.

Disability payments were in some data sets a number between 0-100, with the answer being what degree of disability has the IO. Others had a sum which signified the amount of money. Third variant of the questions was a yes or no questions, for example: do you receive disability payments? To be able to use this data and compare, all variables were simplified to a yes or no questions. A degree of disability > 0 was set to 1, all sums received > 0 were set to 1, and a yes answer is set to 1.

Alder were mixed between year of birth and age at interview time. This was simplified to a general age, and has therefore a error margin of > 12 months on older data sets. The year of birth and actual age creates a possible error margin of < 12 months. For example, if a IO has birthday at 31. December of the year of the interview, but the interview was done on the first of January.

Arb1 This variable was present in all data sets under different variable names.

Samminnt was similar in all data sets.

Helskomb was a combination of separate questions of injury and sickness. This questions was described differently in various data sets. Either they were combined, or separated into two questions. This variables was made to see how health might impact income, without focus on the specific cause nor the specific length of injury or illness.

Utdnivaa represents one of the more challenging variables to compare over the time period in this research. This on the basis of a educational system which has been through multiple reforms and changes. The earliest data from 1973 has IOs with education from the early 20th century, where basic education were much less comprehensive. To compare number of years in education directly would lead to incorrect trends, because the number of years of education has increased. Therefore the stage of education are the point of reference. To help to compare this data, sources such as the history of the Norwegian educational system from Store Norske Leksion was used[25]. But from 1983 and on wards there are some standardization for education in the variables. This was based on Statistics Norway handbook from 1973 with the title NORWEGIAN STANDARD CLASSIFICATION OF EDUCATION[10]. But the definitions had a final revision in 2000 with the new Norwegian Standard for Educational[2].

Machine learning methods

The first data set that the ML methods were tested on consisted of a selection of variables from the EU-SILC data sets, which were the most uniform and consistent data and therefore easiest to work with. The first iteration of the ML models tested this data on several different machine learning algorithms. The results were then compared, and several considerations were made on which to choose for this thesis. To test different methods, I tried out six different linear regression variations for my data. The six were linear, Ridge, Lasso, random forest regression and decision tree. I also optimized using Stochastic Gradient Descent (SGD). The data was normalised using the sci-kit libraries methods for scaling. Data was also one hot encoded. Table 6 presents the best performing numbers from each model in terms of precision.

Type	Mean Absolute Error
Linear	188,803
Ridge	188,617
Lasso	188,617
RFR	198,376
DT	205,987

Table 6: The tested types of regression with their mean absolute error
,

On the basis of this analysis of the different machine learning methods I decided on Ridge regression as my preferred model. It represents the best results and is the easiest to implement on the data set. It also had coefficients which had data which were easy to understand and was the best starting point for visualization the patterns in the data. The decision tree models were able to increase the precision of their predictions and lower their error to that under Ridge originally. But the learned tree structured were complicated and difficult to understand and extract knowledge from. The tree hierarchy created by decision tree model was when unlimited extremely large and complex. When restricted to a tree structure of a less complex form, the precision of the tree regression decreased to below that of ridge. In the end it proved not beneficial to the goal of this study. It was hard to read for experts and impossible for the "average Joe". And when readable for experts its results were less impressive than the much more straightforward Ridge regression model.

The first iteration also includes the initial work on visualizations. As a starting point the last part of the data set were chosen to analyse initially. These are the data sets ranging from 2011 to 2018. These surveys represent the most homogeneous data sets with clear structure which is consistent on a yearly basis. While the surveys have three rotating subjects on a three-year cycle, many of the questions are similar. The data has also been collected with the same variable names, making the process of extracting them easier and less time-consuming. While this data was not part of the final artefacts, it was a great starting point. They enabled easy visualizations and to test the validity of the research. The idea was that a proof by construction, making actual working visualizations, here would make it easier to later apply the same methods and models to the older data set with more inconsistency. The first graphs were crude but gave initial feedback on any errors in the preprocessing and gave validity to the process and models.

5.3.3 Second Iteration

The second iteration started with finding and correcting any preprocessing bugs from the first iteration. Bugs that resulted in nonsensical results caused by missing, wrongly formatted data inputs, general inconsistent data or visualizations which were erroneous. For the second iteration there was also a focus on getting the entire data set combined and processed so it was usable for the models. This meant combining the data sets of 1973, 1983, 1995, 2005, 2013 and 2017. And while these data sets have all been digitized by the NSD, they pose a variety of challenges. The data from the older data sets have different variable names and different types of naming conventions. The questions also often differ in terms of how specific they are. For example, *sivstat*, the variable for marital status, has in 5 of the 6 data sets, five possible answers. These are unmarried, married, widow/widower, separated and divorced. The difference between separated and divorced is that a married couple by law must be separated for 1 year before divorcing. But for a single data set, the year of 1995, the only question regarding this marital status is a yes/no questions, 'are you married?'. Variables such as this create challenges in terms of the tradeoff between detail and compatibility. I want the information that details give, but it must not make the longer trends unusable for visualisation. Each variable had to be assessed and processed in each separate year.

To help with all data exploration, NSD has a overview over the variables, their names and number of replies for each question. And a general description of the variable. The newer data sets have variable names that reflect what they are, such as 'alder'(age). But the older data sets have simple variable names such as v1034, which in itself has no meaning. So the data set needed annotating and streamlining.

In the second iteration the Ridge regression model was the only ml method used, based on the testing in first iteration. The coefficient was processed and converted to percentages based on the median income in the data set. There was generally a main focus on two tasks. Making the data set complete and creating more complex visualizations. The resulting visualizations were the graphs with a combination of trends displaying in one plot. Instead of having six separate graphs with a single line, they were no combined in a multiplot graph.

5.3.4 Final iteration

The goal of the third and final iteration was to increase the complexity of the visualizations on the basis of feedback and suggestions from the supervisor-group. The increased complexity includes wage-groups and the impact of factors based on the different wage groups. The wage groups were based on documentation from Statistics Norway. There are no strict definition of what constitutes a person of low income, but an often used metric is 50 or 60 percent of median income per consumption unit (equivalent income). The OCED standard is 60 percent.[27] This income takes into account the household income in addition to the number of members of the household. The need of a household increases with the number of members, but not linearly. A EU scale is applied where the head of the household is weighted a 1, while the next adult is 0.5 and a child under 17 is 0.3. In the data of this study the IO is always the head of the household, as the next of kin of the IO has been removed from the surveys for privacy concerns. Therefore, these wage groups are based on the OECDs definition of low-income workers. The wage classes are divided into three; low, middle and high class. The low wage group was the group which had a cumulative income of between 0.1 to 0.59 of median income. The middle income class was the IOs between 0.6 and 1.39 of median income. And the high class was for 1.4 and 2.2 of median income. The reason to not include the > 0.1 income is because they skew the data heavily towards 0 and makes the models for low-income individuals perform poorly. For the high-income class the ceiling was set at 60 percent over median income. The reasoning behind setting a ceiling is the same as for the 0 kroner incomes. The outliers skew the results of the data mining which create visualizations that are not representative of the state of society. For example if a small number of IOs have a cumulative income of several multiples higher than any other, the factors they have (are male, of a certain age in a certain life situation), are heavily skewed. This excludes the very wealthy from these final graphs, which takes away some of the finer detail. But in order to keep the visualizations easy comparable, the wage groups were adjusted for both max and min income.

The final visualizations were made using a vertical plotting for dates and a percentage increase or decrease based on the variable being assessed. The chosen factor to use for the final iteration was gender, because it has one of the clearest positive trends, but there is still inequality and I believed that it could be interesting as gender wage gap and income differences is one of the hot topics of today. It is also something I believe most people have an opinion on.

Chapter 6

Results

This chapter summarizes the main artefacts created in this project. It describes the resulting data set, the models created and the resulting pattern and their visualization.

6.1 Data set

The resulting data set is set from 1973 till 2017. It has been filtered so that only rows(individuals) that have worked at least one hour the last week have been included in the research. There is also an age filter which includes interview objects from the ages of 24 to 64. This is to ensure that most rows are working in some capacity and are not part-time students or pensioners. The size of the data set consists of 37287 rows with 9 variables before binary encoding. The final data with the 22 binary encoded variables are:

1. aargang(year)
2. alder(age)
3. arb1_1(worked more than 1 hour last week)
4. Uføretrygdet(receive disability payment)
5. SamletInntekt(cumulative income)
6. helskomb(healthcombination, long-term illness or injury last 12 months)
7. IngenUtdanning(no education)
8. Barneskole(primary education)
9. Ungdomsskole(lower secondary education)
10. VideregåendeGrunn(upper secondary education(basic education))
11. VideregåendeAvslut(upper secondary(final year))
12. Påbygg(post-secondary non-tertiary education)
13. UniversitetBachelor(first stage of tertiary education(undergraduate level))
14. UniversitetMaster(first stage of tertiary education(graduate level))
15. Forskernivå(second stage of tertiary education(post-graduate))
16. Ugift(unmarried)
17. Gift/partnerskap(married/registered partner)
18. Enke/enkemann(widow/widower)
19. Separert(separated)
20. Skilt(divorced)
21. Mann(male)
22. Kvinne(female)

6.2 Models

The machine learning model decided upon and created was the ridge regression model. It produced the most intuitive and easiest to understand patterns. The coefficients created by the model is described in the same value as the inputted value. Meaning that the coefficient shows the value of every factor in value of Norwegian Crowns. For some graphs this value has been converted to percentage to increase comparability over time, since inflation has decreased the real value.

6.3 Visualizations

The first and simplest visualizations that were created where the graphs which describe the development of a certain variable over the 50 years.

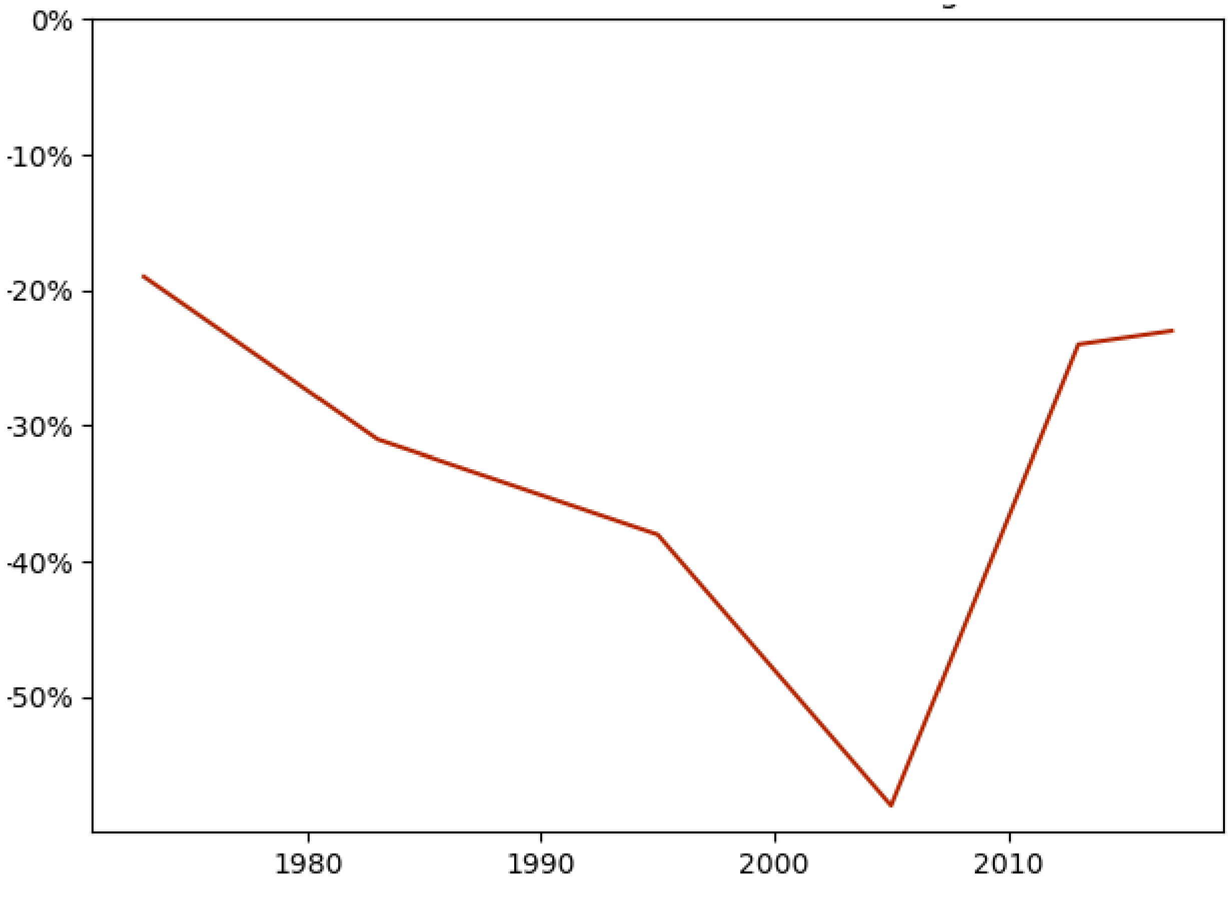


Figure 9: Percentage decrease in cumulative income based on lower secondary education

Figure 9 describes the impact on cumulative income if lower secondary education is ones highest completed education. This graph has an x-scale of years and a y-scale of percentage. The percentage shows how much lower than the median income the factor of having lower secondary education as the highest level of education. The values from the model are shown in percentage instead of numerical in order to be comparable throughout the decades.

The next visualization is a bar chart which describes the impact of the selected variables in that year.

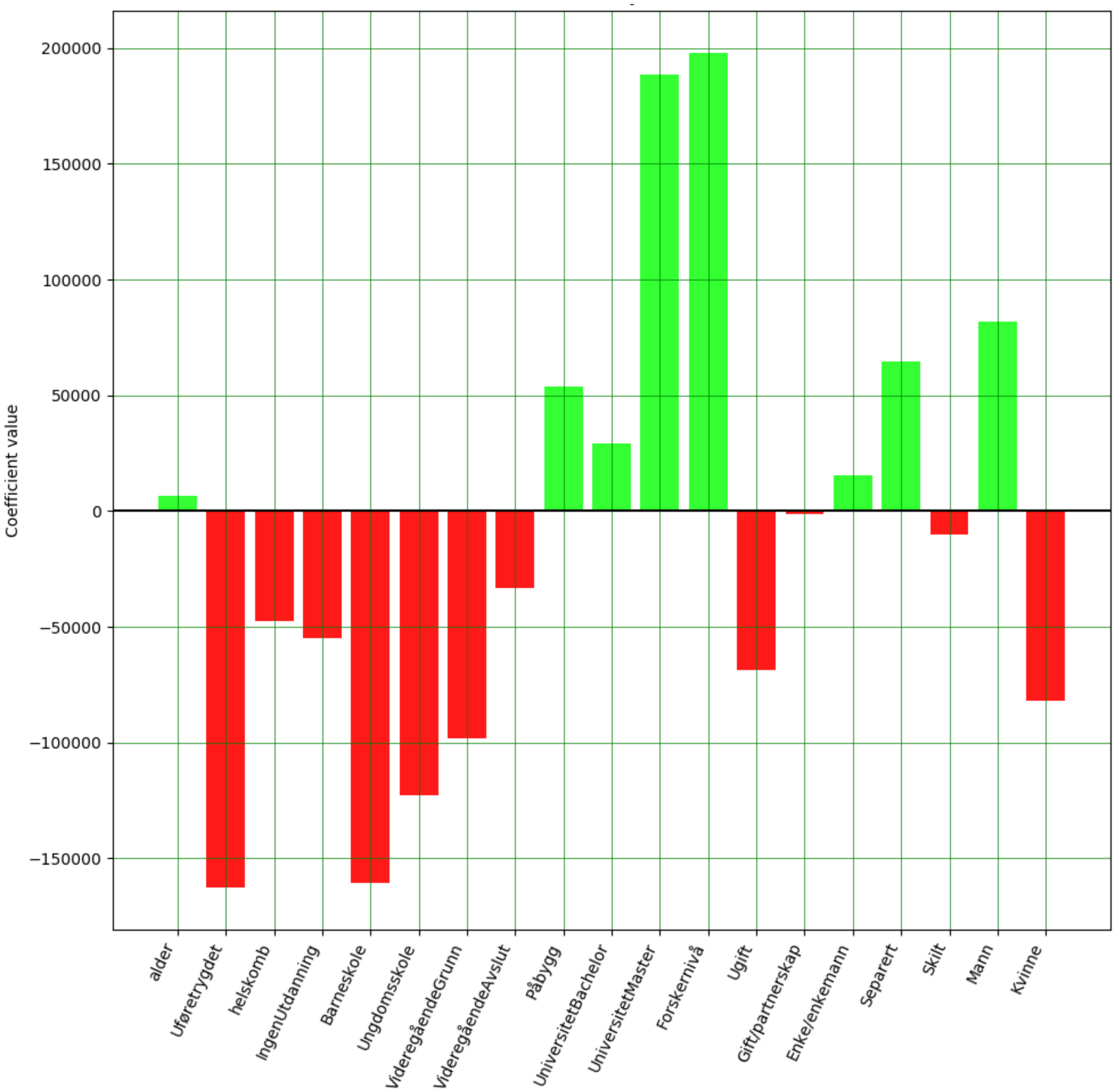


Figure 10: Coefficient of the year 2017, the variables underneath in Norwegian. Described in English in Section 6.3.

In Figure 10 the x-axis are the different factors or variables, and the y-axis are the amount that the model adds or subtracts from the predicted cumulative income. Here the relative importance of each factor is clearly seen. And it gives recognizable value in terms of that the coefficient value is the amount of Norwegian Kroner you are losing/gaining on having a certain factor. For this illustration the variables are given in Norwegian, but are from left to right:

- age
- Receives disability benefits
- Health combination(long-term injury or illness last 12 months)
- no education
- primary education
- lower secondary education
- upper secondary education(basic education)
- upper secondary(final year)
- post-secondary non-tertiary education
- first stage of tertiary education(undergraduate level)
- first stage of tertiary education(graduate level)
- second stage of tertiary education(post-graduate)
- unmarried
- married/registered partner
- separated
- divorced
- male
- female

The third visualization is an overview of the trends in selected variables in a certain income group.

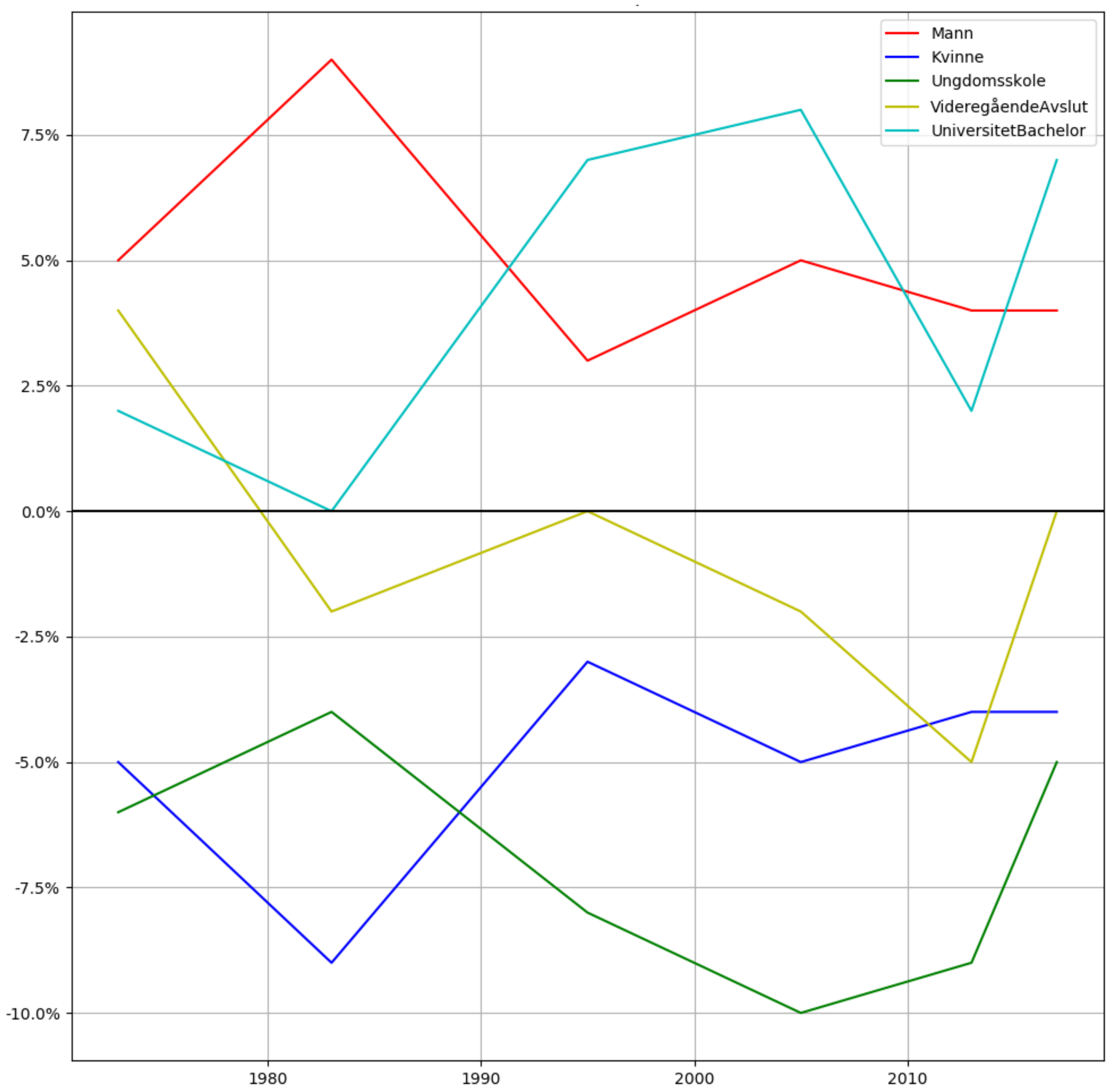


Figure 11: Percentage impact on cumulative impact over the last 50 years in the middle-income class

Figure 11 can be seen as the next iteration of the simple first graphs. As a multi plot graph, it creates an easier way to see the trends and compare the different factors. It's also divided by income groups which makes the resulting percentage differences less impacted by outliers and selection. The variables in the legend of the graph is as follows:

- Red: Male
- Blue: Female
- Green: lower secondary education
- Yellow: upper secondary(final year)
- Light blue: first stage of tertiary education(graduate level)

The fourth type of visualization that were made is a visualization which describes the impact of a variable based on year and income group.

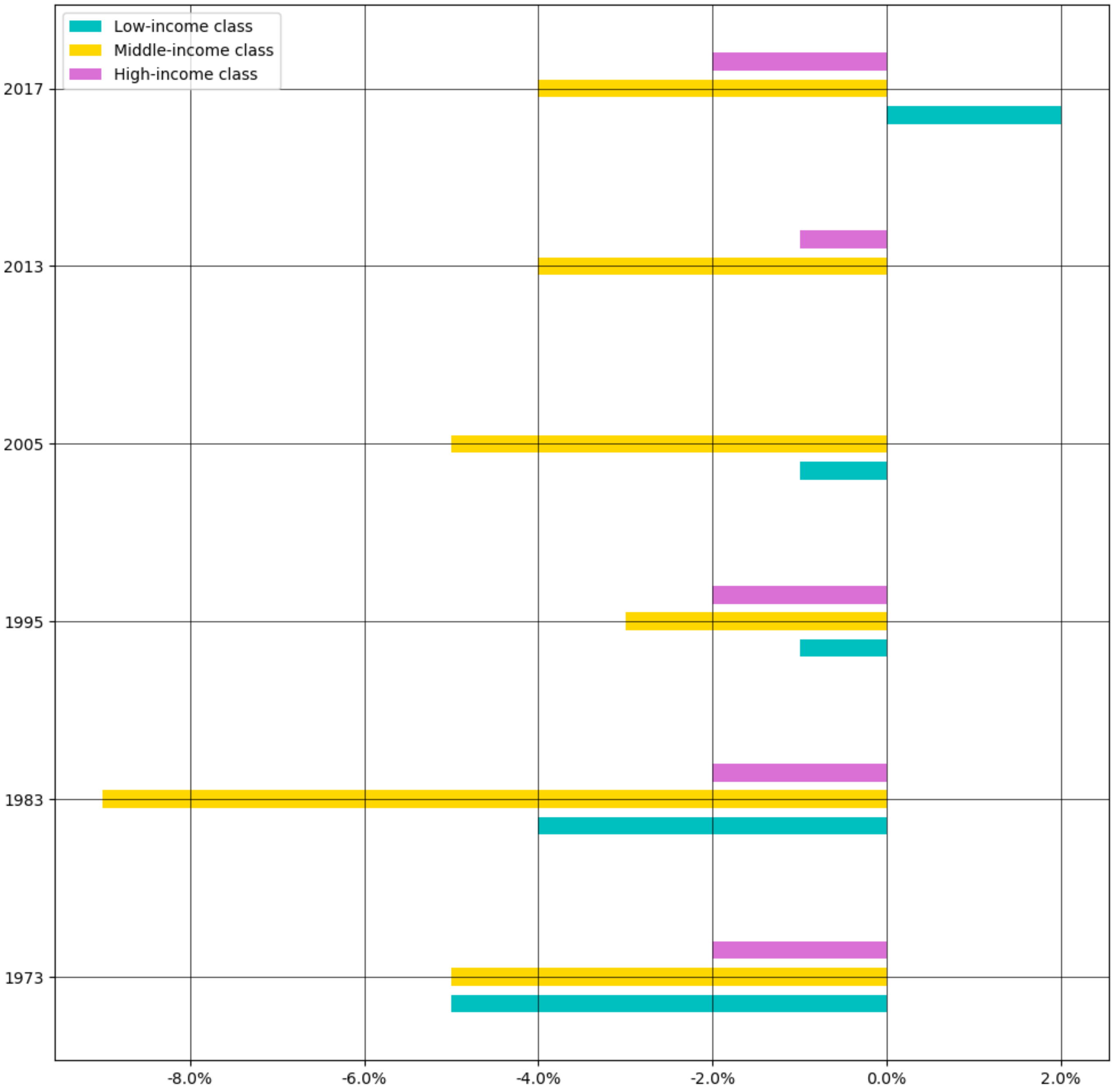


Figure 12: Difference in percentage cumulative income based on the variable: female

In the final visualization type created in this project, Figure 12, we have a number of bar charts grouped by year, with three bars representing the effect of the variable in their respective income class. This visualization style allows for deeper insights into the nature of variable for each income class. It shows for example how being female in the low-income class has had a better trend than for the two others, while for individuals with high-income, being female is about as negative as it was in 1973.

Chapter 7

Evaluation

This chapter presents the evaluation of the finished visualization. The evaluation was performed using the evaluation metrics described in Section 4.4.

7.1 Participants

The intended target group for these artefacts are the common man, meaning no expertise in terms of any social studies field. The evaluations were performed by five master students at Department of Information Science and Media Studies at the University of Bergen and one graduate student in professional studies in psychology. The IT-students represent a user type with a high degree of IT-expertise, but without expertise in general social studies, such as sociology and anthropology. The psychology student represents a person with an average knowledge of IT and no expertise in social studies. All participants have an academic background.

7.2 Semi-structured interview, tasks and observation

The interview started with a general presentation of the context of the project and then shown the visualizations they were going to evaluate. Then afterwards they were given the tools to navigate between the different visualizations that were prepared for them. There was a total of nine visualizations of the different types presented in Chapter 6. After the participants had been given some time to familiarize and understand the visualizations to the best of their abilities, they were given the three tasks. Which would test if the visualizations were easy to read and intuitive to extract knowledge from. Some participants struggled to find the right visualization and where shown the graph which had the information needed to answer the questions.

The first question was based on the graph shown in Figure 13: What is the trend in impact of being female on cumulative income? This questions was quickly answered by most participants. One participant was unsure of the trend before noticing that the y-axis went from negative to zero, which can be an unusual perspective.

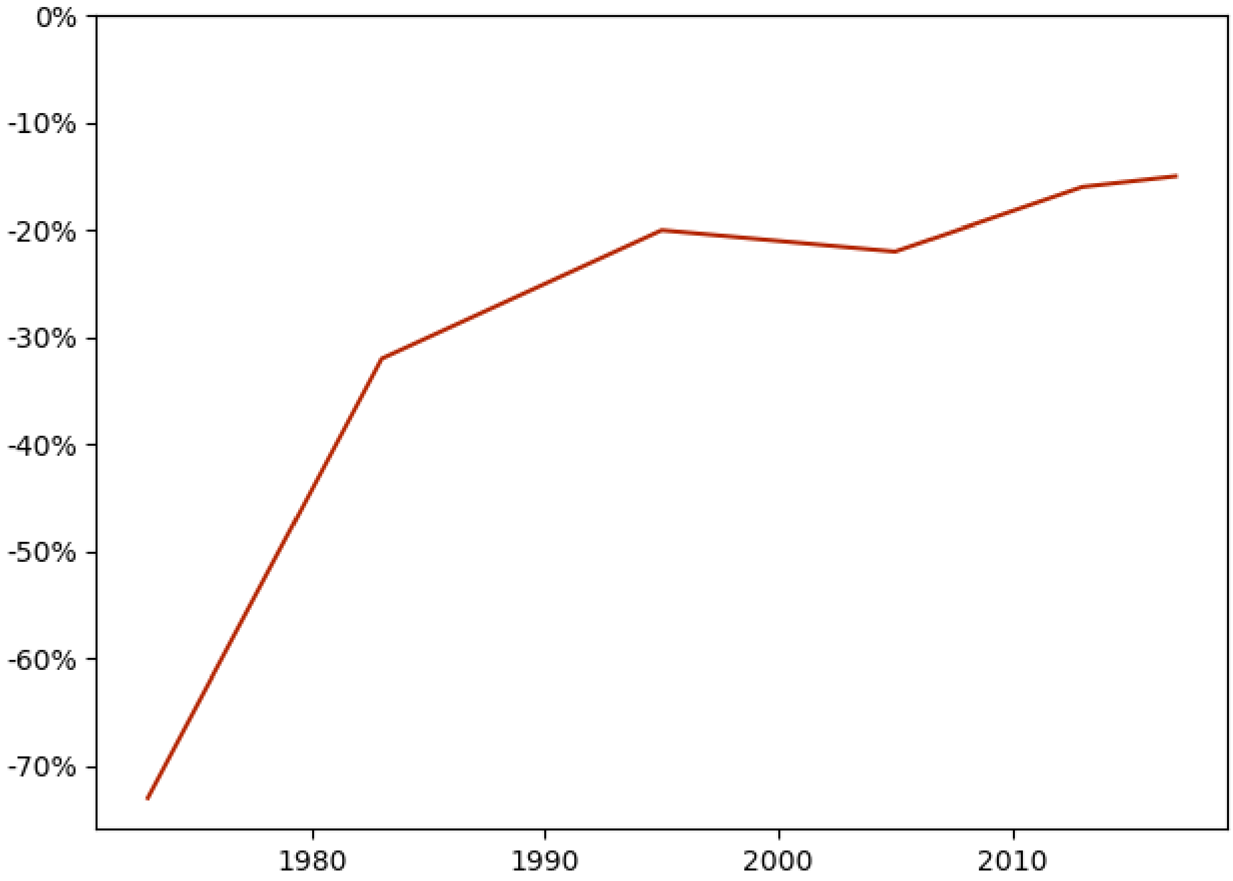


Figure 13: The development of cumulative income compared to median income by the variable female

The second question: In 2017, what were two biggest factors that lead to lower income? And by how much?

This question asked the participants to find a visualizations which showed an overview of multiple variables and had a numerical value as one of its axes, see Figure 10. Some also were unsure what the variables meant, an example being "primary school". It was not intuitive for some that this meant that primary school was the highest completed level of education. This information was part of the quick presentation given before evaluation, but was forgotten by some and they did therefore not understand the variable names. But as soon as they understood the context for the variables the question was answered quickly. This graph was also easier to read for some because it followed the pre-established concept for negative and positive. The colours of the bars are red for negative values and green for positive.

The final question was the hardest to answer for most participants. What year and income group was the only time where being female was not negative for cumulative income? Through observation some struggled to understand the axes and was unsure of the exact year on Figure 12. They wondered if the low income-class was from a different year, as they saw the income class were spread out on the timeline. Also, they didn't immediately understand the different income-classes.

Overall, the tasks were completed correctly by all participants and most found the right answer. Time varied between participants and the final questions took for some more than 2 minutes.

After the tasks, two questions were asked. Under follows the questions and some highlights from the interviews.

Did the visualizations confirm or contradict any beliefs you had?

Regarding the difference in income between man and woman, the visualizations shown strengthen my existing view. That the female variable for the low-income class for 2017 is positive is somewhat surprising.

A. Iden, 25

It confirms most of my views, such as increased equality between the sexes. It reflects what I hear from the media in Norway and from other countries as well.

E. H. Wiik, 28

The fact that the high-income class individuals were negatively impacted by being female surprised me. I believed that the emergence of business-women and that type would decrease the difference, but it didn't look that way on the visualizations. It also surprised me that there is still such a large difference in cumulative income for women overall.

K. Raudstein, 29

They confirm what I already believed. But I was surprised that having a bachelor as highest completed education did not give a higher increase in income in the newer years. But perhaps we have become a society where everyone gets an education and has a bachelor. Otherwise most visualizations reflect my views.

S. Stegane, 25

The visualizations confirmed my established view on inequality between men and women. There was no big surprises. What surprised me a bit was that for the low-income individuals there was a positive link between being female and income. I didn't think there was any groups where women had higher income than men.

T. Dery, 25

I believe for the most part it confirms my established view. I found the differences between the marital statuses unintuitive for me. It doesn't make entirely sense. It makes me wonder why divorced individuals have lower salaries than those married. I was also surprised that being female was a positive for low-income individuals.

V.Johansen, 26

Did you find the result interesting? Why or why not?

Yes, especially the ones were there were multiple graphs or variables in one. In these examples you get quite a lot of information. And when they span over time they are even more useful. A lot of the graphs are very informative.

A. Iden, 25

> *I thought the visualizations were interesting, because they presents something about how society has changed over the years and that is something that is good to know about and have some numbers on.*
>
> *E.H. Wiik, 28*

> *I found them visually appealing and the information was clear and easy to understand. Although the combination graphs where in some cases confusing with converging curves.*
>
> *K. Raudstein, 29*

> *You always hear about the inequalities in society, but when you hear numbers it doesn't seem as dramatic as when you can see the actual differences in visualizations. When I see numbers it can often lead to losing interest or not really getting any information out of it, but when you actually can see it, its easier to interpret and understand the actual differences.*
>
> *S. Stegane, 25*

> *I found it interesting. It was very fun, because one always has certain suspicions on the difference between men and women in income and how it has been throughout the last decades. Very fun to have some research behind ones perception.*
>
> *T. Dery, 25*

> *It was definitively interesting. One has perhaps a crude overview over inequalities in society and in income, but it is interesting to see concrete values and percentages and to get it on a variable level.*
>
> *V. Johansen, 26*

Observing the participants doing the tasks and interviews also raised some questions from the participants that I did not have the answer to. Mostly if they found a variable surprising they would often ask why this is so. While I could explain some potential limitations in selection or some changing definitions throughout the years, I could not answer the question of why do women make more in low-income and why do divorcees make less than married individuals. These are questions that an expert or further research must answer.

7.3 Visualization Usability Scale

All six participants were asked to answer a visualisation usability scale form post interview. The scores were then calculated as with a normal SUS.

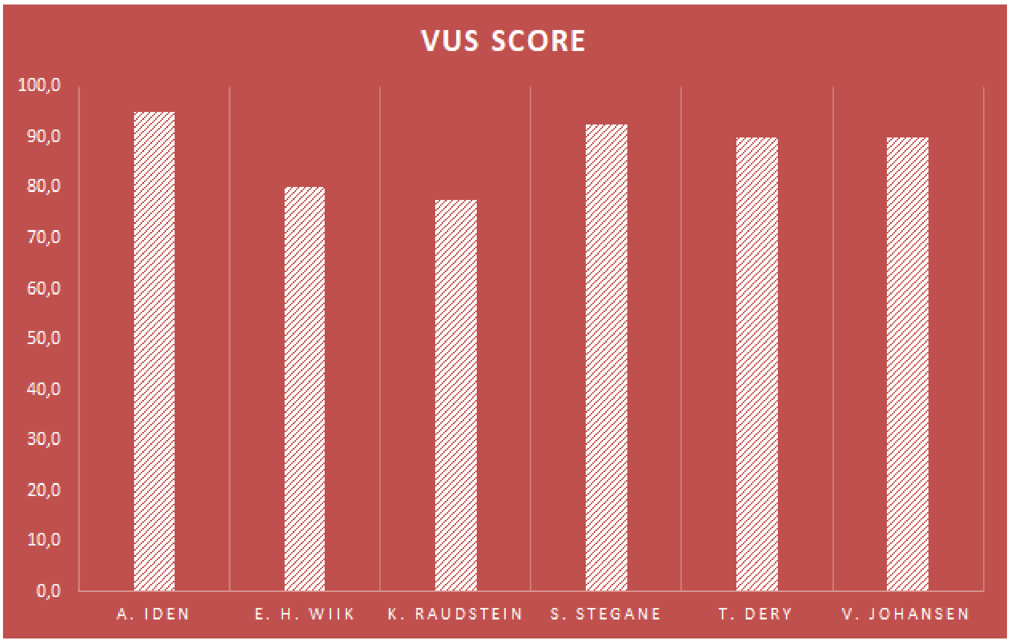


Figure 14: Overview of the VUS-scores

The average VUS-score was 87.5(Figure 14. According to A. Bangor et al. described in Section 4.5.3 it can be represented as acceptable in the acceptability range, as a B in the grade scale and as excellent in the adjective ratings.

7.4 Critical friend

Throughout the process I have had critical friends who have had machine learning projects in their research. At the end of iterations they gave feedback on both the methods applied and the finished visualizations. We were a group of five who had regular meetings with our supervisor, Prof. A. Babic, and presented or talked about the process we were in. This gave the master students in cooperation with our supervisor, the ability and opportunity to be involved in each others project and give helpful feedback. This was especially useful during the period where the university was closed during the Corona-pandemic.

Chapter 8

Discussion

This chapter discusses the methodology and methods employed and the resulting design, development process and limitations. This chapter also answers the research questions.

8.1 Methodologies and methods

8.1.1 Design Science

During this research, the methodological approach of design science research was used in all steps of the research project. The guidelines established in Table 2 have been instrumental in working with the project.

Design as an artefact

The artefacts created in this thesis are the Ridge regression model used to extract patterns and produce results and the visualizations that convert the numerical results to more readable and informative graphs and charts as described in Chapter 5. It allows for novel approach to level-of-living surveys with results that are interesting without the use of manual expert labour as discussed in Chapter 7.

Problem relevance

The domain of knowledge of society and the impact of variables have on income are always relevant in the problem space of inequality and the visualization of knowledge. As discussed in Chapter 3 reports from organizations, both public and private, reflect the work of multiple researchers and many hours of statistical labour. The second problem of this problem is to convey this information in interesting and ways that increase public knowledge. The first data was collected to inform decision makers so they could make decisions on making the level-of-living rise as discussed in Chapter 2.

Design evaluation

The design has been evaluated using an experimental controlled environment with usability testing and functional testing. The evaluations completed in Chapter 7 include semi-structured interviews, tasks and a VUS.

Research contribution

The main contribution is the artefact itself which use previous knowledge in a novel way. The novelty lying in the data set used and the visualizations created on the resulting findings discussed in Chapter 9. The addition to the knowledge base can be used for similar projects on other data sets which need exploration for the benefit of the public domain. This thesis is a proof by construction, which is discussed in Chapter 4.

Research rigor

The main artefacts were made using proven machine learning models with extensive research and use behind them. Ridge regression has been used in many predictive tasks and is one of the most popular regression models. The visualizations were based on clear and easy graphs which has been used in research and reported by NIPH among others as mentioned in Chapter 3.

Design as a Search Process

This research started as an assessment of what patterns and information was possible to extract from the level-of-living studies by machine learning methods. The final solutions discussed in Chapter 6 is the results were the result of targeted and iterative goals, a search in the data and methods. As described in Chapter 5, the first step was to check the newest data and find out what was possible with uniform data and then extend to older and more inconsistent data. The visualizations were also produced iteratively with more every iteration giving more intricate and detailed visualizations. The presented solution of the search process therefore represents a best solution based on the guidance of my peers, my supervisor, the individual work of the researcher and the possibilities in the data.

Communication of Research

For the academic audience, this contribution will be published to the University of Bergen publishing service, www.bora.uib.no. Excluding the results, most of this thesis is of most use for strictly academic readers. The resulting visualizations are meant to be understood by the general public and evaluation participants were not domain experts and on this represented the general public.

8.1.2 The three cycle view

As a part of the design science research the three cycles were also adhered to. It gave the project another helpful way of framing the work and giving ways to understand the value of the different stages of work in design science. The different cycles were used in different stages. The relevance cycle was used to establish the requirements and understanding what was possible to achieve in Chapter 5. This cycle was also used to test it on the application domain, in this case ordinary people, to check if it improved the environment. The rigor cycle was applied when doing literature review in Chapter 3 and finding the right methodologies to use. The contribution the knowledge base is this master thesis, and the novel use of the level-of-living surveys. The internal design cycle are described in the chapters for development, result and evaluation (Chapters 5, 6 and 7). This is the cycle which demanded the most time and effort and was most important for this project.

8.2 Knowledge discovery in Databases

The KDD-process was used to guide and understand the necessary steps during the development and the process from digitized data set to finished visualizations. It was helpful for framing the work needed, the problems encountered, and solutions needed during development. It also gave great clarity to when each specific goal needed to be achieved. It guided all iterations described in Chapter 5. The KDD methodology made it easy to write the backlog in Trello, the Kanban system used for this project mentioned in tools used in Chapter 5.

8.2.1 Data

While NSD provided data that was digitized from the old paper sources, it had not been made homogeneous or was not in its raw comparable across the decades. The data was csv-files derived from the level-of-living surveys, curated by NSD.

8.2.2 Selection

So the selection phase came after the NSD preprocessing. For every iteration selection was done again based on what new features were still usable as the data set was extended back to older surveys. Multiple times during the development data features were added or removed.

8.2.3 Pre-processing

Pre-processing was done for each data set separately as features had different variable names and slightly different formats throughout the years. Outliers, noise, and input which were wrongly inputted, such as values that were nonsensical or missing, were removed.

8.2.4 Transformation

The transformation was the feature engineering during development in Chapter 5. This includes the work done on features to make them more streamlined and comparable. Finding the most useful variables in the data sets and binary encoding the categorical variables and normalizing the numerical.

8.2.5 Data Mining

The data mining process was performed with discovery in mind as described in chapter 4 under KDD-process. The data mining task that was decided on was predictive which includes regression. The predictions were performed using Ridge regression which had the best combination of easy to understand patterns in its coefficients and with an acceptable precision which were tested in the first iteration. The Ridge regression was able to extract patterns as described in Chapter 5. The patterns where judged on what the researcher in cooperation with the supervisor and the master group found most interesting or wanted to know more about. Patterns were also chosen where they contradicted what I believed to be commonly held beliefs, inspired by methods for interestingness described in Chapter 4.

8.2.6 Interpretation/Evaluation

Evaluation was done by the methods mentioned in Chapter 7. They gave good results on both interviews, tasks performed and the VUS-score. Generally people found the results interesting, was able to extract information from the visualizations and found the design pleasing.

8.3 Limitations

Limitations for this project are mainly the limitations of the data. The earliest data set has selection bias for political reasons, with special attention given to vulnerable groups in society. Only the newer data sets from 1995 on are based on a representative selection from the general public. This creates some skewed results in the finished visualizations. Another limitation is that the visualization picked out for evaluation was chosen by a non-expert. An expert could have a better perspective on what features are interesting in the testing of these artefacts. In addition, all participants in the evaluation have an academic background which is not representative of the general public. A more extensive selection of participants would make the conclusions from the evaluation more conclusive.

8.4 Answering Research Questions

This section answers the research questions stated in Chapter 1, Section 1.1that defined the main goals of the research. machine learning is not a usual way of approach the level-of-living surveys in Norway even though there is a wealth of data collected over decades. The usual approach is applying statistics and expert domain knowledge and presenting results to the public in well established forms, such as reports or journals. In this research machine learning explored the feasibility of applying machine learning methods and how well the results could be understood by potential users.

RQ1: Can machine learning models be used to mine information from the Norwegian level-of-living surveys?

This question was tested using different machine learning models, examining their output and precision. The data was explored for different feature selections to mine through the data as extensive as possible (Chapter 5 and Chapter 4). The resulting coefficient values revealed patterns in the data set. Examples include inequality in education, health, disability and marital status (Chapter 6, Chapter 7). Such results encourage further exploration to answer questions that are relevant for different users or information seeker. Clearly there is a potential to apply methods using open source software (Section 5.3.1) and the visualisation tools described in (Section 5.3.1). Data mining included a significant preprocessing step since there were different routines for data collection in the time span of the surveys which extended for almost 50 years (Section 5.3.2). Thus, the data cleaning is an important part of data mining for any future developer.

RQ2: Can visualizations present the information from the level-of-living studies in interesting ways?

To answer this question two lines of enquiry was asked in the evaluation. The first questions was based on Silberschatz and Tuzhilin definition discussed in Section 4.1.2. The second question was open-ended and allowed the users to express their opinion, interpretation, and possible critiques regarding the mined information and its visualisation. The surprise from the interview was that while most of the participants found the visualizations interesting, it was not based on any contradiction to their established beliefs. They found it interesting to see a new perspective on information they already had, they enjoyed seeing the impact and value of the factors and a more tangible view of the inequalities. To see the actual graphs and bars gave a certainty to their conviction and established belief.

Chapter 9

Conclusion and Future Work

9.1 Conclusion

This project has utilized the Design Science research framework together with a Knowledge Discovery in Databases process. This ensured the quality of research and that the resulting thesis presents a tangible, novel, meaningful contribution to the knowledge base, in this case being the level-of-living data in Norway. The main contribution of this research project is the use of data mining and visualizations to discover and present data from level-of-living surveys that previously have only been performed by domain experts. The data was provided by NSD and was essential for this project. All privacy concerns were handled by NSD, and any new data collected during the evaluation was approved by NSD. The combination of design science and KDD was crucial and very helpful in the work. This combination of methods and type of information retrieval can hopefully be used to further explore the level-of-living surveys or other similar public data. The requirements of this project were created iteratively and with input from critical friends and the supervisor. All visualizations were based on these requirements and created with the purpose of communicating discovered knowledge, i.e. mined patterns. The visualizations were tested on with mostly IT-experts and one person with no IT-background. The general evaluation had a positive result and all participants found some part of the visualizations interesting and stirred a discussion towards the interpretation of the acquired knowledge.

9.2 Future Work

Future work should be the inclusion of more features from the available data set. The models derived for this project are usable for many other level-of- living variables comparable to the ones in the research data set and even beyond as long as the proper data preprocessing steps are performed. The main work would be to create an application where the user can filter out variables and retrieve results that would be auto-generated to the optimal visualization for the patterns discovered. The user would then have access to all variables and could easily select the level-of-living variables they find interesting and visualize patterns they find valuable. The users can pick their target label, for example, switching out cumulative income with other variables as a basis for further research into the level-of-living survey data. For example, mining for patterns for individuals with a disability or having primary school as the highest completed level of education. All this is possible by fine-tuning models presented in this thesis and having a user-friendly interface to support users in their independent exploration of data.

Bibliography

[1] A. Bangor, P. Kortum, and J. Miller. Determining what individual sus scores mean: Adding an adjective rating scale. *J. Usability Studies*, 4(3):114123, May 2009. ISSN 1931-3357. 4.5.3

[2] N. Barrabés and G. K. Østli. Classification of education (nus), 2020. URL https://www.ssb.no/klass/klassifikasjoner/36/. 5.3.2

[3] J. Brooke. Sus: A quick and dirty usability scale, 1996. 4.5.3

[4] F. Chollet. *Deep Learning with Python*. Manning, Nov. 2017. ISBN 9781617294433. (document), 6

[5] A. Costa and B. Kallick. Through the lens of a critical friend. *Educational Leadership*, 51, 01 1993. 4.5.4

[6] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11): 27–34, nov 1996. ISSN 00010782. doi: 10.1145/240455.240464. URL https://en.wikipedia.org/wiki/Integrated{_}development{_}environment{%}0Ahttps://en.wikipedia.org/wiki/Waterfall{_}model{%}0Ahttps://en.wikipedia.org/wiki/Minkowski{_}distance{%}0Ahttps://en.wikipedia.org/wiki/List{_}of{_}web{_}browsers{%}0Ahttp://adultmeducation.com/downloads. (document), 5

[7] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996. 4.1.2

[8] S. Few. Data visualization for human perception. *The Encyclopedia of Human-Computer Interaction, 2nd Ed.*, 2013. 3.1

[9] N. C. for Research Data & Statistics Norway. Microdata.no. https://microdata.no/, 2020. Accessed: 2020-05-30. 3.2

[10] O. H. Hansen. *Standard for utdanningsgruppering i offentlig norsk statistikk = Norwegian standard classification of education*. Statistik Sentralbyra, 1973. 5.3.2

[11] A. Hevner, A. R, S. March, S. T, Park, J. Park, Ram, and Sudha. Design science in information systems research. *Management Information Systems Quarterly*, 28:75–, 03 2004. (document), 4.1, 4

[12] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55. 5.3.1

[13] J. Iivari. "A Paradigmatic Analysis of Information Systems As a Design Science. *Scandinavian Journal of Information Systems*, 19(2):39–64, 2007. URL https://aisel.aisnet.org/sjis/vol19/iss2/5. 4.1.1

[14] T. Jefferson. The papers of thomas jefferson, 1787. URL https://founders.archives.gov/documents/Jefferson/01-12-02-0490. (document)

[15] S. Johansson, R. Erikson, J. O. Jonsson, and M. Tåhlin. Swedish level of living survey, 1999. URL https://snd.gu.se/en/catalogue/study/ext0007. 1

[16] M. Klemettinen, H. Mannila, P. Moen, H. Toivonen, and A. Verkamo. Finding interesting rules from large sets of discovered association rules. *Proceedings of the Third International Conference on Information and Knowledge Management*, 02 1995. doi: 10.1145/191246.191314. 4.1.2

[17] H. Kniberg. *Kanban and Scrum - Making the Most of Both*. Lulu.com, 2010. ISBN 0557138329. 5.3.1

[18] L. Lefsrud. Book review: Hans rosling, ola rosling, and anna rosling rönnlund factfulness: Ten reasons were wrong about the world and why things are better than you think. *Organization Studies*, 40(7):1093–1096, 2019. doi: 10.1177/0170840618813918. URL https://doi.org/10.1177/0170840618813918. 1

[19] R. Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 22 140: 55–55, 1932. 4.5.3

[20] W. McKinney. Data structures for statistical computing in python. In S. van der Walt and J. Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51 – 56, 2010. 5.3.1

[21] T. M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997. ISBN 978-0-07-042807-2. 2.1

[22] M. Moshagen and M. Thielsch. A short version of the visual aesthetics of websites inventory. *Behaviour & Information Technology - Behaviour & IT*, 32:1305–1311, 12 2013. doi: 10.1080/0144929X.2012.694910. 4.5.3

[23] M. Moshagen and M. T. Thielsch. Facets of visual aesthetics. *Int. J. Hum.-Comput. Stud.*, 68(10):689709, Oct. 2010. ISSN 1071-5819. doi: 10.1016/j.ijhcs.2010.05.006. URL https://doi.org/10.1016/j.ijhcs.2010.05.006. 4.5.3

[24] A. Müller and S. Guido. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media, 2016. ISBN 9781449369897. URL https://books.google.no/books?id=vbQlDQAAQBAJ. (document), 2.1, 7

[25] L. Mæhlum, E. Pontoppidan, T. Moen, and T. Thune. Norsk utdanningshistorie, 2020. URL https://snl.no/Norsk_utdanningshistorie. 5.3.2

[26] J. Nielsen. 10 usability heuristics for user interface design, 1995. URL https://www.nngroup.com/articles/ten-usability-heuristics/. 4.5.3

[27] S. Norway. Fattigdomsproblemer, levekårsundersøkelsen, Mar 2020. URL https://www.ssb.no/sosiale-forhold-og-kriminalitet/statistikker/fattigdom/aar. 5.3.4

[28] N. I. of Public Health. Social inequalities in health. *Public Health Report*, Oct 2018. URL https://www.fhi.no/en/op/hin/groups/social-inequalities/. 3.1

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 5.3.1

[30] J. Preece, Y. Rogers, and H. Sharp. *Interaction Design: Beyond Human-Computer Interaction*. Wiley, Hoboken, NJ, 4 edition, 2015. ISBN 978-1-119-02075-2. 5.2

[31] T. Rødseth, G. Hernes, and A. Aase. Norwegian level of living study 1973, 2012. 2.2

[32] H. Rosling, A. Rönnlund, and O. Rosling. *Factfulness: Ten Reasons We're Wrong About the World–and Why Things Are Better Than You Think*. Flatiron Books, 2018. ISBN 9781250123817. URL https://books.google.no/books?id=j-4yDwAAQBAJ. 3.1

[33] H. Rosling, A. Rönnlund, and O. Rosling. *Factfulness: Ten Reasons We're Wrong About the World–and Why Things Are Better Than You Think*. Flatiron Books, 2018. ISBN 9781250123817. URL https://books.google.no/books?id=j-4yDwAAQBAJ. 1

[34] A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.*, 3(3):210229, jul 1959. ISSN 0018-8646. doi: 10.1147/rd.33.0210. URL https://doi.org/10.1147/rd.33.0210. 2.1

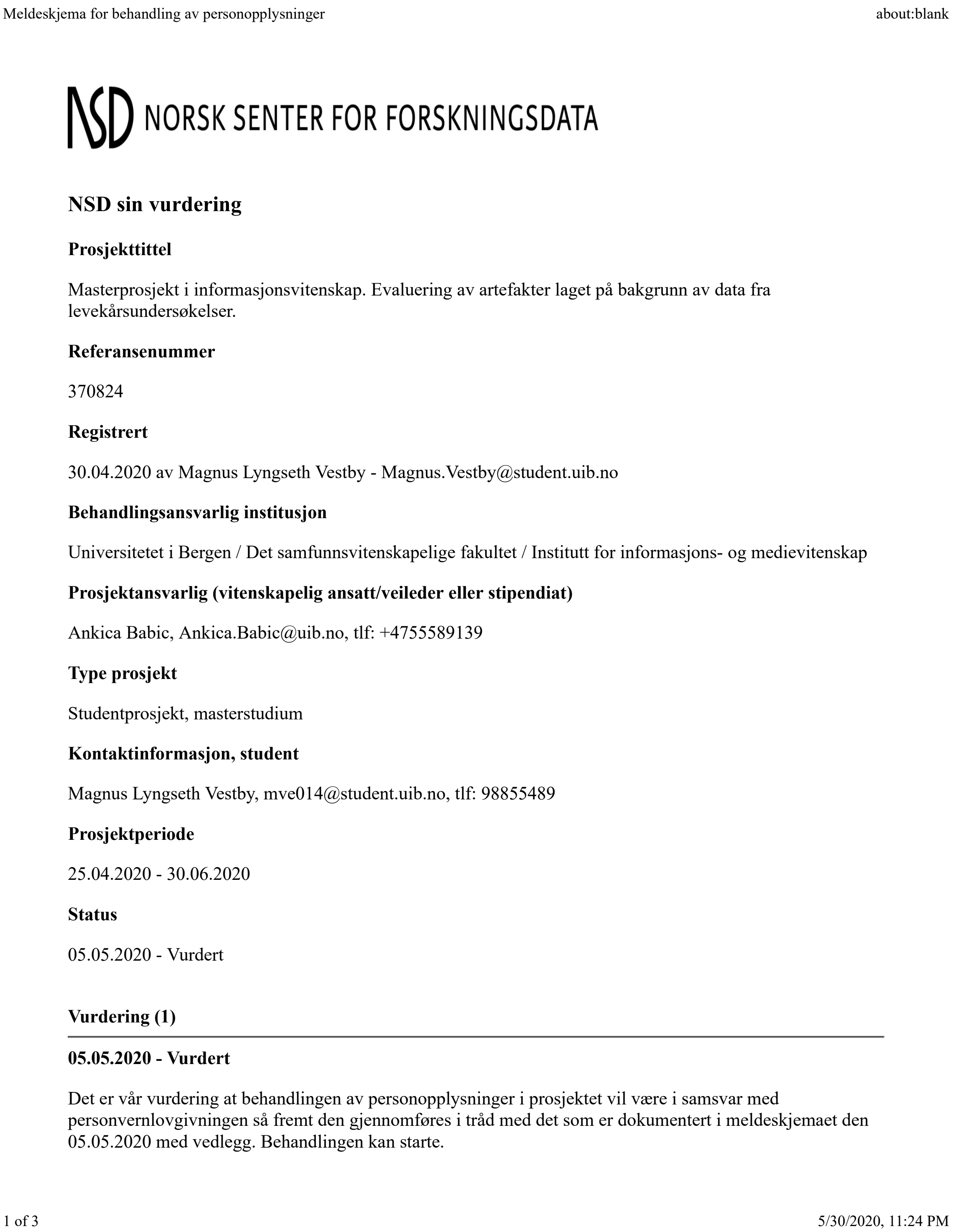
[35] A. Silberschatz and A. Tuzhilin. What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):970–974, 1996. ISSN 10414347. doi: 10.1109/69.553165. 4.1.2

[36] Statistics Norway. Statistisk sentralbyrås historie, 2020. URL https://www.ssb.no/omssb/om-oss/historie. 1

[37] O. Steingrimsdottir, O. Naess, J. Moe, E. Groholt, D. Thelle, and B. Strand. Trends in life expectancy by education in norway 1961-2009. *Eur J Epidemiol*, 27(3):16371, 2012. (document), 1

[38] G. Van Rossum and F. L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697. 5.3.1

A - NSD Approval



NSD NORSK SENTER FOR FORSKNINGSDATA

NSD sin vurdering

Prosjekttittel

Masterprosjekt i informasjonsvitenskap. Evaluering av artefakter laget på bakgrunn av data fra levekårsundersøkelser.

Referansenummer

370824

Registrert

30.04.2020 av Magnus Lyngseth Vestby - Magnus.Vestby@student.uib.no

Behandlingsansvarlig institusjon

Universitetet i Bergen / Det samfunnsvitenskapelige fakultet / Institutt for informasjons- og medievitenskap

Prosjektansvarlig (vitenskapelig ansatt/veileder eller stipendiat)

Ankica Babic, Ankica.Babic@uib.no, tlf: +4755589139

Type prosjekt

Studentprosjekt, masterstudium

Kontaktinformasjon, student

Magnus Lyngseth Vestby, mve014@student.uib.no, tlf: 98855489

Prosjektperiode

25.04.2020 - 30.06.2020

Status

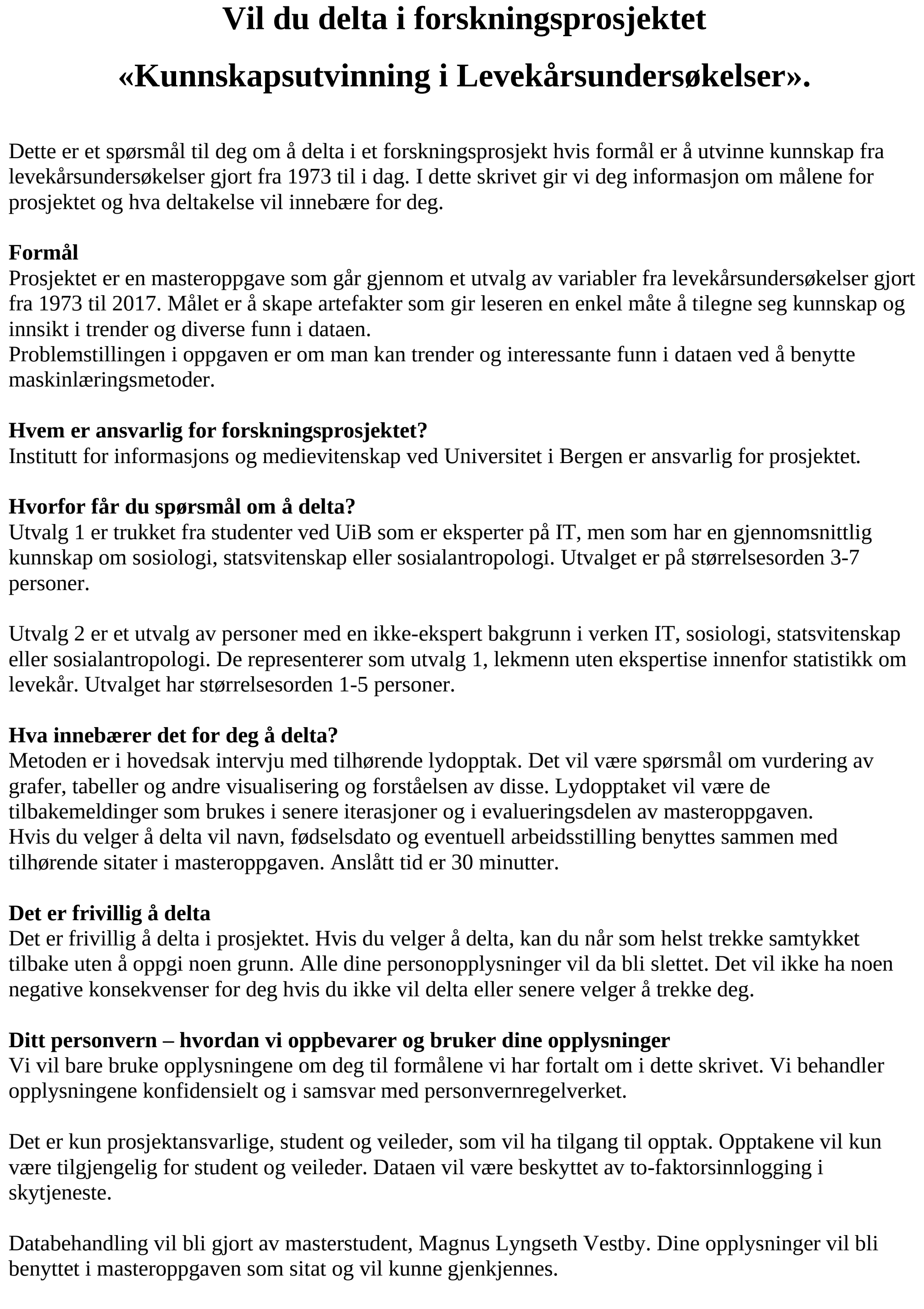
05.05.2020 - Vurdert

Vurdering (1)

05.05.2020 - Vurdert

Det er vår vurdering at behandlingen av personopplysninger i prosjektet vil være i samsvar med personvernlovgivningen så fremt den gjennomføres i tråd med det som er dokumentert i meldeskjemaet den 05.05.2020 med vedlegg. Behandlingen kan starte.

B - Consent form



Vil du delta i forskningsprosjektet «Kunnskapsutvinning i Levekårsundersøkelser».

Dette er et spørsmål til deg om å delta i et forskningsprosjekt hvis formål er å utvinne kunnskap fra levekårsundersøkelser gjort fra 1973 til i dag. I dette skrivet gir vi deg informasjon om målene for prosjektet og hva deltakelse vil innebære for deg.

Formål
Prosjektet er en masteroppgave som går gjennom et utvalg av variabler fra levekårsundersøkelser gjort fra 1973 til 2017. Målet er å skape artefakter som gir leseren en enkel måte å tilegne seg kunnskap og innsikt i trender og diverse funn i dataen.
Problemstillingen i oppgaven er om man kan trender og interessante funn i dataen ved å benytte maskinlæringsmetoder.

Hvem er ansvarlig for forskningsprosjektet?
Institutt for informasjons og medievitenskap ved Universitet i Bergen er ansvarlig for prosjektet.

Hvorfor får du spørsmål om å delta?
Utvalg 1 er trukket fra studenter ved UiB som er eksperter på IT, men som har en gjennomsnittlig kunnskap om sosiologi, statsvitenskap eller sosialantropologi. Utvalget er på størrelsesorden 3-7 personer.

Utvalg 2 er et utvalg av personer med en ikke-ekspert bakgrunn i verken IT, sosiologi, statsvitenskap eller sosialantropologi. De representerer som utvalg 1, lekmenn uten ekspertise innenfor statistikk om levekår. Utvalget har størrelsesorden 1-5 personer.

Hva innebærer det for deg å delta?
Metoden er i hovedsak intervju med tilhørende lydopptak. Det vil være spørsmål om vurdering av grafer, tabeller og andre visualisering og forståelsen av disse. Lydopptaket vil være de tilbakemeldinger som brukes i senere iterasjoner og i evalueringsdelen av masteroppgaven.
Hvis du velger å delta vil navn, fødselsdato og eventuell arbeidsstilling benyttes sammen med tilhørende sitater i masteroppgaven. Anslått tid er 30 minutter.

Det er frivillig å delta
Det er frivillig å delta i prosjektet. Hvis du velger å delta, kan du når som helst trekke samtykket tilbake uten å oppgi noen grunn. Alle dine personopplysninger vil da bli slettet. Det vil ikke ha noen negative konsekvenser for deg hvis du ikke vil delta eller senere velger å trekke deg.

Ditt personvern – hvordan vi oppbevarer og bruker dine opplysninger
Vi vil bare bruke opplysningene om deg til formålene vi har fortalt om i dette skrivet. Vi behandler opplysningene konfidensielt og i samsvar med personvernregelverket.

Det er kun prosjektansvarlige, student og veileder, som vil ha tilgang til opptak. Opptakene vil kun være tilgjengelig for student og veileder. Dataen vil være beskyttet av to-faktorsinnlogging i skytjeneste.

Databehandling vil bli gjort av masterstudent, Magnus Lyngseth Vestby. Dine opplysninger vil bli benyttet i masteroppgaven som sitat og vil kunne gjenkjennes.

Hva skjer med opplysningene dine når vi avslutter forskningsprosjektet?
Opplysningene i oppgaven vil bli stående. Alle opptak og personopplysninger lagret på skytjeneste vil bli slettet ved avsluttet prosjekt, anslått avslutning er 15. Juni.

Dine rettigheter
Så lenge du kan identifiseres i datamaterialet, har du rett til:

- innsyn i hvilke personopplysninger som er registrert om deg, og å få utlevert en kopi av opplysningene,
- å få rettet personopplysninger om deg,
- å få slettet personopplysninger om deg, og
- å sende klage til Datatilsynet om behandlingen av dine personopplysninger.

Hva gir oss rett til å behandle personopplysninger om deg?
Vi behandler opplysninger om deg basert på ditt samtykke.

På oppdrag fra Universitet i Bergen har NSD – Norsk senter for forskningsdata AS vurdert at behandlingen av personopplysninger i dette prosjektet er i samsvar med personvernregelverket.

Hvor kan jeg finne ut mer?
Hvis du har spørsmål til studien, eller ønsker å benytte deg av dine rettigheter, ta kontakt med:

- Institutt for informasjon- og medievitenskap, Universitet i Bergen ved masterstudent Magnus Lyngseth Vestby, epost mve014@student.uib.no og førsteamanuensis Ankica Babic, epost Ankica.Babic@uib.no.

Hvis du har spørsmål knyttet til NSD sin vurdering av prosjektet, kan du ta kontakt med:

- NSD – Norsk senter for forskningsdata AS på epost (personverntjenester@nsd.no) eller på telefon: 55 58 21 17.

Med vennlig hilsen

Ankica Babic   Magnus Lyngseth Vestby

(Forsker/veileder)

Samtykkeerklæring

Jeg har mottatt og forstått informasjon om prosjektet *[sett inn tittel]*, og har fått anledning til å stille spørsmål. Jeg samtykker til:

□ å delta i intervju
□ at opplysninger om meg publiseres slik at jeg kan gjenkjennes ved sitat

Jeg samtykker til at mine opplysninger behandles frem til prosjektet er avsluttet

----
(Signert av prosjektdeltaker, dato)