University of Bergen

Department of Information Science and Media Studies

Masters Thesis

HALE, the Hip Arthroplasty Longevity Estimation system

Author: Per-Niklas Longberg

Supervisor: Prof. Ankica Babic

April 10, 2019

hale^1 *(Adjective) - (of an old person) Strong and healthy.*
"He's only just sixty, very hale and hearty."

Oxford Dictionaries

Abstract

This master thesis presents a Design Science research in which the *HALE* system for total hip arthroplasty prosthesis longevity estimation has been developed. The *HALE* system was developed to explore the use of machine learning techniques on a biomedical dataset motivated by two user groups' needs - biomedical engineers who analyze explanted hip arthroplasty prostheses and physicians who work with patients and want to know what the safe and optimal treatment for each patient is.

The dataset mainly contains biochemical measurements and has a limited number of patient data (demographics). The machine learning techniques are seen as a possibility to quickly and promptly analyze the data and answer questions about specific cases as well as the patient group as a whole.

The machine learning components rely on regression analysis for prediction and estimating the outcome of single patient cases, as well as the group. Two methods were implemented - multiple linear regression and an optimized C&RT decision tree. At this point in development users found multiple linear regression more appealing for its transparency and better performance in comparison to the regression based decision tree counterpart. In the future C&RT trees can be considered as an alternative when the users have more experience and trust rely on the system. The machine learning methods used in the *HALE* system were validated against a comparative linear regression statistical procedure of IBMs SPSS software, resulting in a comparable accuracy, performance and similarly constructed regression model.

User evaluation has shown that the *HALE* system was manageable and appealing to the user groups. The largest current practical limitation is the size of the dataset, however by expanding this dataset and adding new clinical variables it will be easy to improve the performance of the regression models. It is also expected that additional functionality such as discriminant and clustering analysis would be feasible to implement. Thus, the machine learning components of the *HALE* system, as implemented using scikit-learn, have proven to be suitable and easy to utilize even for novice developers.

Acknowledgment

First of all I would like to express my gratitude to my supervisor, Ankica Babic, whose involvement, patience, wisdom, motivation and undeniable positivity were vital to the completion of this project. I would also like extend my gratitude to Doctor Peter Ellison whose assistance, input and involvement in both development and evaluation which proved absolutely invaluable.

My sincere thanks goes also to Doctor Paul Johan Høl for his generous time and imperative feedback and chief physician Professor Ove Furnes who took his valuable time to give me a clinical perspective on work with the arthroplasty implants and introduction to the National Arthroplasty research.

I must also thank my fellow students and friends at 642 Big D-ata Boys for their support, help and motivation throughout the project. There was definitely a lot of action, and we were indeed true survivors.

Per-Niklas Longberg.

Contents

Abstract **ii**

Acknowledgment **v**

1 Introduction **1**

1.1 Motivation 1

1.2 Research Questions 2

1.3 Thesis outline 2

2 Theory **5**

2.1 Related work 5

2.1.1 Use of Machine Learning Theory to Predict the Need for Femoral Nerve Block Following ACL Repair 5

2.1.2 Predicting and Analyzing Osteoarthritis Patient Outcomes with Machine Learning 6

2.2 Knowledge Discovery in Databases 6

2.3 Data Mining 7

2.4 Human-Computer Interaction 8

2.5 Machine Learning 8

2.5.1 Machine Learning and Data Mining in Medicine 8

2.5.2 Supervised Learning 9

2.5.3 Decision Trees 9

2.6 Regression Analysis 11

2.6.1 Multiple Linear Regression 11

2.7 Total Hip Arthroplasty 11

2.7.1 Implant Components 12
2.7.2 Adverse Events 12

3 Methods and methodologies 15
3.1 Machine Learning Models 15
3.1.1 Classification and Regression Trees 15
3.1.2 Simple and Multiple Linear Regression 15
3.1.3 Dataset splitting 16
3.1.4 Classification and Regression Trees in scikit-learn 16
3.1.5 Feature Selection 17
3.2 Validation with SPSS 17
3.2.1 Significance of independent variables 17
3.2.2 Machine Learning Model Evaluation Metric 17
3.3 Design Science 18
3.4 Development Methods and Methodologies 21
3.4.1 Personal Extreme Programming 21
3.5 Usability Evaluation Methods 23
3.5.1 Qualitative Data Gathering 23
3.5.1.1 Semi-structured interviews 23
3.5.2 Quantitative Data Gathering 23
3.5.2.1 Heuristic Evaluation 23
3.5.2.2 System Usability Scale 24

4 Establishing Requirements 27
4.1 Functional Requirements 27
4.2 Non-Functional Requirements 28
4.3 Intended User 28

5 Prototype Development 29
5.1 Tools and Technologies Used 29
5.1.1 Front-End Technologies 29
5.1.1.1 JavaScript 29

5.1.1.2 jQuery 30

5.1.1.3 HTML 30

5.1.1.4 CSS 30

5.1.2 Back-End Technologies 30

5.1.2.1 Python 30

5.1.2.2 Flask 31

5.1.2.3 Scikit-learn 31

5.1.3 Development Tools 31

5.1.3.1 PyCharm IDE 31

5.1.3.2 Git and GitHub 31

5.2 Development Process Method 31

5.2.1 First Prototype Iteration 32

5.2.2 Second Prototype Iteration 33

5.2.3 Third Prototype Iteration 34

5.2.3.1 User Workflow and System Design - Targeted Sample Prediction 34

6 Implementation of Regression Models 43

6.1 PARETO Dataset 43

6.1.1 Dataset splitting 45

6.1.2 Predicting Continuous Longevity Values from PARETO Dataset 45

6.2 Results 46

6.2.1 Leave-One-Out Cross Validation 46

6.2.2 Train_test_split 51

6.3 IBM SPSS Validation 52

6.3.1 Regressor Coefficients of Linear Models 52

6.3.2 Statistical Significance of Regressors 53

6.3.3 SPSS Predicted Prosthesis Longevity 54

7 System Evaluation 56

7.1 Approval for Research 56

7.2 Performing the Evaluation 56

7.2.1 Semi-structured Interviews 57
7.2.2 Heuristic Evaluation 57
7.2.3 System Usability Scale 58
7.3 Evaluation Results 59
7.3.1 Semi-structured Interviews 59
7.3.2 Heuristic Evaluation 63
7.3.3 System Usability Scale 66

8 Discussion **67**
8.1 Dataset Restrictions 67
8.2 Methods and Methodologies 67
8.2.1 Design Science 69
8.2.2 Personal Extreme Programming 69
8.2.3 Usability Evaluation 69
8.3 Machine Learning Model Performance 70
8.3.1 Linear Regression Model 70
8.3.1.1 Evaluation Metrics 70
8.3.1.2 Significance of Regressors 70
8.4 State of System 71
8.4.1 Feature Selection 71
8.4.2 Predict Future Cases 71
8.5 Answering the Research Questions 71

9 Conclusions **73**
9.1 Conclusions 73
9.2 Future work 74
9.2.1 Machine Learning, Data Handling and Improvements 74

Bibliography **75**

A Informed Consent for Semi-Structured Interviews. **82**

B Interview Guide for Semi-Structured Interviews. **86**

C System Usability Scale Questionnaire. 90

D NSD Approval for Processing Personal Data. 92

List of Figures

2.1 Knowledge Discovery in Databases Graph . . . 7
2.2 Example Decision Tree . . . 10
2.3 Radiograph Total Hip Arthroplasty Prosthesis . . . 12
2.4 Total Hip Arthroplasty Prosthesis Components . . . 13
2.5 Radiograph imaging of a dislocated femoral head. . . . 14

3.1 PXP Phases . . . 22
3.2 SUS Score Scales . . . 26

5.1 Iteration 1: Main Menu, Results . . . 32
5.2 Iteration 2: Main Menu . . . 33
5.3 Iteration 2: Target Results . . . 34
5.4 Iteration 2: Training Results . . . 35
5.5 HALE General Design . . . 36
5.6 HALE Workflow Abstraction . . . 36
5.7 Iteration 3: Button Design . . . 37
5.8 HALE Start Page . . . 38
5.9 HALE Input Fields . . . 39
5.10 Patient Information Display . . . 40
5.11 Available Regressors . . . 41
5.12 Longevity Estimation Result . . . 41
5.13 Longevity Estimation Background . . . 42

6.1 SPSS Model . . . 52

List of Tables

3.1 Design Science Guidelines 19
3.2 Nielsens 10 Heuristics. 25

6.1 DT LOOCV Statistics 47
6.2 DT LOOCV Control Subset 47
6.3 DT LOOCV Full Dataset 48
6.4 MLP LOOCV Statistics 49
6.5 MLP LOOCV Control Subset 49
6.6 MLP LOOCV Full Dataset 50
6.7 Manual MLP LOOCV Full Dataset 51
6.8 Scikit-learn Coefficients 52
6.9 SPSS Coefficients 53
6.10 Scikit-learn P-Values 53
6.11 SPSS P-Values 54
6.12 SPSS Full Dataset 55

7.1 Participants: Interview 57
7.2 Participants: Heuristic 58
7.3 Participants: SUS 58
7.4 SUS Task Completion Time 59
7.5 SUS Calculated Scores 66

Chapter 1

Introduction

This thesis wants to solve a problem for an increasingly wide population. The goal is to use information technology to understand and explain what leads to the benefit of people who need total hip arthroplasty. The elderly population is growing in both size and total population percentage (Carone and Costello, 2006). As the elderly population grows the number of performed hip replacement surgeries have increased over the years, with a projected figure of over 500 000 surgeries by 2030 in the United States alone (Kurtz et al., 2007). The prostheses implanted into patients undergoing this surgery have finite lifespans due to a number of reasons including aseptic loosening, infection, instability, component failure, implant fracture and pain. Revision surgeries are expensive and can cause further complications in patients (Ulrich et al., 2008), thus being undesirable for both the healthcare system and the patient. At the current time there is no sure way for physicians at Haukeland University Hospital responsible for patients who require hip arthroplasty surgery to accurately assess how long a prosthesis will be able to last in a patient, they can only refer to statistics and statistical charts gathered throughout the years detailing the implant longevity rate in previous patients.

1.1 Motivation

The rationale for conducting the research carried out in this project is two-fold. Part of the rationale was to explore the field of machine learning and how it can be applied to the field of medicinal informatics in which vast amounts of data can be found (Obermeyer and Emanuel, 2016). The other part of the rationale was to bridge these machine learning techniques to an actual user interface where users can utilize these techniques without being required to be experienced in neither statistics nor usage of statistical software packages available at the time. There has been conducted a lot of research in the field of applying machine learning techniques on biomedical databases to determine which models produce the best results as shown in Section 2.1. These are often in a single specific use-case, but as of this writing none of the research found attempted to bring the resulting models to an end user group.

1.2 Research Questions

The research that was carried out in this master thesis project has attempted to answer the following questions:

1. Is it possible to develop a highly usable longevity prediction module of hip arthroplasty implants based on a biomedical dataset?
2. Can this module produce reliable predictions that are equivalent to the one produced by a well-known, validated statistical module?
3. Are there any guidelines regarding machine learning that could be suggested to software developers that use scikit-learn, an open-source machine learning framework?

1.3 Thesis outline

The following section contains the general outline of this master thesis, excluding this chapter.

Chapter 2: Theory that presents the theoretical groundwork related to this project and expands upon those theories.

Chapter 3: Methods and Methodologies presents the underlying methodology that this research has been based on, as well as details on the methods relating to the development and evaluation of the prototype produced by this master project.

Chapter 4: Establishing Requirements detail the set of requirements that the *HALE* system was based on during its development.

Chapter 5: Prototype Development presents the *HALE* system and its development iterations.

Chapter 6: Implementation of Regression Models describes how the machine learning methods were implemented and the resulting comparative testing and validation with IBMs SPSS as a statistical system.

Chapter 7: System Evaluation presents how the system usability evaluation was conducted and the resulting feedback from the various participants.

Chapter 8: Discussion contains discussion on the research conducted in this project, the methods and methodologies used, the results from testing and validating machine learning models and the evaluation results. The research questions are answered here.

Chapter 9: Conclusions and Recommendations for Future Work concludes the findings of this research and gives recommendations for how to further develop the artifact produced by this research.

Chapter 2

Theory

2.1 Related work

There is ample work being done and research being conducted in the conjoined fields of medicine and machine learning (Faggella, 2018). There is arguably less work being done regarding development of systems utilizing these machine learning techniques in a system designed to be used by physicians, a system that focuses on delivering high usability as well as accurate predictions.

2.1.1 Use of Machine Learning Theory to Predict the Need for Femoral Nerve Block Following ACL Repair

In their publication in the journal *Pain Medicine* a group of researchers explored using machine learning techniques to predict whether or not a patient would require a femoral nerve block after undergoing anterior cruciate ligament repair, a surgery aimed at reconstructing this knee ligament after tearing (Tighe et al., 2011).

In their work the researchers applied a set of machine learning models on a dataset containing 349 patient samples, among the models used were logistic regression for classification, BayesNet, multi-layer perceptron, support vector machines and alternating decision trees. Predictions were performed for each machine learning model and their perfomance were compared to discover the most reliable model. The research carried out in this publication bears similarities to this master thesis project in terms of model evaluation, yet no end-user was considered in their work as their only concern is model performance comparison and whether they are suited for their exact intended use.

2.1.2 Predicting and Analyzing Osteoarthritis Patient Outcomes with Machine Learning

Two master of science students from Lund university based their research upon developing and applying machine learning techniques on patients afflicted by osteoarthritis. Their goal was to discover whether it was possible to predict patient outcomes using various machine learning techniques, as well as discovering which factors contribute to the patient outcomes (Persson and Rietz, 2017).

Among the machine learning models used their research were logistic regression for classification, ensembles of decision trees in random forests, adaptive boosting and gradient boosting, as well as the neural net model multi-layer perceptron. The research carried out in the Lund master thesis project is heavily focused on developing and evaluating the performance and application of machine learning techniques on a larger dataset. The research conducted bears similarities with this project in the fact that several machine learning techniques were explored and evaluated, however the Lund paper delves deeper into the performance of a larger number of models and does not concern any end-users in any way.

2.2 Knowledge Discovery in Databases

As technology advances, data collection methods and storage capacities have increased. More ways to collect and store more data means that processing the increasingly vast amounts of data in search of useful information is practically impossible to accomplish "by hand" (Chakrabarti et al., 2006; Fayyad et al., 1996). The field of Knowledge Discovery in Databases (KDD) refers to a collection of tools, methods and processes used to enable extracting knowledge and useful information from these growing sets of data. KDD is defined by Fayyad, Piatetsky-Shapiro and Smyth as *"the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data."* (Fayyad et al., 1996). Knowledge Discovery in Databases uses intersecting methods from fields such as machine learning, data mining, databases, artificial intelligence, statistics, data vizualisation and so forth.

Data mining, while being a field described as *the science of extracting useful knowledge from such huge data repositories* (Chakrabarti et al., 2006) and defined by Encyclopædia Britannica as *"Data mining, also called knowledge discovery in databases, in computer science, the process of discovering interesting and useful patterns and relationships in large volumes of data."* (Clifton, 2010). Data mining is a step in the process of KDD. While the aforementioned descriptions of data mining might overlap with the definition of KDD and in some cases being called synonymous with each other, the latter encompasses a bigger picture by incorporating a larger methodological framework for its process with more detail such as data selection, preparation and cleaning, the incorporation of appropriate prior knowledge and the proper interpretation of the data mining results (Fayyad et al., 1996). These additional steps are taken to ensure that the knowledge gleaned from the entire process is useful as there is a risk of finding invalid and/or meaningless patterns if data mining is applied without consideration. All steps of Knowledge Discovery in Databases are depicted in Figure 2.1.

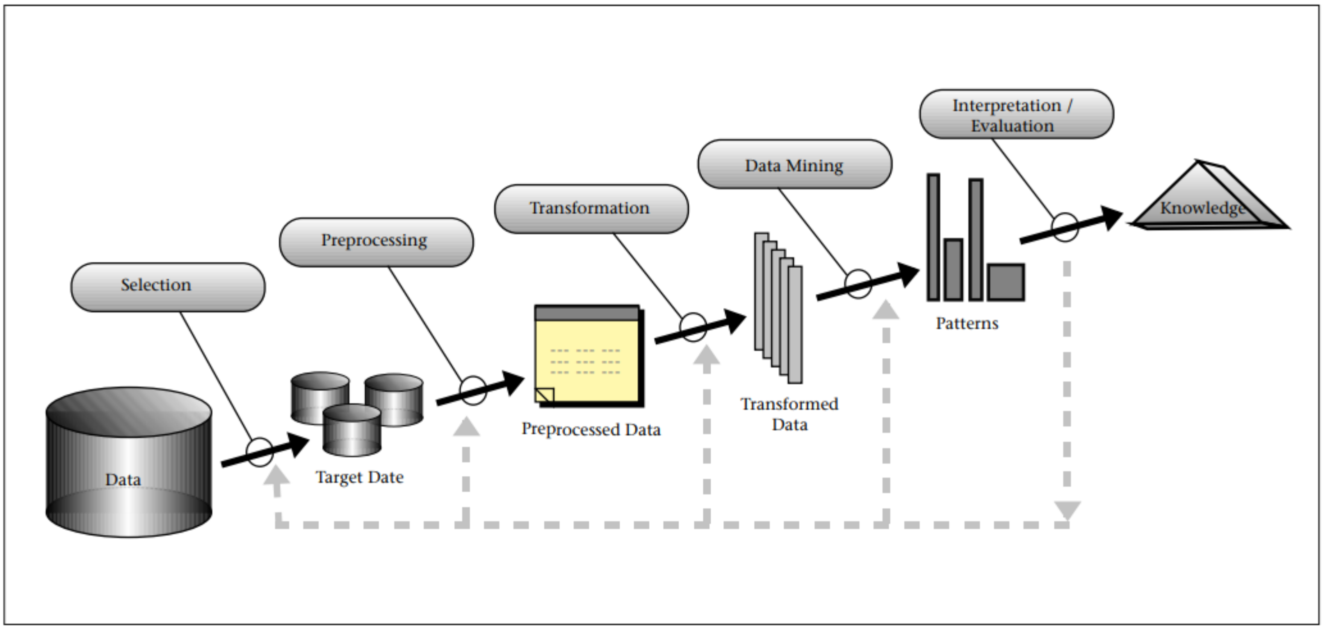


Figure 2.1: Graphical representation of Knowledge Discovery in Databases and its steps (Han et al., 2011)

2.3 Data Mining

Data mining is the application of methods and algorithms from fields of machine learning, artificial intelligence, database systems and statistics in order to extract patterns from data (Chakrabarti et al., 2006). As computing progresses and the vastness of available data continues to expand (Hilbert and López, 2011), we have long since passed the coining of the term big data. The subfield of computer science known as data mining has been developed as a response to the increasing difficulty of creating information from the amounts of data using the interdisciplinary processes of database systems, statistics and machine learning Chakrabarti et al. (2006). Data mining is considered to be the analysis step of the Knowledge Discovery in Databases process where application of data analysis and discovery algorithms should produce an enumeration of patterns (models) over the data (Fayyad et al., 1996).

These patterns (models) can include cluster analysis, anomaly detection, classification and dependencies (Ma et al., 2008). The data being mined can come from a variety of sources such as the internet, databases or data warehouses and so on (Han et al., 2011). In the process of KDD, the patterns discovered from data mining are subsequently evaluated in order to determine whether they are valid or not. While the model produced by data mining might show a seeming relationship between higher infant mortality rates and higher amount of ice cream sold by grocery stores, these may not actually be related at all - correlation does not imply causation (Aldrich, 1995).

2.4 Human-Computer Interaction

Human-computer interaction is a field that concerns the relationship between a computer system and its users, emphasizing on the interfaces and interactability in this relationship (Preece et al., 2015). This relationship can take on many forms today, such as graphical user interfaces displayed on a computer screen, vibration motors used to alert mobile phone users of notifications and alerts, voice-activated personal assistant systems integrated in speakers and so forth. In their publication the Association for Computing Machinery defines human-computer interaction as *"a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them"* (Hewett et al., 2009, p.5).

2.5 Machine Learning

The term machine learning denotes the subfield of artificial intelligence that enable computer information systems to *learn* through statistical techniques. In his book Machine Learning, Tom Mitchell broadly defines machine learning as *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* (Mitchell, 1997). This means that computer software can without being explicitly programmed gain experience and improve performance on a task by doing that specific task, i.e. *learn by doing*.

2.5.1 Machine Learning and Data Mining in Medicine

Data collection and storage methods are increasingly growing in the field of medicine as well, enabling more data to be gathered on each single patient. The data can be vast in terms of variables, samples or a combination of both. In their paper Obermeyer and Emanuel (2016) discuss how machine learning techniques can be applied to medicinal data to create information and ultimately knowledge, especially in terms of predicting patient outcomes. However, the common theme of *correlation does not imply causation* (Aldrich, 1995) is present in applying machine learning algorithms that predict these patient outcomes - while machine learning methods are data-hungry in the term of requiring large datasets to perform to a satisfactory degree, including more variables can help a model predict more accurately but the variables themselves may not be relevant for a given patient outcome (Obermeyer and Emanuel, 2016).

Nevertheless Obermeyer and Emanuel (2016) believe that correctly applying machine learning techniques to medicinal data will transform the field of medicine in three areas. They predict that prognosis will be dramatically improved as more input variables can be assessed by a machine learning system than the humans currently tallying the scores. They also predict that applying machine learning techniques for image recognition and analysis will displace much of the work currently done by radiologists and anatomical pathologists. Lastly, they predict that machine learning will lead to an improvement in diagnostic accuracy, having algorithms generate diagnoses that would likely be better at suggesting high-value testing and lower the rate of testing overuse.

2.5.2 Supervised Learning

In artificial intelligence and machine learning some distinct types of learning can be found - reinforcement learning, unsupervised learning, supervised learning and semi-supervised learning. Of these four types, supervised learning is relevant for this project. Using supervised learning for the machine learning model means for the model to learn from example input-output pairs, then mapping an input to an output based on the learned examples (Russell and Norvig, 1995). The input-output pairs consist of the data that the machine learning system is given to learn from, split into two parts - one for training the model and another for testing.

2.5.3 Decision Trees

A common method of applying machine learning methods is through the use of decision tree learning (Rokach and Maimon, 2008). The general goal of the decision tree learning approach is predicting the value of a dependent variable by constructing a decision tree using several independent variables. The machine learning method earned its name from the tree-like structure depicted in Figure 2.2.

There are several components and steps that make up a decision tree model. The tree itself is made from nodes and branches, and the steps of splitting, stopping and pruning decide how the branches and nodes are created (Song and Lu, 2015).

- **Nodes** are the decision points in the decision tree. There are three nodes - the root node, the internal nodes and the leaf nodes. The root node is the first node, through which all the data samples passes before being split ut into subdivisions. Internal nodes are like the root node but at lower levels of the decision process. At the end of the decision process the data samples end up in a leaf node which represents the final result of decisions and/or events.
- **Branches** are what ties the nodes together to form the decision tree. The branches

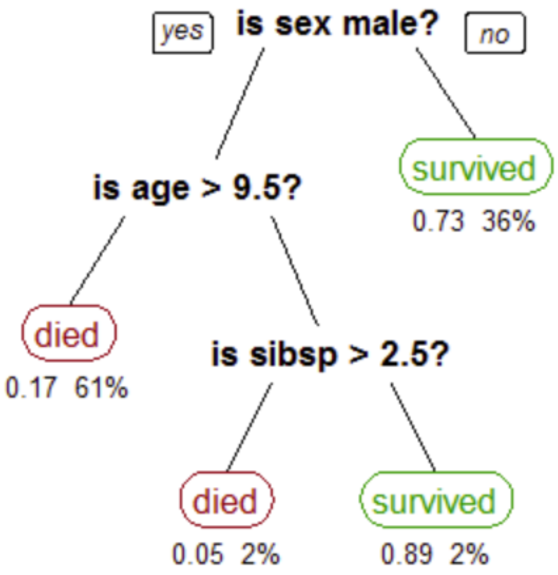


Figure 2.2: An example of a decision tree based on the survivors of the Titanic, in which the leaf nodes show survival probability of a person based on several data features.

represent the decision outcomes from the root node to the leaf nodes, much like if-statements in programming (if x, then go along this branch to the next node - if not x, go to through that branch to another node).

- **Splitting** refers to the process of creating child nodes from the root to the leaf nodes of the tree structure. Independent variables that are related to the dependent variable are used to split a parent node into purer child nodes of the dependent variable. Many methods are used for determining the relation between an independent variable and the dependent variable, such as entropy, information gain or the Gini index (Song and Lu, 2015).

- **Stopping** refers to the set of rules generated to hinder a decision tree from being built too large by demanding that leaf nodes remain as pure as possible. This is to prevent extreme cases of overfitting (Song and Lu, 2015).

- **Pruning** is the process that will be utilized should the aforementioned stopping rules not have an impact that is significant enough to avoid overfitting. It is an alternative method of constructing the tree in which a large tree is initially grown, then its leaf nodes pruned based on whether or not they provide a satisfactory amount of information to the model.

2.6 Regression Analysis

Regression analysis is a method used for estimating the relationships between variables. It is most commonly used to predict or forecast an expected value in a dependent variable, given some independent variables. An illustrative example of regression analysis is predicting house pricing (dependent variable) using the characteristics of a house such as number of bedrooms, total square size, garage or parking spaces and so on (independent variables). The independent variables affect the outcome of the price prediction of a specific house. In this example, the characteristics are the explanatory variables that have a relation to the dependent variable (Chatterjee and Hadi, 2006). The method of regression analysis is one of the most widely employed statistical tools due to its simplistic method of establishing a functional relationship between variables, as well as how extensive its real life subject areas for application are (Chatterjee and Hadi, 2006).

2.6.1 Multiple Linear Regression

Using regression analysis methods to find the relationship between a dependent variable and two or more independent variables is called *multiple regression analysis* (Bremer, 2012). The regression equation for multiple regression is

$$Y = b_0 + b_1 * X_1 + b_n * X_n$$

where the dependent variable Y equals the intercept b_0 plus the regression coefficient b of each independent variable $_n$ times X, the value of each independent variable.

2.7 Total Hip Arthroplasty

Total Hip Arthroplasty, commonly referred to as total hip joint replacement, is the surgical procedure of replacing both the femoral head and implanting a cup in the acetabulum of the pelvis in a patient (Fargon and Fischer, 2015).

Common causes for a patient to require a total hip arthroplasty surgery include arthritis, injury or fracture, or diseases or tumors that can affect bone in joints (Fargon and Fischer, 2015). Arthritis is the most common cause of chronic hip pain, the most common forms of arthritis being rheumatoid and traumatic arthritis, and osteoarthritis. The latter form can often be referred to as *wear and tear* arthritis that can typically occur in individuals passing the age of 50, whose family has a history of arthritis (Fargon and Fischer, 2015). These causes can contribute to reducing a patients ability to accomplish simple everyday tasks that in turn can greatly decrease their quality of life.

According to the American Academy of Orthopaedic Surgeons there is no age restrictions

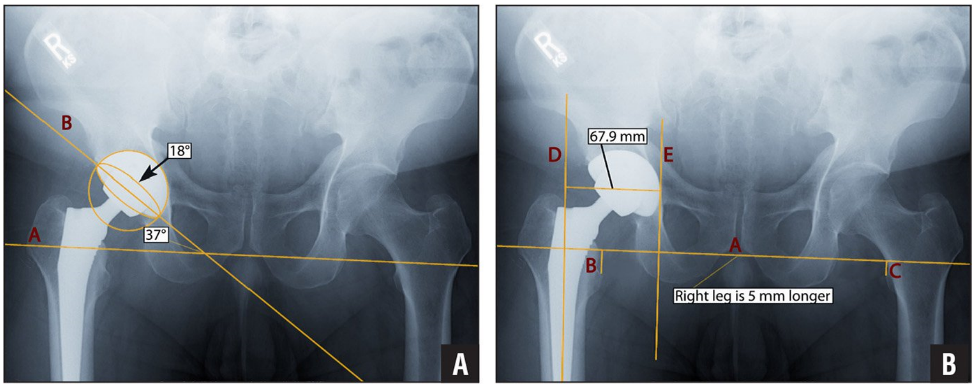


Figure 2.3: Radiograph displaying a patients implant. Left image (A) details the implants acetabular cups X and Y angles. Right image (B) details the acetabular inclination of the patients hips.

on total hip arthroplasty (Fargon and Fischer, 2015). Due to arthritis being a major cause behind total hip arthroplasty and osteoarthritis being one of the most common forms of arthritis, the majority of patients who have undergone this surgery have been between 50 to 80 years of age. In some cases much either younger patients or even older may require this surgery (Fargon and Fischer, 2015). The primary goal of performing hip arthroplasty surgery is relieving pain and increasing or restoring joint mobility so that the patient can return to an unhindered everyday life.

2.7.1 Implant Components

The replacement prosthesis used in Total hip arthroplasty surgeries can be described as a four-part bridge as seen in Figure 2.4. The four parts of the implant are *stem*, *femoral head*, *acetabluar cup lining* and *acetabular cup* (Nieuwenhuijse et al., 2014). The stem is fastened to the patients femoral bone, either through cementing the stem in place or press-fitting, in which the stem has a porous surface allowing for bone ingrowth. The femoral head is either metal or ceramic. Between the femoral head and the acetabular cup is a lining of either plastic, metal or ceramic. This lining allows for smooth motion between the femoral head and the acetabular cup. The acetabular cup is the component fastened in the pelvis of the patient, replacing the natural femoral socket.

2.7.2 Adverse Events

There are several complications that can occur in a patient after undergoing a total hip arthroplasty surgery. These complications are known as *adverse events*, defined as any untoward medical occurrence related to medical management rather than disease, such as all aspects of medical care, diagnosis and treatment (World Health Organization, 2005). A revision surgery is required in the presence of any adverse event, and aims to relieve the prob-

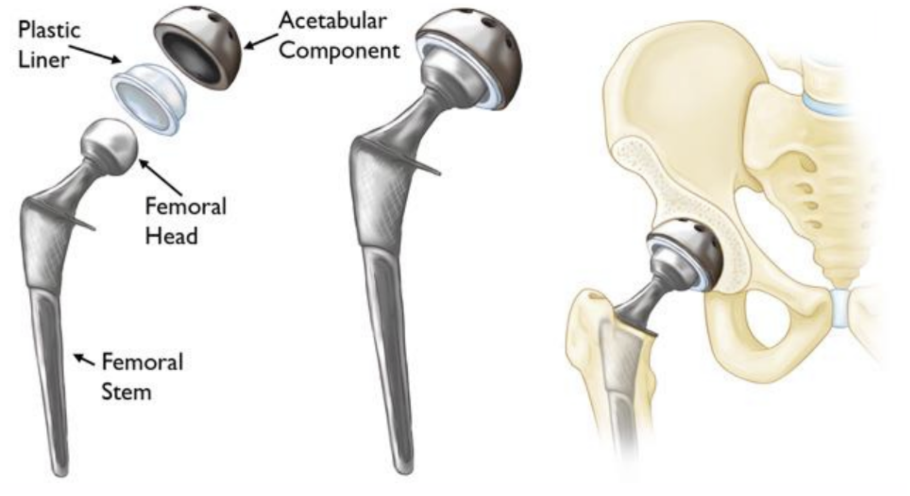


Figure 2.4: A graphical depiction of the components used in total hip arthroplasty surgery before and after insertion.

lems from this event. The surgeries are costly and carry their own set of risks and possible complications, and are therefore an undesired outcome after a total hip arthroplasty surgery for both patient and the healthcare system as a unit. Some examples of which complications can lead to an adverse event is listed below:

- **Blood clots.** One of the most common post-surgery complications, the blood clots can be life-threatening if they are allowed to travel to a patients lungs.
- **Dislocation.** When the tissue around the inserted prosthesis is healing after the surgery, dislocation of the femoral head and acetabular cup can occur.
- **Infection.** Infection is one of the most serious post-operative complications that can lead to revision surgery or in the worst cases to removal of the prosthesis as the infections can spread to the implants.
- **Implant wear and loosening.** Over time the implanted prosthesis will wear down from everyday use. This can cause particles from the materials used in the prostheses to leak into the patients surrounding tissue and bloodstream, triggering osteolysis which can cause bone death around the prostheses.

While there are other complications that physicians need to take into account after a total hip arthroplasty surgery, they were not relevant for this project. The most important complication for this project is also the most common one - implant wear and loosening leading to what is called *aseptic loosening* of the implant.

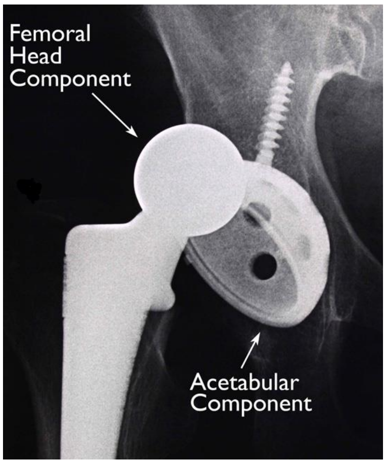


Figure 2.5: Radiograph imaging of a dislocated femoral head.

Aseptic Loosening

The term *aseptic loosening* is used for the adverse effect in which an implanted prosthesis is loosened from the patients bone while no infection is present. Aseptic loosening can be caused by mechanical loss of fixation over time, inadequate initial fixation during surgery or biological loss of fixation due to osteolysis induced by particle debris of the implant itself (Abu-Amer et al., 2007). Aseptic loosening can occur from 10 to 20 years after the primary hip arthroplasty surgery (Abu-Amer et al., 2007). According to the data gathered from the Swedish total hip arthroplasty register the primary reason for patients requiring revision surgery has been periprosthetic osteolysis, this being the cause in over 75% of revision cases (Malchau et al., 2002).

Chapter 3

Methods and methodologies

This chapter details the methods and methodologies that was used in this research project.

3.1 Machine Learning Models

Two different approaches to estimating a continuous dependent variable were used in this project, Decision Tree Regression and Multiple Linear Regression. The Decision Tree model was chosen for its reputation for being easy to interpret and understand yet yield accurate results for regression problems (Seif, 2018), as well as it being a widely adopted method for predictions. Multiple Linear Regression was chosen as a comparative regression model which would be tested against the performance and accuracy of decision tree (). These models were implemented through the use of scikit-learn (Pedregosa et al., 2011).

3.1.1 Classification and Regression Trees

Scikit-learns ***Decision Tree Regression*** (Pedregosa et al., 2011) module was used for this project. This regression model is based on an optimized Classification and Regression Tree (CART), an algorithm that constructs binary decision trees made by the pruning method - using independent variables and thresholds that yield the most information gain at each node (Scikit-learn, b).

3.1.2 Simple and Multiple Linear Regression

Scikit-learns ***Linear Regression*** (Pedregosa et al., 2011) model was used for conducting both simple and multiple linear regression. This module from the machine learning framework creates a predictor object using Ordinary Least Squares Linear Regression that automatically adopts either simple or multiple forms of regression based on the passed regressors when using the model (Scikit-learn, a).

3.1.3 Dataset splitting

Conducting machine learning techniques on any given data is dependent on splitting that given dataset into two parts - a **training** subset and a **testing** subset (Pedregosa et al., 2011). The machine learning component uses the training data to learn and thus generating a model based on the training set. After training the model will use the testing set for validation by performing the desired prediction techniques on a subset of the testing set stripped of the actual values the machine learning model is trying to predict (Reitermanová, 2010). The result is a comparative set of data which can be used to measure the model's predictive accuracy, goodness of fit and other metrics.

Scikit-learns train_test_split

Scikit-learn, the framework for machine learning, contains a subpackage for model selection that can split arrays and matrices into training and testing subsets (Scikit-learn, e). The train_test_split function takes a set of data that is either a list, a numpy array, scipy-sparse matrices or pandas dataframes. Based on the *random_state* parameter the function will return either a random split or a fixed split based on the parameter value (Scikit-learn, e).

Leave-One-Out Cross-Validation

Leave One Out cross-validation (LOOCV) is conceptually similar to scikit-learns training and testing split function in that *leave one out* creates one subset of the dataset for training and another subset for testing (Kohavi, 1995). However, as the name implies *leave one out* will use all samples but one to train the model, then test the model on the remaining one sample. This method of training and testing a machine learning model can quickly become computationally expensive on larger datasets (Kohavi, 1995).

3.1.4 Classification and Regression Trees in scikit-learn

Scikit-learns regression models allows for hyperparameter tuning. A hyperparameter is a parameter for a machine learning model that is set before the learning occurs, rather than learned by the model itself (Claesen and De Moor, 2015). Hyperparameter tuning is crucial to developing accurate, well-fitting models for a given dataset (Koehrsen, 2018).

3.1.5 Feature Selection

A core process in machine learning methods is feature selection. In this process a subset of available features from a given dataset are compared using methods such as k-fold cross-validation, the subset that has the highest contribution to prediction accuracy and the lowest amount of dimensionality will be used in the machine learning model (Bermingham et al., 2015). While feature selection bears some similarities to machine learning model selection it is a separated process done prior to evaluation of model performance.

3.2 Validation with SPSS

IBMs statistical software package called IBM SPSS Statistics serves as a validation tool for this project. SPSS is a statistical analysis tool that is widely used in a variety of business and research fields (Piatetsky, 2013), offering a comprehensive set of tools for decision making, predictive analysis and data mining techniques (Quintero et al., 2012). This statistical package has decades of history and has been developed by one of the most well-known information technology companies in the world, making it a well-validated software package whose results can be reliably depended on.

3.2.1 Significance of independent variables

The significance of each independent variable was calculated by using scikit-learns *f_regression* submodule found in the *sklearn.feature_selection* module. Each independent variable in the dataset was passed into f_regression along with the desired dependent variable. How the significance (p-value) is calculated, from scikit-learns documentation page:
The correlation between each regressor and the target is computed, that is,

$$((X[:,i] - mean(X[:,i])) * (y - mean_y))/(std(X[:,i]) * std(y))$$

It is converted to an F score then to a p-value. (Scikit-learn, c).

In order to validate scikit-learns P-values for significance, the SPSS system was used to calculate p-values of each independent variables correlation to prosthesis longevity. These were presented as soon as SPSS had fitted the model to the data.

3.2.2 Machine Learning Model Evaluation Metric

For this project an adjusted calculation of the coefficient of determination was used as the metric used for evaluating the machine learning model performance. The coefficient of determination is denoted as R-squared or R^2 and it is a key output in regression analysis

(Rao, 1973). The coefficient of determination is widely used for linear regression models as a goodness-of-fit metric (Cameron and Windmeijer). Goodness-of-fit is a term used for how well a statistical model fits, or explains, a set of observations. The calculated coefficient of determination in multiple regression analysis is between 0.0 and 1.0 representing the proportion of the variance in the dependent variable that is predicted from the independent variable(s). A R^2 score of 0.5 can be interpreted as that 50% of the variance in the dependent variable can be explained by the independent variable(s).

The coefficient of determination metric has been criticized for not sufficiently telling the whole story of how well a linear regression model fits a set of observation (Stone et al., 2013). R^2 scores can be increased by increasing the number of independent variables used for the model (Minitab, 2018). This increase can be artificially heightened if the independent variables do not significantly contribute to the dependent variable, introducing noise to the prediction. An extension to calculation of R^2 in which the number of regressors used are taken into account is called *adjusted R^2*. This penalizes overuse of independent variables and provides an unbiased estimate of the population R^2 (Minitab, 2018).

3.3 Design Science

According to Hevner et al. (2004, p.83) two paradigms are present in the field of information system research - behavioural science and design science. Behavioural science is concerned with the explaining or predicting human or organizational behaviour, while design science is concerned with developing artifacts that extend the boundaries of human and organizational capabilities.

As the goal of this master thesis project was to design and develop an artifact in the form of a software application that can benefit total hip arthroplasty surgeons and doctors by creating a bridge between performing machine learning techniques on biomedical data and a user-friendly, simple interface the project falls within the boundaries of design science research. Because of this, the master thesis project was carried out following the guidelines established by Hevner et. al as its base values. While the guidelines are not necessarily a strict set of rules to follow (Hevner et al., 2004) conducting the research with the help of the guidelines, and for each guideline to be addressed in some manner are by Hevners own words vital for the design science research to be complete. The seven guidelines of design science research can be found in Table 3.1 below.

These guidelines were applied to the research conducted in this master project in order to develop the HALE system and its comparative evaluations as a design artifact.

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

Table 3.1: The seven guidelines for conducting design science research.

Design as an artifact

This master project has produced some artifacts in the form of the HALE system for longevity estimation, and the comparative data generated on evaluating the performance and accuracy of two approaches for continuous value estimation through use of regression model.

Problem relevance

In the relation to the design science research guidelines a problem is defined as “the differences between a goal state and the current state of the system” (Hevner et al., 2004). The current state is represented by that doctors and surgeons do not have a reliable way of producing accurate estimates of how long a hip prosthesis will last in a given patient, and the goal state is that they would have a tool capable of providing this in a useful, understandable and user-friendly manner by using reliable, well-proven data mining methods.

Design evaluation

The usability of the artifact was evaluated using well-proven evaluation methods being user testing supplemented with semi-structured interviews and heuristic evaluation by experts. These methods are elaborated on in Section 3.5 and their results available in chapter 7 of this thesis.

Design contributions

The fourth guideline requires clear contributions in the area that the artifact falls within, its evaluation knowledge and / or construction knowledge. While the master project chiefly contributes with its artifact in the field of medicinal informatics, the artifact produced in this master project was developed using proven methods and methodologies whose knowledge contributes to the field of system development research and system usability evaluation.

Research rigor

The artifact has been developed by applying a software development methodology designed specifically for single-person development teams (see Section 3.4). This methodology builds on other well-known methodologies and has been proven superior to ad-hoc development. Evaluation of the artifact, as aforementioned, is done using well-known efficient and constructive evaluation methods elaborated on in Section 3.5.

Design as a search process

The research regarding the design artifact, its development and subsequent evaluations were all carried out in accordance to principles set in the used methods and methodologies.

Communication of research

As the research conducted throughout this master project has been documented through this thesis which will be publicly available through University of Bergens open research archive as well as in two outlined scientific publications, this satisfies the communication of research guideline.

3.4 Development Methods and Methodologies

3.4.1 Personal Extreme Programming

Personal Extreme Programming (PXP) is an agile system development methodology that has its roots in two other development methodologies, it is a modification of Personal System Development (PSP) that adds additional concepts from Extreme Programming (XP) (Dzhurov et al., 2009).

As an agile process PXP aims to reduce time spent documenting work, thus adopting some but not all scripts from PSP (Dzhurov et al., 2009) so that PXP Uses the core principles without overburdening the developer with documentation. Personal Extreme Programming is based on the following principles:

1. Developers need a disciplined approach to the development process, they need to follow the process principles and practices.

2. Developers need to track, measure and analyze their work daily.

3. Developers are required to learn from performance variations and need to focus on improving performance based on the collected project data.

4. Developers are required to do continuous testing.

5. Developers need to fix defects early rather than late in the development process.

6. Developers should focus on automating of their daily work as much as possible.

These principles are accompanied by fourteen practices. Six of these practices are adopted from PSP and another six are borrowed from XP. The combination of practices is designed to emphasize on the disciplined project structuring from PSP while embedding the agile

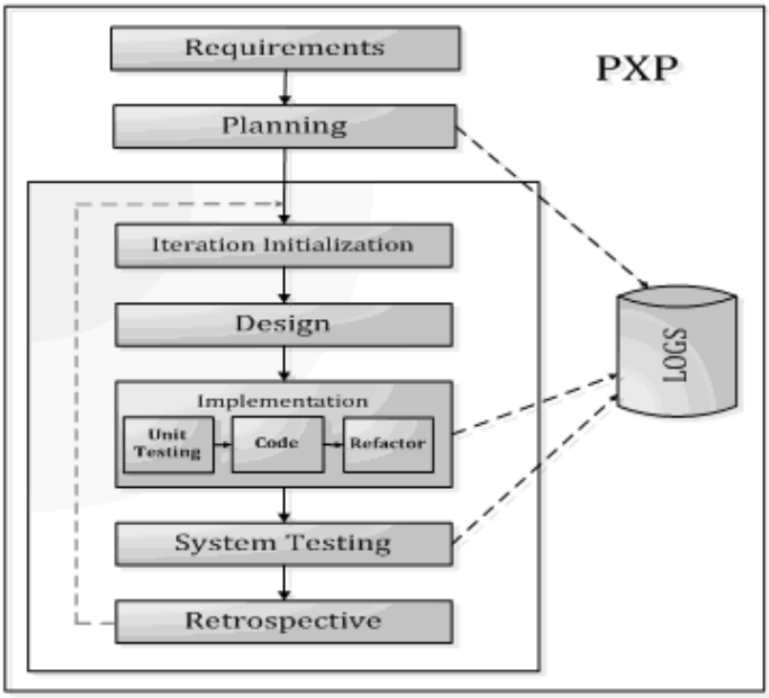


Figure 3.1: Personal Extreme Programming phases of development

practices that embrace change and iterative development. In PSP developers are required to write a planning script prior to the development process, this script will act as a guide throughout the development (Humphrey, 2000). PSP focuses on extensive time management for planning and reporting throughout the development process. In PXP his focus that has been diminished in favor of focus on general productivity in-line with the agile manifesto. PXP still requires time management and this is largely reliant upon experience from prior projects (Dzhurov et al., 2009).

Phases of PXP

Personal Extreme Programming is an iterative development methodology. Initial tasks planning and requirement establishment does not reiterate after project initiation, but the remaining phases are reiterated throughout development until the project is complete (Dzhurov et al., 2009). Data such as time spent on each phase is noted for retrospect. These phases are:

1. **Iteration initialization** that starts by selecting a set of tasks to complete during this iteration.

2. **Design** regarding the system architecture, its modules and classes.

3. **Implementation** in which the coding is conducted. This phase consists of three subphases that is conducted chronologically; *unit testing, coding* and *refactoring.*

4. **System testing** consists of ensuring that all unit tests written in the prior phase are passed.

5. **Retrospective** signifies the end of each iteration cycle, in which the data collected throughout the phases is analyzed.

During the retrospective phase the developer has to measure the development process and whether or not the system is fulfilling the established requirements. If the system does meet the requirements this marks the end of the project, if it does not (and there is time) a new iteration cycle starts. The full cycle of PXP phases process is depicted in Figure 3.1.

3.5 Usability Evaluation Methods

The usability of an information system can be highly subjective, depending on factors such as the users knowledge and skill with other information systems and attitude towards potential problems faced when using said systems (Longo and Dondio, 2016). Several methods of gauging the usability of the artefact developed in this project were conducted, each targeting their own set of users in order to assess the system usability from several subjective perspectives.

3.5.1 Qualitative Data Gathering

3.5.1.1 Semi-structured interviews

For gathering of qualitative data, semi-structured interviews is one of the most commonly used forms of data gathering (Kallio et al., 2016). In comparison to the rigorous set of questions found in structured interviews, a semi-structured interview allows for deviation from the scheduled list of questions in order to pursue new ideas, topics or themes based on what the interviewee provides during the process. This openness can lead to interesting and useful information that may not be explored during a structured interview.

3.5.2 Quantitative Data Gathering

3.5.2.1 Heuristic Evaluation

A heuristic evaluation of an information system is an evaluation of the usability of its user interface. The evaluation itself is based on Jakob Nielsens 10 heuristics that can be found in Table 3.2, and the evaluation is an informal method of assessing the usability of a system.

These heuristics are meant to help identify usability problems in computer software regarding the user interface and its design, often while the software is under development (Nielsen, 1994).

The method employs the use of experts. These experts are people who are knowledgeable and skilled in using various information system user interfaces. The experts evaluate and judge the system according to their own subjective opinions in regards to the ten heuristics (Nielsen and Molich, 1990).

3.5.2.2 System Usability Scale

System Usability Scale (SUS) is a self-proclaimed *'quick and dirty'* scale for measuring the perceived usability of computer system Brooke (1996). The evaluation method is a five-level Likert scale in which a ten-item questionnaire is answered by the participants. Despite its self-proclaimed description the evaluation method has been proven to be robust and reliable Brooke (1996).

Evaluation of a computer system is conducted through exposing a participant to the system through completion of a set of tasks. When all tasks are completed or as completed as can be, participants answer the ten-item questionnaire ranging from 1 to 5, where 1 represents ***strongly disagree*** and 5 represents ***strongly agree***. The ten items are

1. **I think that I would like to use this system frequently.**
2. **I found the system unnecessarily complex.**
3. **I thought the system was easy to use.**
4. **I think that I would need the support of a technical person to be able to use the system.**
5. **I found the various functions in this system were well integrated.**
6. **I thought there was too much inconsistency in this system.**
7. **I would imagine that most people would learn to use this system very quickly.**
8. **I found the system very cumbersome to use.**
9. **I felt very confident using the system.**
10. **I needed to learn a lot of things before I could get going with this system.**

These ten items are designed to be as generalized as possible so that the evaluation method can be used on a broad set of computer systems but specific enough to provide relevant

Table 3.2: Nielsens 10 Heuristics.

Heuristic	Description
Visibility of system status	The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.
Match between system and the real world	The system should speak the user's language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.
User control and freedom	Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.
Consistency and standards	Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.
Error prevention	Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.
Recognition rather than recall	Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.
Flexibility and efficiency of use	Accelerators—unseen by the novice user—may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.
Aesthetic and minimalist design	Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.
Help users recognize, diagnose, and recover from errors	Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.
Help and documentation	Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

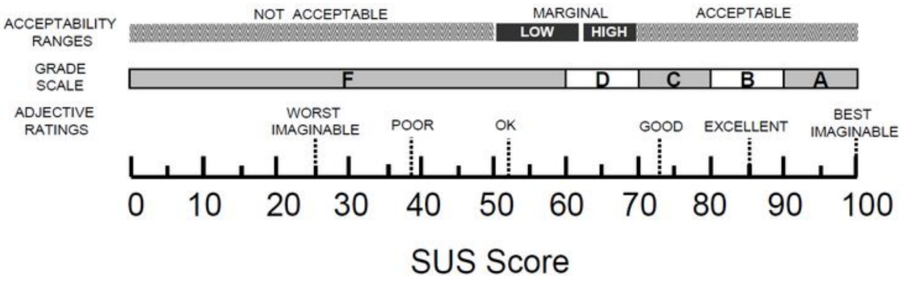


Figure 3.2: The System Usability Scale final score scale with appropriate grades.

usability feedback (Brooke, 1996). Every even question is positively loaded and each odd question is negatively loaded by design.

The system usability scale scoring of systems ranges from 0 to 100. Lower scores indicate lower usability, higher scores indicate higher usability in a system. The score is calculated by summing the score from each of the ten items. Each item contributes between 0 and 4. Items 1, 3, 5, 7 and 9 contribute their respective items Likert scale value minus 1. Items 2, 4, 6, 8 and 10 contributes 5 minus the scale value. This alternating of sums where odd numbered scale items are positive contributions and even numbered scale items are negative contributions has been designed to keep participants from mindlessly checking a sum for all items (Bangor et al., 2009). The sum of all items is then multiplied by 2,5. Although the scores range from 0 to 100 they should not be considered a percentage, rather a percentile.

As evident in Figure 3.2 a score below 60 is deemed unacceptable (Brooke, 1996), but research conducted in 2009 show that the total mean score of 1433 web-pages is 68.2 (Bangor et al., 2009) indicating that a score below 68 would be lower than average, thus indicating that the general usability of the user interface is less than satisfactory which is in turn undesirable.

Chapter 4

Establishing Requirements

In software development requirements are described as the statements of the intended product that acts as a specification for how the product should perform (Preece et al., 2015). These requirements should be as clear, concise and unambiguous as possible to avert any misinterpretation on any part from anyone involved in the development process. Establishing requirements is a core practice of well-executed system development as these requirements will lay the foundation of what is to be achieved by the developed system. Two types of requirements have proved traditional in software development - functional requirements that detail the specifics on what the system should do, while non-functional requirements that detail specific restrictions for the product and its development (Preece et al., 2015).

The system development methodology utilized for this project emphasizes an early establishment of requirements that is stable throughout the project cycles. While the requirements for developing the HALE system were established in the early stages before the coding cycles, they were revisited in each iteration. While most requirements were established through conversations with expert users from the biomaterial laboratory at Haukeland University Hospital, some were designed for validation through comparative testing with SPSS which was a feature not intended for the end user.

4.1 Functional Requirements

The functional requirements set for the HALE system were resolved in collaboration with some expert users. These requirements made up the foundation of the systems capabilities for its intended users and are as follows:

1. The system must be able to predict the longevity of a given patients prosthesis.

2. There must be a way for users to input patient information.

3. The systems machine learning components must obtain its information from an expandable source.
4. The system must allow users to view previously entered user information during the workflow.
5. The system must allow users to edit previously entered user information during the workflow.
6. The user must be allowed to decide which data columns should be used for prediction.
7. The system must provide some statistical background for its predictions to its users.
8. The user must be allowed to reset the system at any point in the workflow.

4.2 Non-Functional Requirements

The non-functional requirements established for the HALE system were:

1. The system must be easy to use.
2. The system must have a short and simple workflow.
3. The system must be compatible with older and/or less powerful computers.
4. The system must be verifiable.

These two sets of requirements set precedence of the core functionality offered to the end user in regards to single patient prosthesis longevity estimation, as well as restrictions on system complexity and combability with older, slower computers. Additionally the models that estimate longevity were required to be transparent and verifiable as this is required for healthcare related information technology.

4.3 Intended User

The user interface and functionality of the system was developed towards a target demographic throughout the development process. The target demographic consists of two groups. The first group was the researchers at the Biomatlab at Haukeland University Hospital, the second group was the physicians responsible for total hip arthroplasty patients both before and after surgery.

Chapter 5

Prototype Development

This chapter presents the development process that produced the design science artifact using the methods discussed in Section 3.4. The artifact itself is presented as well, with details on its workflow and design.

5.1 Tools and Technologies Used

Several freely available technologies and tools enabled the development of the design artifact produced by this master project. These were sectioned into two parts - front-end development and back-end development. The front-end development section concern the technologies used to construct the graphical user interface of the artifact. The back-end section concerns the technologies used to implement the machine learning and data handling aspects of the artifact. While some of the technologies overlap in their capabilities they were sectioned as to how this project utilized them.

5.1.1 Front-End Technologies

5.1.1.1 JavaScript

According to Flanagan (2011) *JavaScript* is one of the three core technologies that any aspiring web developer need to learn. The other two technologies are HTML to specify the content of web-pages and CSS to specify their presentation. JavaScript greatly enhances the dynamic capabilities of an otherwise static HTML document. Content can be dynamically hidden and shown at appropriate intervals such as time-gating or expanding a chosen content container.

5.1.1.2 jQuery

JavaScript can be extended by the use of libraries. The library *jQuery* is designed to be a light-weight, efficient and feature-rich extension (jQuery Foundation) that simplifies certain JavaScript functionalities such as HTML document manipulation, event-handling, animations and AJAX (Asynchronous JavaScript And XML) calls.

5.1.1.3 HTML

HyperText Markup Language, more commonly referred to as HTML, is a mark-up language for efficiently structuring world-wide-web documents, pages and applications (Flanagan, 2011). The HTML documents contains HTML elements (tags) that describe the content type and its structure, which in turn contain the information content.

5.1.1.4 CSS

Cascading Style Sheets (CSS) are style sheet documents that describe the presentation of a HTML document (Flanagan, 2011). CSS documents can determine the shape, font size and type, color, position and additional effects of HTML elements such as shadows, opacity, transitional effects and so on.

5.1.2 Back-End Technologies

5.1.2.1 Python

Python is a programming language designed to be a high-level general purpose language. The language has a wide variety of applications, can be extended into C and C++ for better computing speeds on intensive tasks and provides strong structuring constructs that enables clear and logical application for large and small tasks (Kuhlman, 2009). In the Python Enhancement Proposals (PEPs), (available on python home page https://www.python.org/dev/peps/) *The Zen of Python* states the core philosophy of Python as a language, including aphorisms such as *explicit is better than implicit*, *simple is better than complex*, *complex is better than complicated* and *readability counts* (Peters).

Python was chosen as programming language for the artifact due to the languages' focus on simplistic-yet-powerful syntax, as well as its extensive standard library (Python Software Foundation, 2012) and powerful third party libraries.

5.1.2.2 Flask

Flask is a microframework for web development with Python. This framework acts as a connective bridge between HTML and Python, enabling Python code to execute from a web-based user interface all while being light-weight, easy and extensible (Ronacher).

5.1.2.3 Scikit-learn

The machine learning elements to the system utilizes the free machine learning library **scikit-learn**. Scikit-learn is a framework that integrates various machine learning algorithms. The framework is intended to bring the most state of the art algorithms for medium-scale supervised and unsupervised problems to non-specialists through Python (Pedregosa et al., 2011).

5.1.3 Development Tools

A set of development tools were used to utilize the technologies listed above when constructing the artifact. These were specialized tools created for the purpose of both simplifying the development process as well as extend the developers capabilities.

5.1.3.1 PyCharm IDE

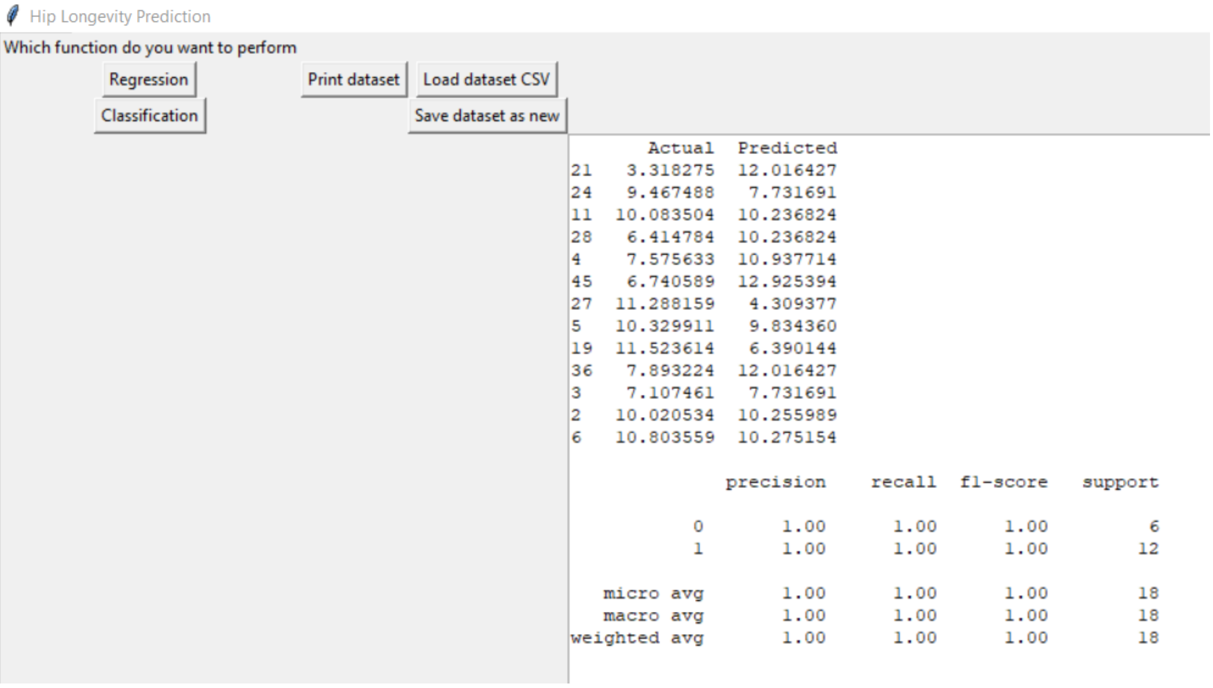
The PyCharm Integrated Development Environment by JetBrains was used for all coding during the development process. PyCharm was chosen due to its extensive integrated development tools as well as general experience and familiarity with its Java counterpart, IntelliJ. PyCharm is not restricted to Python as a language, downloadable packages enables multiple programming language support such as all technologies mentioned in the front-end section.

5.1.3.2 Git and GitHub

Git is a free, open-source distributed version control system used for system development (Torvalds, 2017). GitHub is an online service that offers hosting of Git repositories. Git was used in conjunction with GitHub to allow for continuous development on several workstations as well as enabling instantaneous sharing of any state of the system during development.

5.2 Development Process Method

As mentioned in chapter 3 the system was developed using Personal Extreme Programming. Three prototypes were developed in three iterations of the development phase. Each iter-



```
        Actual   Predicted
21   3.318275  12.016427
24   9.467488   7.731691
11  10.083504  10.236824
28   6.414784  10.236824
4    7.575633  10.937714
45   6.740589  12.925394
27  11.288159   4.309377
5   10.329911   9.834360
19  11.523614   6.390144
36   7.893224  12.016427
3    7.107461   7.731691
2   10.020534  10.255989
6   10.803559  10.275154

             precision    recall  f1-score   support

           0       1.00      1.00      1.00         6
           1       1.00      1.00      1.00        12

   micro avg       1.00      1.00      1.00        18
   macro avg       1.00      1.00      1.00        18
weighted avg       1.00      1.00      1.00        18
```

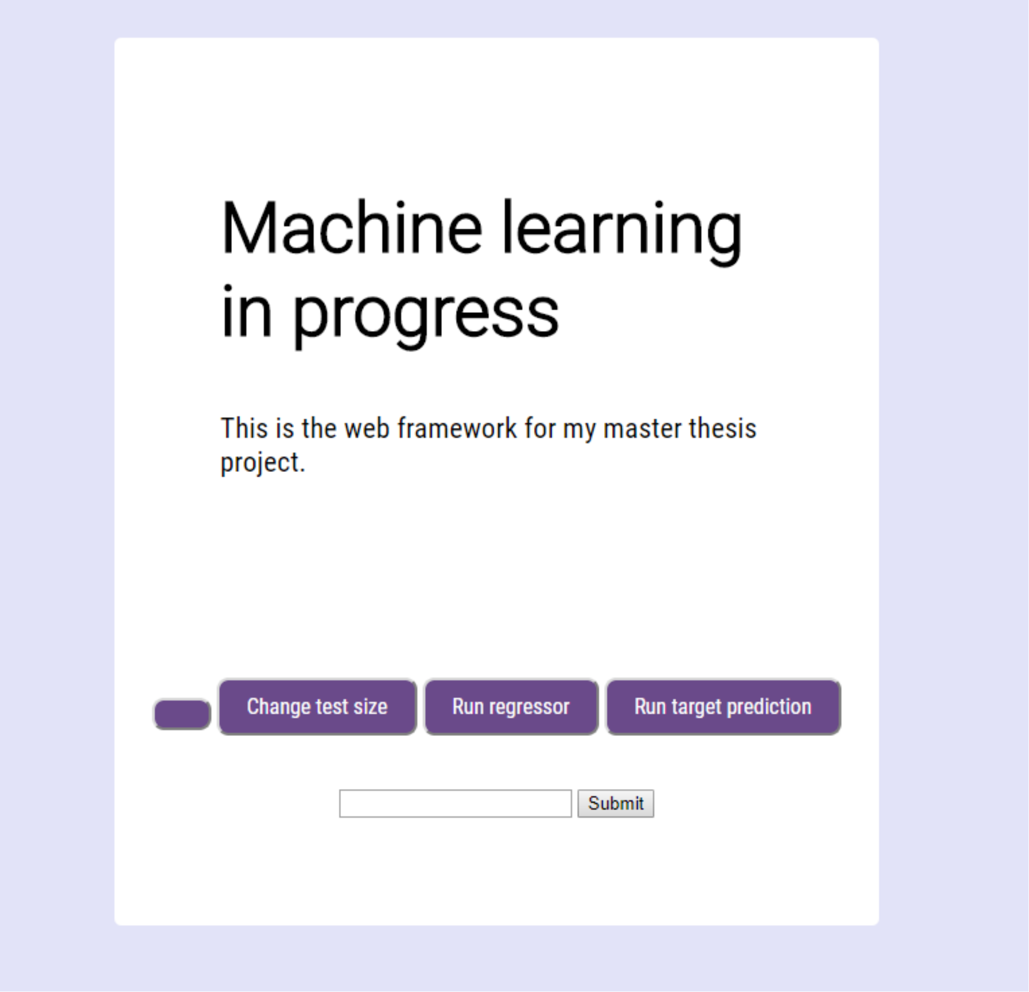
Figure 5.1: First iteration: Main menu

ation was informally presented and discussed with a user from the biochemical laboratory user group, after which new requirements for the next iteration were set. The first iteration prototype was developed with only vaguely abstract requirements in place - a somewhat exploratory iteration to get an overview of what can feasibly be completed in regards to this system and the goals of the project. The two next iterations were developed with an increasingly concrete set of requirements.

5.2.1 First Prototype Iteration

This iteration was focused on the basic requirements of the project - read the dataset, perform some machine learning technique(s), and present the results. The developed prototype was a highly bare-bones graphical user interface that enabled some decision tree regression as well as decision tree classification, as seen in Figure 5.1.

The user interface consisted of a set of interactible buttons and an output text field. Some classification was performed on which samples would have a revision surgery, in which their *case* value would equal 1 if they did, and the results of performing longevity estimations on a test set using the trained regression model was present through their respective buttons. The interactive elements, their layout and the general look and feel of this user interface had an archaic quality to it that was not well received.



Machine learning in progress

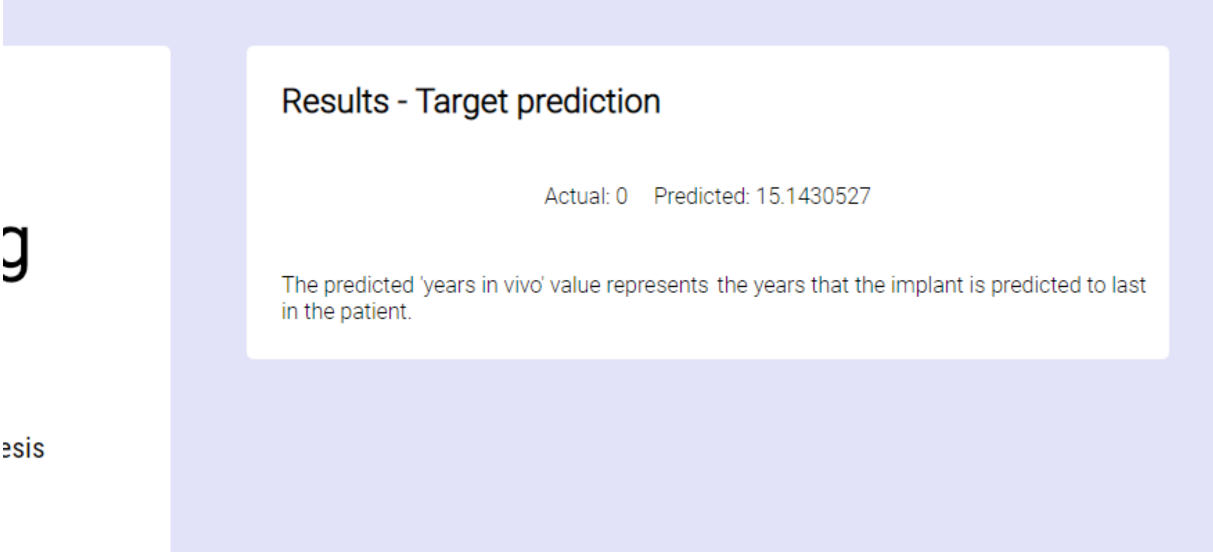
This is the web framework for my master thesis project.

Figure 5.2: Second iteration: Main menu

5.2.2 Second Prototype Iteration

The second iteration prototype of the system was created to explore possible graphical user interface options as well as desired functionality. As shown in Figure 5.2 the second iteration prototype allowed for training and testing a decision tree regression model whose hyperparameters were set to default values, as well as using the same model to estimate the longevity of a patient whose information was stored locally in a separate file from the rest of the data. The only mutable element presented to the user was changing the testing and training split percentage, a feature not necessary for any end-user.

This iteration was the first to implement Flask for bridging a web-based user interface to the underlying Python machine learning modules. The decision to use a web-based user interface was taken in parts due to the familiarity most people have with using web browsers as content providers and in parts due to the possibilities for moving the computing process heavy calculations away from the clients (users) computer to a more powerful remote system. A specialized computer constructed for dealing with machine learning techniques and larger dataset scaling can potentially be used for multiple users.



Results - Target prediction

Actual: 0 Predicted: 15.1430527

The predicted 'years in vivo' value represents the years that the implant is predicted to last in the patient.

Figure 5.3: Second iteration: Result from performing a target prediction

Evident from the training results seen in Figure 5.4 test data had leaked into the training subsets for the regression model. Most of the predictions were to the point equal to their actual values while a few predictions were genuine results of the model. Design-wise, the redesign was lauded in comparison to the first iteration. When the user would start the system only the main menu in which all the interactive elements were bundled would be visible, while the results of running predictions would appear to the right of the menu.

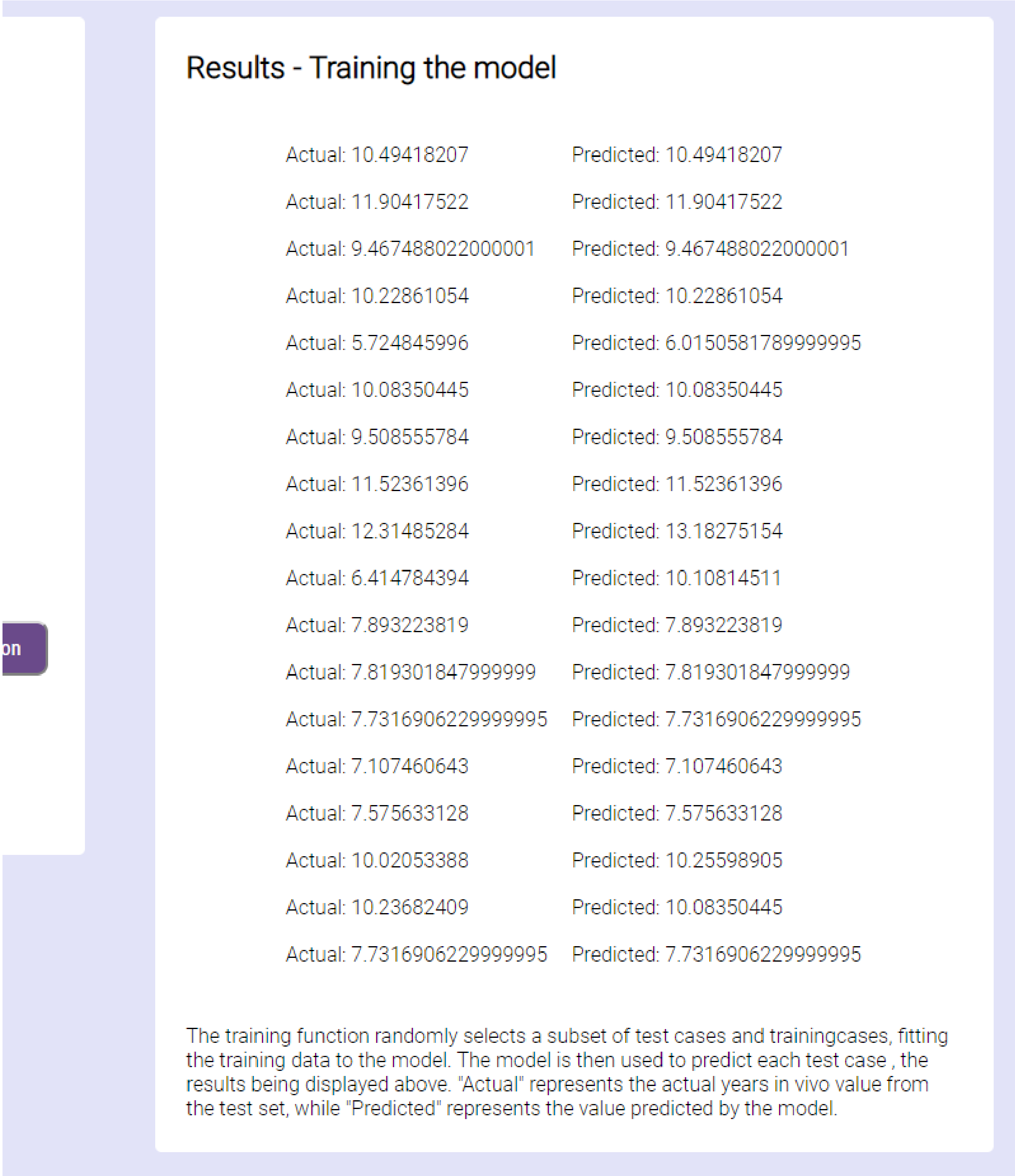
5.2.3 Third Prototype Iteration

The third prototype was the last that was developed during this project. Compared to its prior iteration this prototype has a somewhat extended functionality in comparative machine learning model evaluation, and significantly improved target sample estimation functionality. The general design can be seen in Figure 5.5.

5.2.3.1 User Workflow and System Design - Targeted Sample Prediction

The workflow of the system is designed to be as short as possible for the user. An abstraction of the workflow is modeled in Figure 5.6. In a general use case the user would start the system, enter their desired patient information before starting the longevity estimation process. The design is intended to accommodate both experienced and inexperienced users alike. The steps of the system workflow will be referred to as pages.

As buttons are used as tools for the core functionality in the system they carry some general continuity. The most vital buttons are styled with a striking dark orange color, have a large surface and light up when hovered (1a, 1b in Figure 5.7). Navigation buttons carry similar color but are less pronounced, their functionality implied in their location rather than their

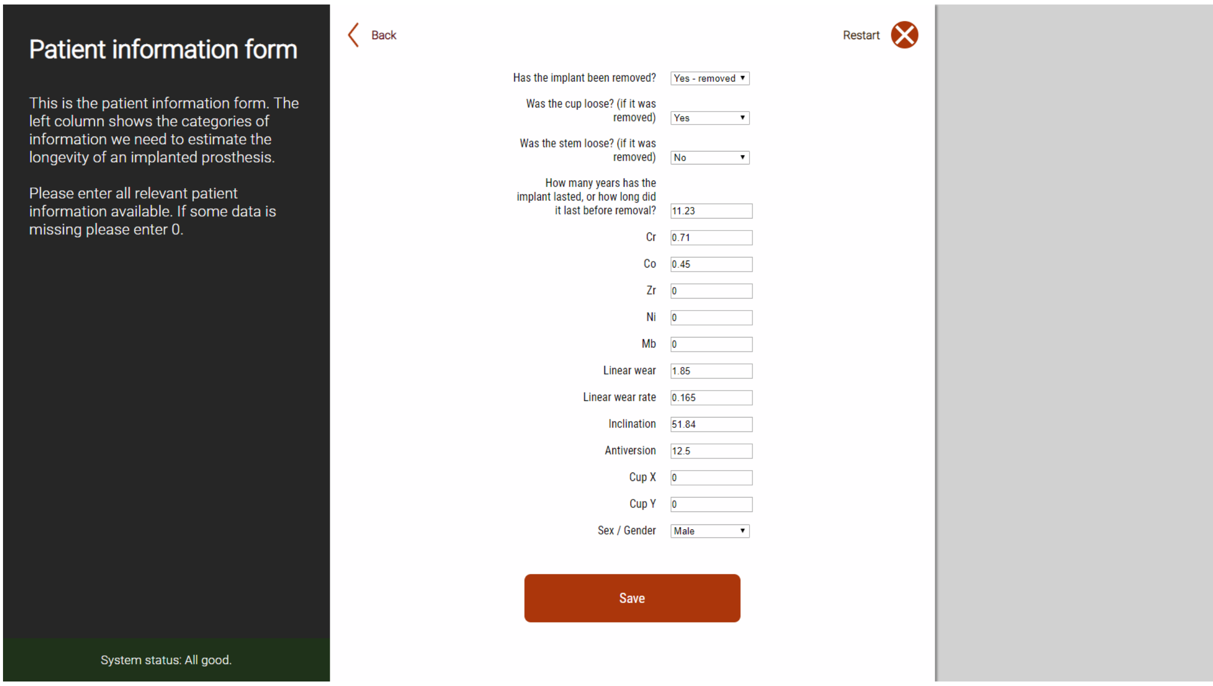


Results - Training the model

```
Actual: 10.49418207          Predicted: 10.49418207
Actual: 11.90417522          Predicted: 11.90417522
Actual: 9.467488022000001    Predicted: 9.467488022000001
Actual: 10.22861054          Predicted: 10.22861054
Actual: 5.724845996          Predicted: 6.0150581789999995
Actual: 10.08350445          Predicted: 10.08350445
Actual: 9.508555784          Predicted: 9.508555784
Actual: 11.52361396          Predicted: 11.52361396
Actual: 12.31485284          Predicted: 13.18275154
Actual: 6.414784394          Predicted: 10.10814511
Actual: 7.893223819          Predicted: 7.893223819
Actual: 7.819301847999999    Predicted: 7.819301847999999
Actual: 7.7316906229999995   Predicted: 7.7316906229999995
Actual: 7.107460643          Predicted: 7.107460643
Actual: 7.575633128          Predicted: 7.575633128
Actual: 10.02053388          Predicted: 10.25598905
Actual: 10.23682409          Predicted: 10.08350445
Actual: 7.7316906229999995   Predicted: 7.7316906229999995
```

The training function randomly selects a subset of test cases and trainingcases, fitting the training data to the model. The model is then used to predict each test case , the results being displayed above. "Actual" represents the actual years in vivo value from the test set, while "Predicted" represents the value predicted by the model.

Figure 5.4: Second iteration: Results from training and testing a non-parametric scikit-learn decision tree regression model.



Patient information form

This is the patient information form. The left column shows the categories of information we need to estimate the longevity of an implanted prosthesis.

Please enter all relevant patient information available. If some data is missing please enter 0.

System status: All good.

Has the implant been removed?	Yes - removed
Was the cup loose? (if it was removed)	Yes
Was the stem loose? (if it was removed)	No
How many years has the implant lasted, or how long did it last before removal?	11.23
Cr	0.71
Co	0.45
Zr	0
Ni	0
Mb	0
Linear wear	1.85
Linear wear rate	0.165
Inclination	51.84
Antiversion	12.5
Cup X	0
Cup Y	0
Sex / Gender	Male

Figure 5.5: The general design of the system visualized through the patient information input form section.

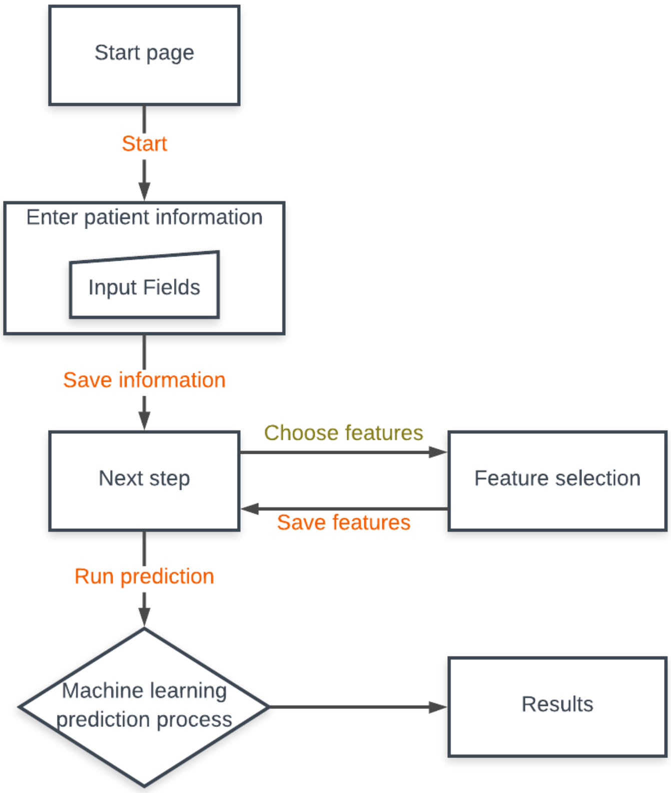


Figure 5.6: A workflow abstraction of the steps available and required in the system.

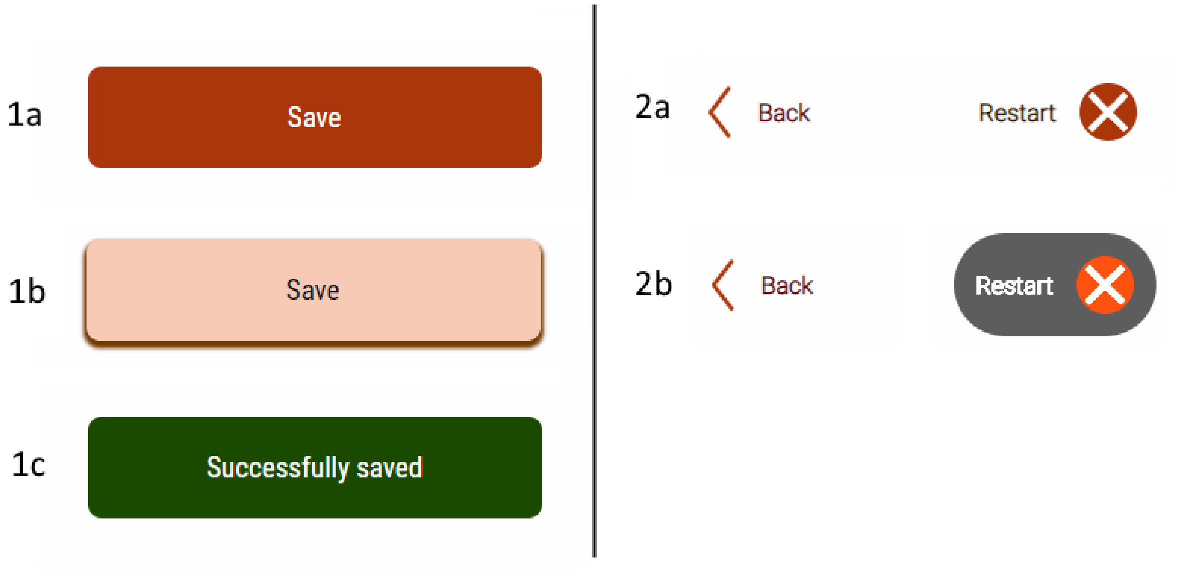


Figure 5.7: Various buttons found in the HALE system. Button 1a through 1c show the *save* button from the patient information form during its three phases idle, hovered and clicked.

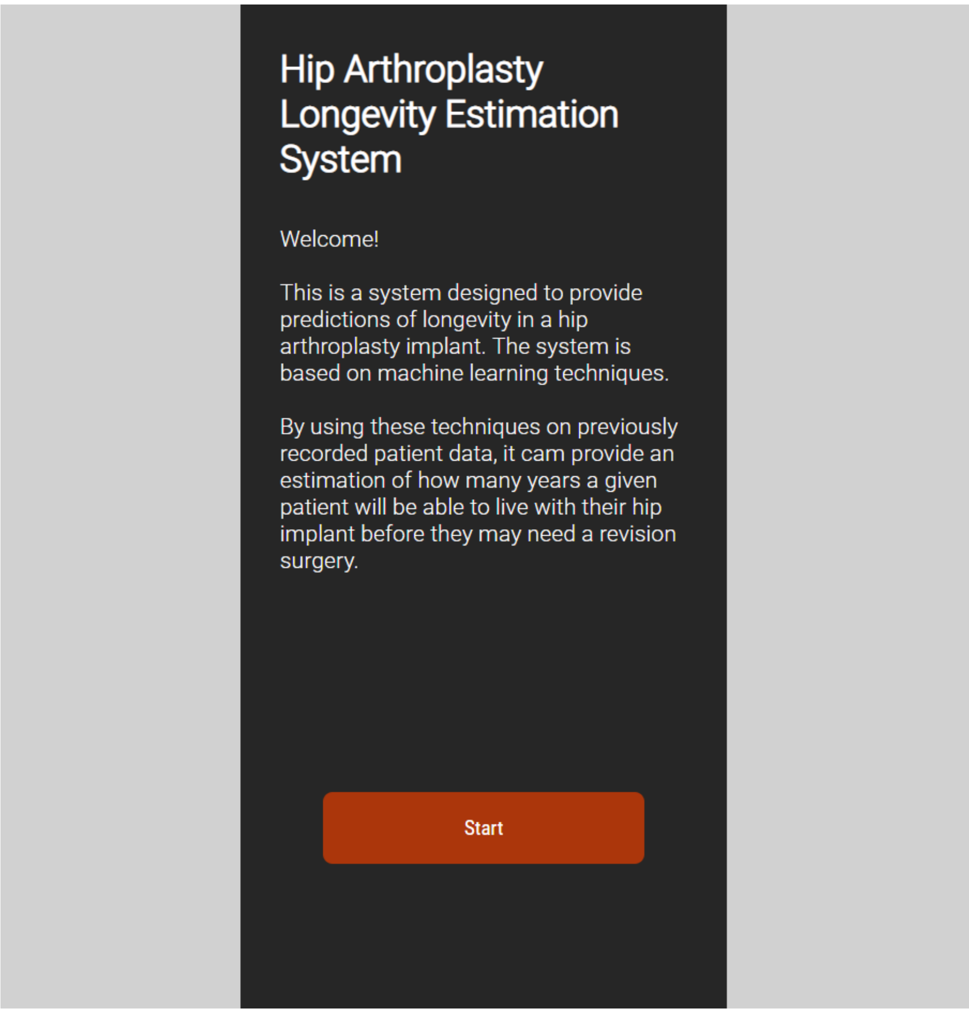
contrast. These have a much graver hover effect (2a, 2b in Figure 5.7).

Start

Start is, as its name implies, the starting point of the workflow in HALES. It is depicted in Figure 5.8. This is the initial page that explains to the user what the system does, with the intention of inaugurating the user to HALEs contextual explanatory texts as well as its interactive elements, presented in the dark column. Pressing the *start* button will slide the dark column to the side, making place for a white column that will contain most of the systems functions.

Patient Information Form

The patient information form page requires the user to enter all the information available on the patient. The fields are HTML input fields that either only accept integer and float inputs, or is a drop-down style menu with predetermined values. Some of the fields can be seen in Figure 5.9. When the user has saved their patient information a new element is introduced right outside the primary content column, as seen in Figure 5.10. When the user has entered their desired data pressing the *save* button will take them to the next page, dubbed **next step** due to its crossroad-like nature. It's on this page that the user may choose to either run the prediction process and produce an estimation for longevity on the given patient data, or manually set the desired regressors to be used in the machine learning model before running the prediction.



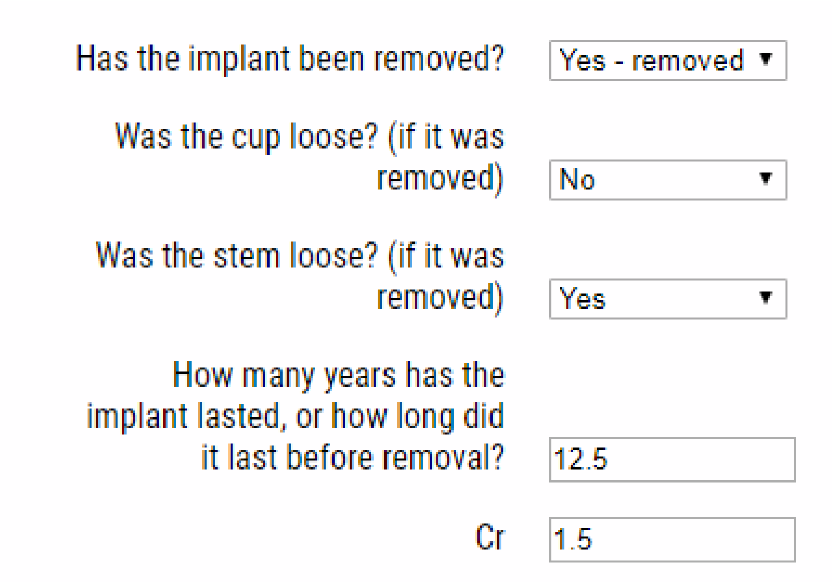
Hip Arthroplasty Longevity Estimation System

Welcome!

This is a system designed to provide predictions of longevity in a hip arthroplasty implant. The system is based on machine learning techniques.

By using these techniques on previously recorded patient data, it cam provide an estimation of how many years a given patient will be able to live with their hip implant before they may need a revision surgery.

Figure 5.8: By running machine learning procedures on previously recorded patient data, for each specific patient a number of years is estimated for which patient should not need a revision surgery.



Has the implant been removed?	Yes - removed ▼
Was the cup loose? (if it was removed)	No ▼
Was the stem loose? (if it was removed)	Yes ▼
How many years has the implant lasted, or how long did it last before removal?	12.5
Cr	1.5

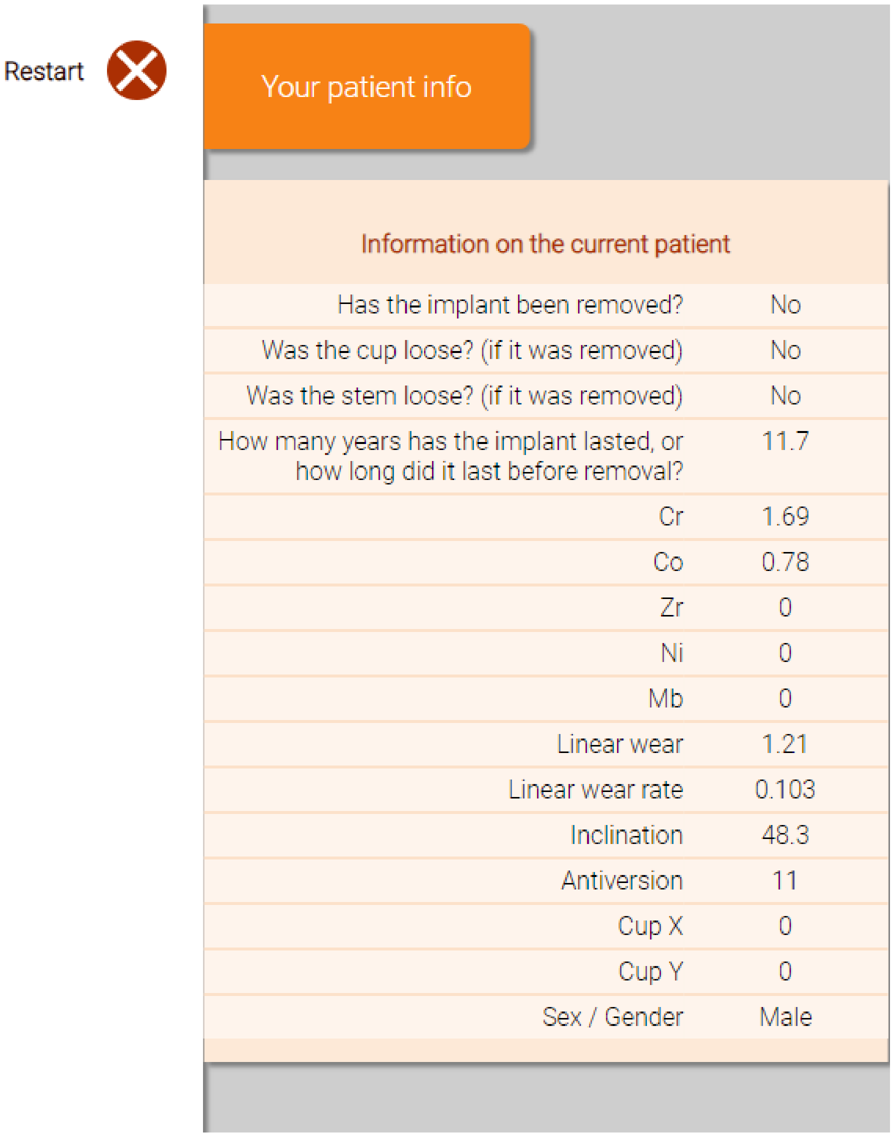
Figure 5.9: This is a crop from the patient information form. The implant longitude is based on the variables in the left columns.

Optional Regressor Selection

If the user wants to edit the regressors used during the prediction process they will be presented with the list of regressors, or features (see Figure 5.11). These are fetched from the dataset file and fed into a list of checkboxes, displayed for the user to edit at will. As with the input form the user needs to save their desired inputs through the same *save* button as they used before. The contextual menu on the left side of the screen will display an explanation as to what editing which regressors are chosen means for the underlying machine learning model, as well as a list of default regressors. If the user decides to not change anything they are always able to press either *back* or *reset* at any given time.

Prediction Results

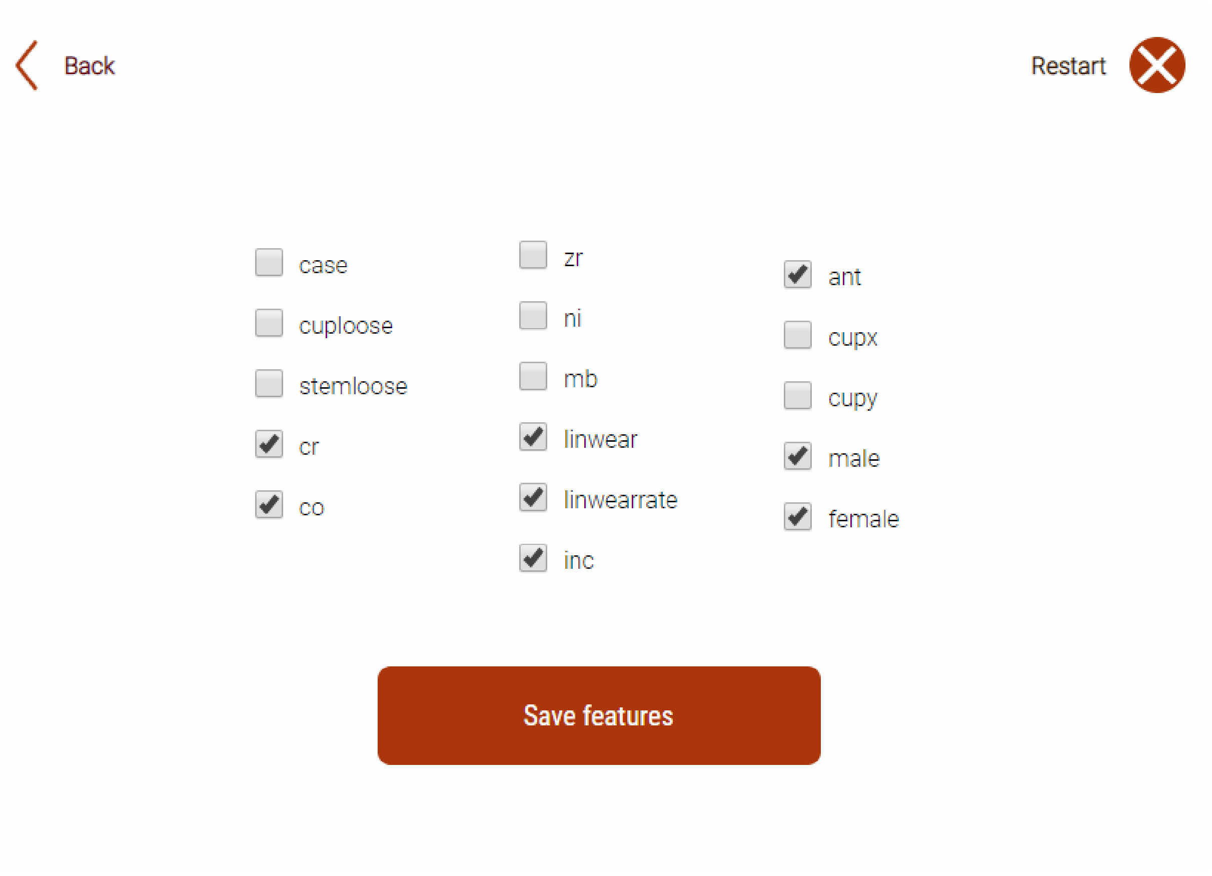
When the user has completed the necessary steps they will be presented with the hip prosthesis longevity estimation results after a brief animated loading screen. The results are displayed in the center of the white content column, the predicted years of longevity highlighted with the strong orange color as seen in Figure 5.12.



Information on the current patient

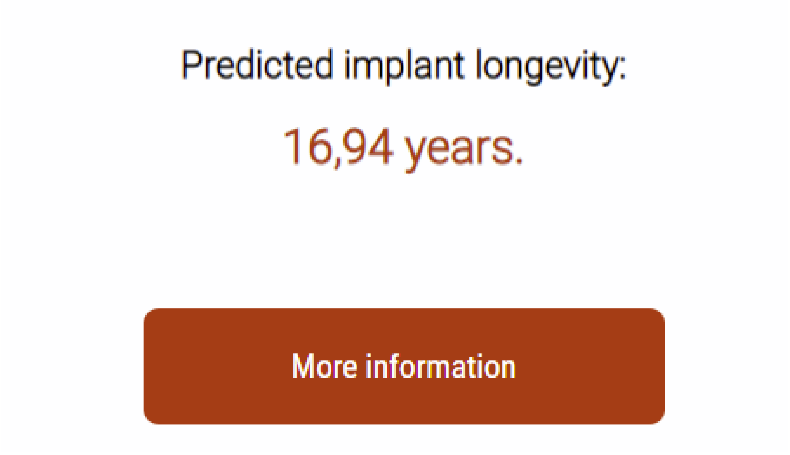
Has the implant been removed?	No
Was the cup loose? (if it was removed)	No
Was the stem loose? (if it was removed)	No
How many years has the implant lasted, or how long did it last before removal?	11.7
Cr	1.69
Co	0.78
Zr	0
Ni	0
Mb	0
Linear wear	1.21
Linear wear rate	0.103
Inclination	48.3
Antiversion	11
Cup X	0
Cup Y	0
Sex / Gender	Male

Figure 5.10: Patient information display box, accessible throughout all pages (including loading screen) in the system.



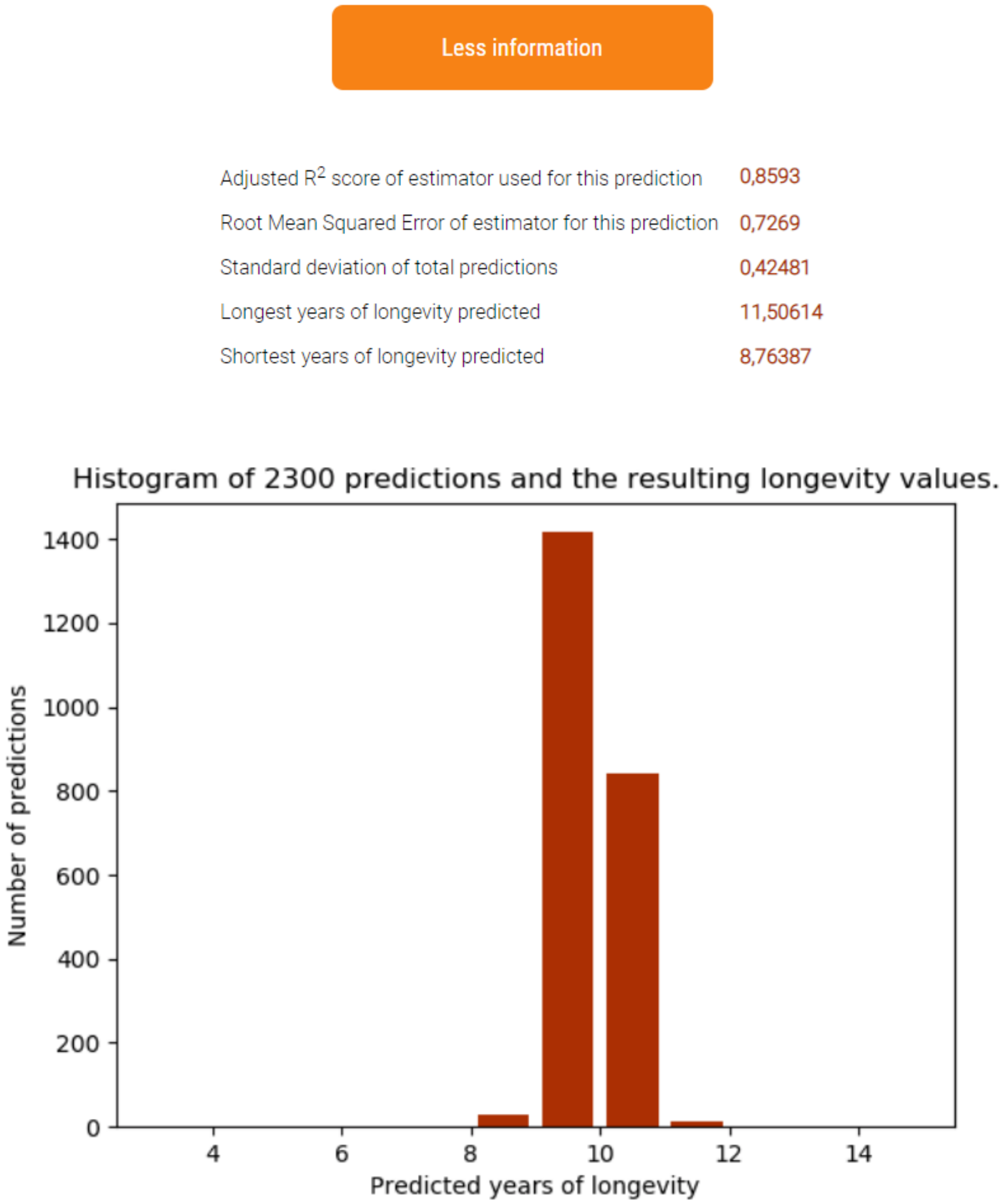
- □ case
- □ cuploose
- □ stemloose
- ☑ cr
- ☑ co
- □ zr
- □ ni
- □ mb
- ☑ linwear
- ☑ linwearrate
- ☑ inc
- ☑ ant
- □ cupx
- □ cupy
- ☑ male
- ☑ female

Figure 5.11: A list of available regressors whose column name values were fetched directly from the dataset and populated as a list of checkboxes. For this Figure the default values are enabled, the rest disabled.


Predicted implant longevity:

16,94 years.

Figure 5.12: Total Hip Arthroplasty prosthesis longevity estimation result in years.



Adjusted R^2 score of estimator used for this prediction	0,8593
Root Mean Squared Error of estimator for this prediction	0,7269
Standard deviation of total predictions	0,42481
Longest years of longevity predicted	11,50614
Shortest years of longevity predicted	8,76387

Histogram of 2300 predictions and the resulting longevity values.
Number of predictions
1400
1200
1000
800
600
400
200
0
4
6
8
10
12
14
Predicted years of longevity

Figure 5.13: Expanded information display containing statistical background for the longevity estimation and its performance.

Chapter 6

Implementation of Regression Models

This section explains how the aforementioned methods were applied in this project, as well as the results gained from applying these methods.

The developed system is comprised of two parts or modes. One mode is focused on users, usability and single longevity prediction on a target sample whose data is input by the user.

The other mode consists of comparative testing between methods in which the outcome generated by the system is fixed. This mode requires modification of the systems code to produce new results, and therefore this mode is reserved for calculating prediction model accuracies, performance, calibration of hyperparameters and so forth.

6.1 PARETO Dataset

A dataset was provided by the Biomatlab Research Group of the orthopedic clinic at Haukeland University Hospital. The dataset contains a set of samples from patients, a set which information was gathered and processed in conjunction with a research project dubbed PARETO. 49 samples were present in the dataset, of which 17 of the samples came from a control group who had not yet needed a revision surgery. The remaining 32 samples were patient records from revision surgery patient samples, all whose implants faied due to aseptic loosening. In all samples gathered the patient were implanted with the *Spectron EF* prosthesis developed by Smith-Nephew (Brien et al.).

Excluding observation identification, the PARETO dataset was comprised of 18 features per sample. These features are listed below:

1. **Case** determines whether the patient has had a revision surgery or not. 0 if the patient has not had revision surgery, 1 if they have.

2. **cupLoose** determines whether or not the cup component of the implant came loose before a revision surgery, 0 if it did not and 1 if it did.

3. **stemLoose** determines whether or not the stem component of the implant came loose before a revision surgery, 0 if it did not and 1 if it did.

4. **sex** represents the gender of the patient. 0 represents undefined, 1 is male and 2 is female.

5. **years in vivo** is the numerical value of how many years the implant has been inside the patient - the time of either the patients checkup or revision surgery minus the time of implant insertion surgery.

6. Patients underwent a blood sample analysis in which four metals in the bloodstream were measured. Higher measures of these metals in the blood samples indicates that the implant is wearing down and its particles are leaking into the patients bloodstream.

   (a) **Cr** is an abbreviation of the metal chromium.

   (b) **Co** is an abbreviation of the metal cobalt.

   (c) **Zr** is an abbreviation of the metal zirconium.

   (d) **Ni** is an abbreviation of the metal nickel.

   (e) **Mo** is an abbreviation of the metal molybdenum.

7. Wear is measured and recorded as debris from wearing down the polyethylene used as liner between the femoral stem and acetabular cup implants can be problematic for the human body. Too much wear can lead to aseptic loosening of the prosthesis which will require a revision surgery.

   (a) **linWear** is the linear wearing of the plastic lining the implants cup, measured in millimeters.

   (b) **linWearRate** is a measure of the rate of how fast the implant wears down per year.

   (c) **volWear** represents a numerical result of a calculation based on the linear wear in the implant. When using linear wear, this data is redundant.

   (d) **volWearRate** represents a numerical result of a calculation based on the linear wear rate. When using linear wear rate, this data is redundant.

8. **Inc** is short for acetabular inclination, a measure of positioning the femoral stem component of the implant according to the angle of the cup and pelvis axis (Vanrusselt et al., 2015).

9. **Ant** is short for the anteversion which represents the acetabular component's positioning in the femural bone (Park et al., 2018).

10. **CupX** is the acetabular cups position on the X-axis in millimeters.

11. **CupY** is the acetabular cups position on the Y-axis in millimeters.

The dataset was provided in a Comma Separated Values (CSV) filetype that was parsed into a pandas dataframe in Python, an easy-to-use datatype package for Python with built-in data structure and analysis tools (PyData).

6.1.1 Dataset splitting

As a machine learning package, scikit-learn offered built-in functionality for automated generating of training and testing subsets. This functionality was found in its *model_selection* submodule.

In order for a machine learning algorithm to learn from the PARETO dataset it was split into two parts, training and testing. Scikit-learns splitting functionality requires a feature (or data column) to base the split on. In the case of this project the feature chosen was ***years in vivo***.

Scikit-learns train_test_split

The built in split function can take a variety of data types. In this particular project pandas dataframes were used, primarily due to the build in functions of replacing missing values, removing all samples with missing values and the straightforwardness of mutating the dataframes.

This function was used to split the PARETO dataset into a training subset that consisted of 85% samples while the remaining 15% were used for the testing subset. This split occurred every time the prediction function was called by the system. During calibration of Decision Trees a *random_state* value was passed to this function for consistent results. When the regression model from the user interface called for the dataset split, the function returned a random split for the dataset.

Leave-One-Out Cross-Validation

While Leave-One-Out cross-validation can be computationally extensive it presented no problem for the PARETO dataset due to its limitation on sample size. LOOCV was conducted in two parts. First on the control group subset that consisted of 17 samples, then on the complete dataset that consisted of 49 samples.

6.1.2 Predicting Continuous Longevity Values from PARETO Dataset

Feature Selection

While scikit-learn offers many methods for automated feature selection none were used in this project. Instead of having automated the process, two experts in the specific field were consulted with regards to which features of the dataset to use. The recommendations for features to use were **cobalt**, **chromium**, **linear wear**, **linear wear rate**, **gender**, **inclination**, and **anteversion**.

Classification and Regression Trees

Scikit-learns Decision Tree Regressor machine learning algorithm was initially used for estimating implant longevity for the end users, based on what the patient information the user fed the system. Grid Search Cross-Validation was used for tuning the hyperparameters available for scikit-learns Decision Tree Regressor. A set of valid hyperparameters such as maximum tree depth, criteria for measuring node split quality, minimum samples per split and so on were passed into the cross-validator, each with their own subset of values. The validator performed an exhaustive search across the grid of values, generated a regression model for each hyperparameter value and measured the performance of the model on the training and testing datasets using the R-squared metric. The estimator with the highest R-squared was returned from the function to be used in the system.

Ordinary Least Squares Linear Regression

Scikit-learns Linear Regression algorithm was used as a generalized linear model for both simple linear regression as well as multiple linear regression. In contrast to scikit-learns decision trees, the linear regression submodule offers no hyperparameter tuning. Instead of this the submodule offers whether or not to calculate an intercept for a given models fit, and whether or not to normalize the data by subtracting the mean and dividing it by the l2-norm (Scikit-learn, d). Among the two available options calculation of intercept was enabled while normalization was disabled due to producing no discernible difference in results.

6.2 Results

This section contains the results of using Multiple Linear Regression. Calculation of leave-one-out cross-validation R^2 values were conducted by passing two lists to an R^2 calculation function found in scikit-learns *metrics* submodule. These two lists were one that contained the true prosthesis longevity values and one that contained the predicted values.

6.2.1 Leave-One-Out Cross Validation

Testing was done using Leave-One-Out Cross-Validation for splitting the dataset and comparatively estimating the longevity of each sample using two different regression models: the

Table 6.1: Leave One Out Cross-Validation with Decision Tree Regression model estimation results from the control group subset (n = 17) and the entire dataset (n = 49).

n = 17	n = 49
Actual mean longevity in years: 9.4702 Predicted mean longevity in years: 10.1753 R^2: −0.2633	Actual mean longevity in years: 9.3413 Predicted mean longevity in years: 8.9935 R^2: −0.0233

Table 6.2: Decision Tree Regression results for actual longevity and predicted longevity in every single sample using Leave-One-Out cross-validation on the control group subset.

ID	Actual	Predicted
1	10.19	10.16
2	10.49	10.15
3	10.02	10.2
4	7.11	10.2
5	7.58	10.2
6	10.33	10.15
7	10.8	10.15
8	10.28	10.15
9	10.26	10.15
10	10.11	10.2
11	10.21	10.15
12	10.08	10.2
13	5.72	10.2
14	10.24	10.15
15	9.83	10.2
16	10.23	10.15
17	7.51	10.2

Decision Tree Regression model and the Multiple Linear Regression model, both acquired through the use of scikit-learn.

Decision Tree Regression

The average results from performing Leave One Out cross-validation with Decision Tree regression on the dataset is available in Table 6.1.

When the control group subset was used the estimator generated one of two predictions for a given sample, resulting in an average not far off from the average longevity found in the subset. While the estimator was somewhat close on the average longevity, by only generating two different values for the dependent variable the predictions were often off by many years on samples compared to their true value. In Table 6.2 it is apparent that the regressor estimates a number close to the statistical average for all samples.

Table 6.3: Decision Tree Regression results for actual longevity and predicted longevity in every single sample using Leave-One-Out cross-validation on the entire dataset.

ID	Actual	Predicted
1	10.19	10.28
2	10.49	7.11
3	10.02	7.58
4	7.11	6.74
5	7.58	7.73
6	10.33	10.26
7	10.8	7.58
8	10.28	10.26
9	10.26	10.28
10	10.11	6.74
11	10.21	7.58
12	10.08	10.28
13	5.72	6.84
14	10.24	10.28
15	9.83	7.65
16	10.23	17.87
17	7.51	10.26
18	8.08	10.28
19	6.39	7.73
20	11.52	10.26
21	7.73	10.28
22	3.32	10.28
23	7.73	7.58
24	10.94	10.26
25	9.47	7.58
26	7.82	6.74
27	6.02	6.84
28	11.29	10.26
29	6.41	6.84
30	7.2	7.73
31	4.31	6.84
32	3.25	7.73
33	7.28	10.28
34	12.02	10.26
35	11.04	6.74
36	11.9	10.26
37	7.89	10.28
38	9.03	10.28
39	12.07	7.58
40	9.51	7.11
41	8.11	10.26
42	6.84	6.74
43	15.14	10.26
44	11.38	10.26
45	12.93	10.26
46	6.74	6.84
47	12.31	10.26
48	17.87	10.33
49	13.18	10.26

Performing the same operation on the entire dataset with all 49 samples produced better results. Average prediction results was closer to the average true values of the dataset than the control group subset was. The estimator predicted a wider variety of longevity values for the samples with better accuracy on some samples as seen in Table 6.3, with an increased R^2 score compared to the control group subset. Samples 8, 9, 23 and 42 among others were close to that samples true value, yet 22, 32, 39, 43 and 48 estimated a value that was highly inaccurate.

Multiple Linear Regression

Multiple Linear Regression was applied on the data using the dataset features recommended by a scientist at the biochemical lab that provided the dataset. These two methods were given an R^2 measuring at -0.7752 meaning that the independent variables do not explain

Table 6.4: Leave One Out Cross-Validation with Multiple Linear Regression model estimation results from the control group subset ($n = 17$) and the entire dataset ($n = 49$).

n = 17	n = 49
Actual mean longevity in years: 9.4702 Predicted mean longevity in years: 9.1880 R^2: −0.7752	Actual mean longevity in years: 9.3413 Predicted mean longevity in years: 9.3476 R^2: 0.0441

Table 6.5: Multiple Linear Regression results for actual longevity and predicted longevity in every single sample using Leave-One-Out cross-validation on the control group subset.

ID	Actual	Predicted
1	10.19	11.41
2	10.49	10.78
3	10.02	9.66
4	7.11	8.98
5	7.58	9.33
6	10.33	10.58
7	10.8	10.25
8	10.28	10.44
9	10.26	9.53
10	10.11	7.89
11	10.21	9.79
12	10.08	9.56
13	5.72	-1.03
14	10.24	9.67
15	9.83	10.36
16	10.23	10.12
17	7.51	8.87

the longevity of the implant very well. However, when increasing the sample amounts to include the complete dataset the R^2 value was significantly increased to measure at 0.0441. Additionally bringing a greater amount of samples for the machine learning algorithm gave an output of average longevity estimation that was closer to the average of the actual longevity recorded for each patient as seen in Table 6.4, signifying that simply having more samples can increase the performance of the regression model.

All implant longevity predictions are available in the tables below. Table 6.5 contains the control group samples, Table 6.6 contains the full dataset.

Repeated Loops for Random Training and Testing Splits

A manual leave-one-out cross-validation was performed using the same final approach as the user-centered predictions used. The results are available in Table 6.7. The average adjusted R^2 for these estimations were 0.719. Every estimation sample produced its own ad-

Table 6.6: Multiple Linear Regression results for actual longevity and predicted longevity in every single sample using Leave-One-Out cross-validation on the entire dataset.

ID	Actual	Predicted
1	10.19	8.98
2	10.49	10.42
3	10.02	9.08
4	7.11	8.76
5	7.58	9.24
6	10.33	10.89
7	10.8	9.67
8	10.28	9.88
9	10.26	9.38
10	10.11	9.25
11	10.21	9.89
12	10.08	10.2
13	5.72	7.0
14	10.24	9.75
15	9.83	10.2
16	10.23	10.03
17	7.51	14.51
18	8.08	4.42
19	6.39	8.45
20	11.52	9.98
21	7.73	7.96
22	3.32	9.17
23	7.73	8.5
24	10.94	11.01
25	9.47	7.15
26	7.82	8.29
27	6.02	7.33
28	11.29	8.81
29	6.41	8.08
30	7.2	8.7
31	4.31	9.45
32	3.25	1.81
33	7.28	10.83
34	12.02	8.56
35	11.04	6.22
36	11.9	9.39
37	7.89	10.71
38	9.03	9.73
39	12.07	13.67
40	9.51	9.9
41	8.11	9.68
42	6.84	9.57
43	15.14	8.96
44	11.38	8.51
45	12.93	9.28
46	6.74	10.43
47	12.31	12.83
48	17.87	10.84
49	13.18	12.71

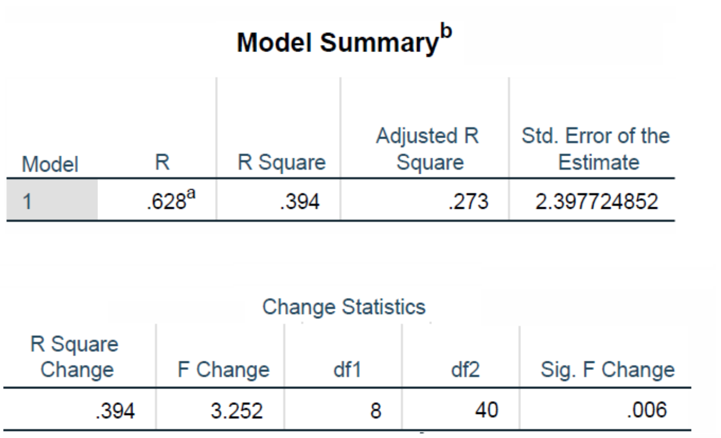
Table 6.7: Manual LOOCV and 1000 runs per sample for best R2

ID	Actual	Predicted
1	10.19	7.88
2	10.49	10.61
3	10.02	8.49
4	7.11	11.16
5	7.58	8.98
6	10.33	10.65
7	10.8	8.08
8	10.28	1.63
9	10.26	9.57
10	10.11	9.22
11	10.21	10.45
12	10.08	8.99
13	5.72	6.15
14	10.24	15.0
15	9.83	9.7
16	10.23	10.2
17	7.51	14.38
18	8.08	4.94
19	6.39	6.84
20	11.52	10.32
21	7.73	6.8
22	3.32	9.65
23	7.73	9.65
24	10.94	9.77
25	9.47	5.59
26	7.82	12.57
27	6.02	5.98
28	11.29	8.99
29	6.41	6.89
30	7.2	8.03
31	4.31	13.66
32	3.25	4.42
33	7.28	11.23
34	12.02	9.31
35	11.04	2.32
36	11.9	8.69
37	7.89	10.19
38	9.03	9.86
39	12.07	14.52
40	9.51	10.34
41	8.11	9.74
42	6.84	7.4
43	15.14	5.21
44	11.38	11.21
45	12.93	6.2
46	6.74	14.38
47	12.31	13.72
48	17.87	11.69
49	13.18	12.58

justed R^2 value which in turn was added to a list of values, from which a mean value was calculated. In addition to this metric, the two lists containing the actual longevity values and predicted longevity values respectively were passed to the R^2 calculation function found in scikit-learns metrics submodule. The resulting R^2 value was -0.91 indicating a seemingly very poor fit.

6.2.2 Train_test_split

When using the built-in dataset splitting offered in scikit-learn the parameter *test_size* allowed for determining what at what percentage to split the data. Intuitively, increasing test size resulted in reduced training size. The default value set the split at 75% training / 25% testing. Through trial and error the split was set at 85% training subset and 15% testing subset. The more training data used in predicting a separate, previously unseen target sample,



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.628[a]	.394	.273	2.397724852

Change Statistics

R Square Change	F Change	df1	df2	Sig. F Change
.394	3.252	8	40	.006

Figure 6.1: A summary of the model generated by SPSSs Linear Regression functionality.

Table 6.8: Coefficients of scikit-learns multiple regression model with the highest adjusted R^2.

Regressor variable	Coefficient
Chromium	-0.498
Cobalt	0.211
Linear wear	2.934
Wear rate	-21.445
Inclination	0.020
Anteversion	-0.046
Male	-3.234
Female	-3.932
Intercept	12.020

the higher the R^2 score yielded by the regression model.

6.3 IBM SPSS Validation

This section details the results of performing the same set of operations through IBMs statistical software package as were performed through scikit-learn. The model used in these operations is shown in Figure 6.1.

6.3.1 Regressor Coefficients of Linear Models

Scikit-learns linear regression model objects allowed for retrieval of coefficients after the model had been fit to the data. This functionality was used after excluding the regressors that were not recommended by the medical expert. The resulting coefficients for that regression model can be found in Table 6.8.

Table 6.9: Coefficients of regressors from SPSS linear model.

Regressor variable	Coefficient
Chromium	-0.347
Cobalt	0.164
Linear wear	2.683
Wear rate	-21.821
Inclination	0.020
Anteversion	-0.046
Male	-3.328
Female	-3.863
Intercept	12.178

Table 6.10: P-values showing the statistical significance of each regressor variable in scikit-learns model.

Regressor	P-Value
Chromium	0.018
Cobalt	0.417
Linear wear	0.079
Wear rate	0.420
Inclination	0.296
Anteversion	0.537
Male	0.214
Female	0.903

6.3.2 Statistical Significance of Regressors

The significance of independent variables was discovered for the dependent variable implant longevity *(years in vivo)*. The resulting significance of each dependent variables correlation to prosthesis longevity is available in Table 6.10 for significance of correlation in independent variables in scikit-learns model, and Table 6.11 for significance of correlation in independent variables in SPSSs model.

Lower P-values indicate more statistical significance. The results show that when prediction was done for implant longevity the independent variables with the most significance was chromium. Linear wear of the cup plastic proves nearly significant. These regressors are highly related to implant failure as more chromium in a patients bloodstream can lead to bone death which leads to aseptic loosening of the prosthesis. These regressors statistical significance in the regression model is validated by an expert user from the bioengineering laboratory at Haukeland University Hospital. IBMs statistical software tool SPSS was used to generate a linear regression model from the same dataset, using the same dependent and independent variable and filling the empty dataset values with the mean value of each respective data column. The resulting P-values of this model

Table 6.11: P-values showing the statistical significance of each regressor variable in SPSSs model.

Regressor	Prosthesis longevity
Chromium	0.009
Cobalt	0.209
Linear wear	0.039
Wear rate	0.210
Inclination	0.148
Anteversion	0.268
Male	0.107
Female	0.452

6.3.3 SPSS Predicted Prosthesis Longevity

By adding the dataset to SPSS, the statistical package could generate a linear regression model in which columns (features) could be chosen for the dependent variable and independent variable(s) as desired. The software was used to produce a multiple linear regression model containing the same set of dependent and independent variables as the set used in the implementation of scikit-learn and its submodules. Applying the SPSS-generated regression model on the dataset an estimated longevity value for each sample was generated. The resulting values can be found in Table 6.12.

According to the model summary generated by SPSS this regression model had an adjusted R^2 score of 0.273 (as seen in Figure 6.1). This is a significantly lower adjusted R^2 score than produced by the scikit-learn implemented model. To compare the metrics for these predictions these results were also tested through the two lists of true longevity values as well as predicted values to scikit-learns R^2 calculation method. Where scikit-learns multiple regression model scored -0.91, SPSS scored a value of 0.058 for this method.

Table 6.12: SPSS Predicted prosthesis longevity compared to the actual longevity from the dataset.

ID	Actual	Predicted
1	10.19	10.28
2	10.49	10.64
3	10.02	9.4
4	7.11	9.02
5	7.58	9.25
6	10.33	10.94
7	10.8	9.76
8	10.28	10.17
9	10.26	9.79
10	10.11	10.2
11	10.21	9.97
12	10.08	10.02
13	5.72	7.42
14	10.24	10.1
15	9.83	10.71
16	10.23	10.56
17	7.51	8.99
18	8.08	5.83
19	6.39	8.56
20	11.52	10.82
21	7.73	8.12
22	3.32	9.34
23	7.73	8.3
24	10.94	10.46
25	9.47	7.22
26	7.82	8.94
27	6.02	7.84
28	11.29	9.27
29	6.41	8.61
30	7.2	8.91
31	4.31	9.34
32	3.25	4.19
33	7.28	9.33
34	12.02	9.34
35	11.04	11.28
36	11.9	9.34
37	7.89	9.34
38	9.03	9.34
39	12.07	10.95
40	9.51	10.1
41	8.11	9.74
42	6.84	10.64
43	15.14	9.34
44	11.38	9.34
45	12.93	9.34
46	6.74	9.34
47	12.31	9.34
48	17.87	9.34
49	13.18	9.34

Chapter 7

System Evaluation

This chapter explores the results of conducting the system usability methods previously detailed in chapter 3. Though the methods used were primarily focused on general usability and user-friendliness the semi-structured interviews uncovered some elements in additional aspects of the system.

7.1 Approval for Research

Permission was granted from *Norsk Senter for Forskningsdata* (NSD) to handle personal data for this research project. This was required for conducting semi-structured interviews as recording the interviews for later use constituted as handling personal data. The approval can be found in Appendix D.

7.2 Performing the Evaluation

A set of tasks were designed to expose the participants of all evaluation methods to all parts of the system. These tasks were to be completed chronologically. The tasks were used in all three usability evaluation methods in this project. The tasks were:

1. **Start the system and fill the patient information form with mock data. Save the information.**

2. **Check that the patient information is correct (according to what you entered).**

3. **Perform a prediction (based on the information from task 1).**

4. **Start over again and change some input form values. Save the new information.**

Table 7.1: Participants in semi-structured interviews.

Participant ID	Background
P1E	Orthopedic Surgeon
P2E	Biomaterial Research Group
P3E	Biomaterial Research Group

5. **Change which features (dataset columns/patient information categories) are used in the predictions before performing a new prediction.**

6. **Explore the statistics of the prediction.**

7.2.1 Semi-structured Interviews

All participants of the semi-structured interviews consented to participating in evaluation of this project. The consent form is available in Appendix A. A set of questions were prepared for the interviews (see Appendix B). The introductory questions were aimed at exploring the interviewees own perceived technological expertise and uncover the possible need for a system as the one developed in this project. The interviewees were then given the tasks and would to the best of their ability complete them with as little intervention as possible. When the interviewees completed the set of tasks the usability of the system was explored through the questions that can be found in Appendix B. These questions were intended to uncover potential user interaction flaws, particularly in regards to how the interviewees perceived the user-friendliness of the system and its workflow. After all questions were answered and possible diverging topics were explored the interviewees were asked if they had anything to add in general.

Semi-structured interviews were conducted on what qualified as the target demographic of this information system. All participants were representative of the two user groups this project focused on. The participants had backgrounds in medicine, research and arthroplasty and were involved with hip prostheses in their work. One orthopaedic surgeon for hip arthroplasty and two bioengineers agreed to participate in this evaluation, as seen in Table 7.1. Participants P2E and P3E were employed in the Biomaterial Research Group at Haukeland University Hospital.

7.2.2 Heuristic Evaluation

Four people who had a high level of technological expertise participated in the evaluation, as seen in 7.2. The participants shared the same background, being master degree students in the same field. Participants completed the aforementioned tasks before tasked with evaluating the system according to Hevners heuristics. Each participants task completion time

Table 7.2: Participants in heuristic evaluation.

Participant ID	Age	Background	Gender
P1H	24	MSc student, Information Sciences	Male
P2H	25	MSc student, Information Sciences	Male
P3H	24	MSc student, Information Sciences	Female
P4H	24	MSc student, Information Sciences	Male

Table 7.3: Participants in System Usability Scale evaluation.

Participant ID	Age	Background	Gender
P1H	24	MSc student, Information Sciences	Male
P2H	25	MSc student, Information Sciences	Male
P3H	24	MSc student, Information Sciences	Female
P4H	24	MSc student, Information Sciences	Male
P5S	25	Medical Bioengineering	Female
P6S	28	Nurse, Polyclinic	Female
P7S	39	Surgeon	Male
P8S	26	Law Degree	Female
P9S	28	IT Product Sales	Male

was recorded, the recorded time is available in Table 7.4 (participants P1H through P4H).

7.2.3 System Usability Scale

A set of 9 people with various backgrounds and levels of technological expertise participated in the evaluation, as seen in Table 7.3. As the number of participants required to accurately evaluate the usability of a system is 8 to 12 the number of participants for this project validate the results. Time completion task was also recorded for this evaluation method.

Four of the participants (participant 1 through 4) completed the system usability scale evaluation in parallel with the heuristic evaluation. Participants in both evaluations filled out the system usability scale form before completing the heuristic evaluation.

The participants were instructed to fill out the evaluation forms (see Appendix C) as quickly and accurately according to their experience with the system as they could, directly after they completed the list of tasks. Each participants individual time per task and the average time per task can be found in Table 7.4.

Table 7.4: Task completion time for each participant

Participant ID	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6
P1	01:10	00:06	00:28	00:51	00:39	00:38
P2	00:59	00:07	00:30	00:28	00:35	00:29
P3	02:24	00:11	00:29	00:41	00:43	00:33
P4	01:53	00:10	00:31	00:39	01:19	00:34
P5	01:36	00:08	00:28	01:12	00:56	00:13
P6	03:23	00:20	00:27	01:29	01:54	01:01
P7	03:11	00:08	00:33	00:47	00:43	00:26
P8	01:31	00:09	00:28	00:55	00:51	00:25
P9	01:24	00:06	00:31	01:16	00:44	00:26
Average	02:07	0:09	00:29	00:56	00:56	00:32

7.3 Evaluation Results

7.3.1 Semi-structured Interviews

Three participants from the two user groups were interviewed in order to evaluate the developed system. Two participants were researchers at the bioengineering laboratory at Haukeland University Hospital, responsible for providing the PARETO dataset. The last participant was an orthopedic surgeon at another hospital in Hordaland, whose area of specialization was hip arthroplasty.

Participant one

The first interviewee gave ample feedback to the entirety of the system, not just its usability. As an orthopedic surgeon P1 had extensive knowledge and experience with hip arthroplasty and the patient information that was at the time of this project stored during each surgery.

Feedback on functionality and features

- Most of the data in the PARETO dataset was either unknown to the participant or not relevant from his perspective.
- While the PARETO data was interesting, it does not contain many of the variables that are considered highly important when estimating prosthesis longevity and whether or not a patient should undergo surgery.

- A prediction of implant longevity could be useful, but the participant would also like to see the system decide whether or not a patient should be operated on using classification techniques on the data from the National Arthroplasty registry.

User interface and system usability feedback

- While entereing patient information the interviewee would like an explanation as to what the data required is, both its medicinal context as well as descriptive information like measurements of the input values and what values are normal for any given patient.
- Most of the statistical information regarding the performance of the regression model was unclear to the interviewee. They had not heard of the terms adjusted R^2 and root mean squared error.
- The interviewee found the additional segregated step of feature selection for the model unnecessary. The interviewee would rather have moved the checkboxes for each feature into the patient information form.
- The interviewee noted the lack of a help section in the system.

Some observations were made during the task completion process. The interviewee was clearly well versed in the use of various information systems and had had little qualms with navigating through the system. The interviewee dismissed the contextual descriptions found throughout the system, both headers and their accompanying texts. It seemed like the interviewee may have underestimated their helpfulness or had a habit of just clicking the buttons and learning by doing rather than fully understanding the system, or maybe relied on the presence of interactive contextual help section (such as a button labeled with a question mark). This led to some confusion during the task completion stage, in which the interviewee required a small amount of assistance.

Participant two

The second interviewee provided a thorough evaluation of both usability and functionality.

Feedback on functionality and features

- The required steps to complete a prediction was satisfactory, however the interviewee would have liked to see a different approach to feature selection. The interviewee suggested either adding checkboxes next to the respective input features during the step in which the user has to enter the patient information, or an integration of automated feature selection based on inputs from the user in which the form fields left blank would be discarded from the regression model.

- The interviewee noted that while restrictions were set on the input variables in form of input type, a set of restrictions for input values would improve the system. A user could enter patient information that is not only exceeding the extremes of that value (such as recording 5000 millimeters of linear wear) but also the impossible (such as recording a negative linear wear value). Suggestions were to inform the user of manual restrictions by adding a description for each input field detailing the allowed values. This would restrict the systems ability to make wrongful predictions.

User interface and system usability feedback

- The interviewee was overall pleased with the user interface and design of the system, noting that it looked *quite nice*. Additionally the system was perceived as simple yet clean and functional.
- The statistics that described both prosthesis longevity estimations and the model behind it had too many decimals. Additionally the statistics themselves had no context in terms of relating to the model. The interviewee would like to have some further description of these statistical elements, the interviewee suggested having an information button next to the results that could expand when clicked and present the user with a more descriptive explanation as to what the numbers meant.
- A lack of statistical result metrics were noted. The interviewee suggested adding units to the numbers. An example introduced was the standard deviation of all 2300 estimations calculated by the system during the task completion. This deviation was presented as a floating point decimal and the interviewee felt that adding the unit "years" behind the decimal would greatly clarify the context.
- In relation to statistical results, the interviewee pointed out that there should be a section in the expanded information that tells the user what the prediction is based on, which machine learning model produced the results and some description of the model itself.

Participant three

The third and last interviewee reported a modest 7 out of 10 for their estimation of self-perceived technological prowess. Out of the three participants in this evaluation method, this was the only interviewee who took their time with reading the descriptive contextual text in the system. However, after completing task 2 this interviewee stopped reading the context.

Feedback on functionality and features

- Having background from the biochemical research group responsible for the PARETO dataset, the interviewee provided insight into what the data means and how it should be used.
  - The dataset features molybdenum, cup X and cup Y should be removed from the system.
  - When choosing regressors a single checkbox item should be listed for gender instead of having two separate items for each gender. This one checkbox would enable or disable both male and female features.
  - During the prediction process the regression model could predict a negative value for prosthesis longevity. The interviewee suggested removing all negative predictions from the displayed results, and if all results were negative the results should display a message to reassess the situation for the patient or the values given in the input form.
  - Same as the prior evaluation participant, the interviewee would like some restrictions for input values in the patient information form. Suggestions were disabling or highlighting incorrect values or providing information on valid value range.
  - The interviewee noted that while the system is restricted to a small database where all samples have worn the same prosthesis type, functionality to select which prosthesis type to be used or has been used would be appreciated.

User interface and system usability feedback

- The interviewee noted that the system could be somewhat technical for surgeon standards.
- Some paragraphs in the descriptive text could be rephrased, an example was that the introductory explanation of the longevity prediction could be interpreted as being the total remaining longevity of the patient. The phrasing of the description of *case* (whether the patient has removed their prosthesis or not) could also be improved.
- The interviewee noted that there were no metrics being used throughout the system and suggested were given. Measurements or units should accompany all input fields, whether in-line with the input field or as a documentation element that can be displayed at the users leisure either through hovering the input fields or having a help

section nearby. Degrees should be used for inclination and anteversion. Milligrams per liter should be used for all blood sample values.

- The interviewee commented that some grouping of input fields could be helpful, such as having some visual indication that the input fields for blood sample analysis were one category and the degrees of inclination and anteversion another.
- The metals from blood samples used in the PARETO dataset were measured in milligrams per liter while (according to the interviewee) most doctors were used to measuring nanograms per liter. Additionally some information should be given that the blood samples need to be measured from a whole blood sample.
- While entering the mock patient information the interviewee was uncertain as to which separation symbol to use for decimal inputs. The HTML5 input element restricts the format to using period instead of comma, but having both would be preferrable to increase usability.
- The interviewee completed the second evaluation task swiftly but noted that the patient information display contained more information than what the interviewee was asked for. This was due to two binary input fields on whether the cup or the stem of the prosthesis had come loose in the event of a revision surgery being dependant on actually having the revision surgery (case = 1). The interviewee did not choose yes for that input field and thus never saw the two subsequent fields. Suggestions for this was to remove them from the patient information display or to include them in the initial input form.
- When tasked with going back to the start of the system process to edit the patient information form the interviewee pressed the back button twice instead of pressing the reset button. This was reportedly due to a concern that the original patient information that was already entered would be reset, leaving the interviewee to enter all the information from scratch. As the system was designed to maintain the information throughout possible use cases, a suggestion was given to rephrase the reset button to something along the lines of "Go back to patient information form".

7.3.2 Heuristic Evaluation

The participants were exposed to the system using the aforementioned tasks. Immediately after task completion they were asked for their subjective perception of the system according to Hevners ten heuristics (see Table 3.2), provided in an unformal discursive manner.

Visibility of System Status.

All participants were in general content with the information presented in the system. They all were happy with the headers displaying which part of the system they were working on. Two participants noted that the interactive buttons gave appropriate feedback and that they were never unsure whether something was happening or not.

Match between system and the real world.

Every participant of this evaluation noted that there was an amount of medical terms in the system that they were not familiar with. Additionally the statistical measures describing the model caused some confusion. As for the system documentation available they were all happy with how the system described its processes, two of the participants noted that use of natural language was well executed.

User control and freedom.

All participants praised the system for its use of its navigational button. They liked that no matter where in the process they were they could always return to the start of the process. One participant noted that the *restart* button should delete the previously saved patient information.

Consistency and standards.

Every participant were happy with the consistency of the system, that the interactive elements of the system that progresses through the process are equally sized and colored. Two participants noted that the navigational buttons could be based on well-established standards from other systems (namely iOS and Windows), thereby having no issues discerning their intended use.

Error prevention.

One participant commended the system for its unseen error preventions regarding user input and process cancellation. Another participant noted that the restart button should have a confirmation dialog when clicked to avoid accidental emergency exits.

Recognition rather than recall.

All participants commend the system for its capabilities in displaying the patient information entered earlier in the process. One participant noted that chosen regressors for the machine learning process are not displayed at any point except when choosing between them, and that they would like to see that along with the patient information. Another participant noted that if the patient information display was toggled to visible, nothing (except the display toggle button) in the system would change its visibility no matter what the participant did.

Flexibility and efficiency of use.

Every participant noted that they could not think of any process being accelerated due to the simple flow in the system. One participant mentioned using the tab key to switch between input fields.

Aesthetic and minimalist design.

The general consensus among the participants was that the design was minimalist with simple and elegant aesthetics directed at contrasting areas and maintaining focus in the center field. One participant thought the text accompanying each step of the workflow, explaining what the system does in that part was unnecessary to display at all times. Additionally that while the text did clutter the design it did so in a subtle, unrestrictive manner, but the participant would rather that it was hidden and only displayed when necessary.

Help users recognize, diagnose, and recover from errors.

All participants noted that the error message was constructive in suggesting what needed to be done to fix a problem if it occurred and that it did so in an adequately natural language. However they all thought that the error message did not specify exactly what caused the error.

Help and documentation

Every participant desired some clarification on the medicinal aspects of the input fields, as well as the statistical background for the machine learning model and its prediction metrics. Three participants would like to see a hover-for-description or similar solution in the input field. One participant noted that the descriptive paragraphs detailing the current parts of the system flow should be initially hidden, or rather moved into a help section detailing the entire process and each step to achieve a result from the system. This would declutter its design and present new users with knowledge of the total requirements to complete a prediction.

Table 7.5: Each participants individual calculated SUS score

Participant ID	SUS Score
P1	60.0
P2	82.5
P3	90.0
P4	90.0
P5	85.0
P6	45.0
P7	72.5
P8	50.0
P9	65.0
Average	71.1

7.3.3 System Usability Scale

Results were somewhat varied across the board of participants, as seen in Table 7.5. Most of the participants reported after completing the evaluation that they assumed they were supposed to take the role of an expert user, that they would be expected to know more about the medical and statistical aspects of the system. A select few participants (notably P1, P6 and P8) gave poor scores for whether they would use the system often, needed technical help to use the system and felt confident using the system. Because of the disparity between participants' assumptions the resulting scores are somewhat skewed.

One participant from the target demographic for the developed system (P7) evaluated the system with a score of 72.5. This indicates that the system falls within the bounds of acceptable usability yet implies that improvement can be made for this specific person. Although loosely related to this project, another participant (P5) who had a background in biochemistry and was at the time of this research employed at a hospital in Norway evaluated the system with a score of 85.0.

Chapter 8

Discussion

In this master project we have explored idea of using easily available machine learning methods to solve a relevant clinical problem which to predict an orthopedic prothesis longevity. This is a question that interests several expert groups. The request came from the Laboratory of Biomedical Engineering at the Haukeland University Hospital which analyses explanted devices. The same question is of vital interest for treating surgeons who want to implant the most suitable prothesis that will last and improve patients' life quality. Although the interest in the device longevity is very important for both the expert groups, they use a different approach and data to estimate it. In our work we have looked at how to predict longevity using data mining on the biomedical engineering group's database. We have built a system that delivers both individual and group predications using xx software. We have validated methods using SPSS statistical package. The whole system was developed using design science approach. We are discussing in this chapter the most important issues that occurred.

8.1 Dataset Restrictions

The database was not large, which would be expected in a relatively newly started data mining project. However, this data is representative of the research in the field. It is in the initial phase and user group would to start with data mining from the beginning to avoid later data migration from diverse systems into one database. So the motivation was to start building a system even if the data size was limited. That way users would be engaged from the beginning which in turn would contribute to the system adaption.

8.2 Methods and Methodologies

Among all the possible methods we have chosen are two data mining approaches. One was multiple regression analysis and the other was an optimized classification and regression tree (see Section 3.1.4 and 2.6.1), both applied through the use of scikit-learn, an open-source machine learning framework that provides simplified implementation of methods that are easy to manage even by a novice developer. The idea was to explore two general approaches capable of predicting a continuous value, whose resulting predictions could be then be compared and evaluated. During the course of this master project tuning and cross-validation of the best set of hyperparameters for the particular biomedical dataset used in this project has been conducted on both regression models for continuous value prediction as seen in Section 6.2. Use of these hyperparameters for both regression models has been carried out on both single sample prediction as well as prediction on all samples available (see tables in Section 6.2.1), resulting in multiple linear regression performing better than decision tree regression for this set of data, this was more appealing to the user group. This led to multiple linear regression serving as the primary regression model for predicting a single samples longevity in the user-centered part of the HALE system.

To validate the methods we opted to use the SPSS statistical package as discussed in Section 3.2. This was essential to achieve comparative results in terms of prediction (calculated longevity) and statistical significance when applicable. Given the database size we had no expectation to have predictions that would hold for the whole patient population. However, we needed to be sure that the development we have done is understandable and replicable. That is why we validated the results of scikit-learns regression models against a well-established method, which is a commonly seen approach in design science.

Results are presented in the form that was easy for the users to understand. Moreover, two main sets of results were delivered: a single case prediction and the complete dataset samples prediction. That is in line with the established way of looking at the data which was appealing to the user group and even surgeons who evaluated the system. The open-source scikit-learn offers data mining solutions that ought to be mention since the user might want to expand the machine learning capabilities of the HALE system with functionalities such as discrimination and clustering to name just a few principal methods. That also means that when machine learning methods are learned and established, they can easily be applied on the same dataset.

Since the user and their understanding and satisfaction were important, we had to consider several ways of evaluating how these chosen machine learning procedures appeal to the user. That is why heuristic and system usability scales were used in addition to the semi-structured interviews. Experts of the two different user groups have provided valuable comments and critique that not only identified problems, but gave constructive feedback that can inform for future development. The potential of data mining was clear to them and they came with new ideas and request to include additional data and develop more applications.

8.2.1 Design Science

Design Science (see Section 3.3) is a powerful framework that provides seven guidelines which are instrumental for conducting research and developing solid scientific artifacts. Following all the Hevner's seven guidelines as basis for this research, resulting is the HALE system that was methodically evaluated by potential users and IT experts. In addition, the machine learning part was validated using the well-established statistical package. The resulting artifact (the HALE system) has been instrumental in bringing across the potential of data mining for total hip arthroplasty. The evaluation has shown that the artifact could be easily managed and operated on the real-world data in a novel way as it would be a sought outcome within Design Science research. An implemented system even as a prototype provides a unique user experience and feedback that could be hard to obtain in any other way. Results are given as the answers to the research questions relevant for this research. This concerns advantages of using here developed regression models to predict the longevity of the arthroplasty implants. The main user groups to utilize the results are biomedical engineers and surgeons with patients for whom the surgeries are carried out. Design Science is therefore providing a great framework to conducting research in a systematic way to provide results (artifacts) that users can interpret and understand, while not being pressured to have prior knowledge and understanding of all the underlying methods.

8.2.2 Personal Extreme Programming

Many agile system development methods are designed for teams, but they often used by single developers. They provide a sense of progress, control; they are helpful to document the main development steps. The development in this project has followed the principles of personal extreme programming. The advantage is the iterative structure allowing for continues improvements and a sense of retrospective evaluation. The demanding part is that the planning builds on the prior knowledge gathered from the previous projects which could be seen as a disadvantage when the developer is novice and cannot reflect on prior experience from previous projects. Regardless, this methodology could be recommended to novice developers as its advantages outweighs its disadvantages.

8.2.3 Usability Evaluation

Usage of different usability evaluation methods proved fruitful in uncovering both positively and negatively perceived elements in the developed system. All three system evaluation methods were conducted with a satisfactory number of participants. While chiefly directed at system usability, the semi-structured interviews with experts resulted in feedback concerning improvements and new ideas for system functionality to be implemented in further iterations of the system.

System Usability Scale was obtained from the homogenous user group that included only a third of healthcare personnel. This was done with the intention to test the systems usability as broadly as possible which might explain a certain skewing of results. However, the evaluation has resulted in many useful comments and provided a possibility to observe new users executing the set of tasks.

8.3 Machine Learning Model Performance

8.3.1 Linear Regression Model

The project development started with implementing a regression based decision tree but the model was outperformed by the multiple linear regression model, which was validated by using IBMs SPSS statistical package. This comparison depended on calculation of R^2 and adjusted R^2 that are generally accepted as a standard goodness-of-fit metric.

8.3.1.1 Evaluation Metrics

Further evaluation metrics for each machine learning model should be considered. While R^2 and adjusted R^2 are generally accepted as standards for goodness-of-fit, some statisticians consider the use of them incomplete in regards to total model performance and have suggested conjoined use of multiple metrics to get a better overall idea of the models performance (Stone et al., 2013). For the evaluation of HALEs implemented models adjusted R^2 was used in conjunction with root mean squared error, as well as displaying the standard deviation of multiple (2300) predictions resulting prosthesis longevity. This can give a fair idea of prediction performance and accuracy for this system.

8.3.1.2 Significance of Regressors

A distinct disparity between the calculated statistical significance of regressors in SPSS and scikit-learn is evident. While the cause of this disparity is unknown, differences in method of calculation is suspected. Despite being a widely used statistical tool, no specifics on how SPSS computes the statistical significance of each regressor could be found. Comparatively scikit-learns p-value calculations are relatively similar to those of SPSS, but each independent variable calculated from this method had almost exactly double the value compared to SPSSs p-values for the same independent variables.

8.4 State of System

8.4.1 Feature Selection

The features of the dataset (also known as regressors, or independent variables) were chosen based on recommendations from an expert user from the bioengineering laboratory user group. One of the features that held the highest impact on prosthesis longevity prediction was *linear wear*. This feature can only be recorded after the prosthesis has been implanted and been worn for some time, rendering it useless for estimating how long a prosthesis will last in a patient that has not yet undergone surgery. Similarly, there are two other features, *chromium* and *cobolt*, whose values were measured by analyzing blood samples that contained traces of the prosthesis' metals.

8.4.2 Predict Future Cases

The current set of variables might not be the most optimal ones for the prediction since it is based only on retrospective data of explanted prostheses. Additional information would come from the patient records where other clinical and medical patient parameters would be included in the prediction. That is something that has to be addressed in the development of this project.

8.5 Answering the Research Questions

As mentioned in Section 1.2 the research conducted in this master thesis project aimed to answer three research questions.

- **RQ1:** Is it possible to develop a highly usable longevity prediction module of hip arthroplasty implants based on a biomedical dataset?

Yes. The developed prototype HALE was based on a small biomedical database produced by and retrieved from the Biomatlab Research Group at the Orthopedic Clinic, Haukeland University Hospital. HALE can produce estimations on prosthesis longevity for any single patient based on a selected combination of data features. By using established methods, and methodologies for software development and usability evaluation, the HALE system has proven usable to a satisfactory degree. Overall impression was that the prototype was simple, clean and with a straightforward workflow. This had appeal for new users. User evaluation has identified weaknesses such as lack of measurement units and a limited presentation of the feature selection. Attending to these issues would make the system more appealing to the users.

- **RQ2** Can this module produce reliable predictions that are equivalent to the one produced by a well-known, validated statistical module?

Yes. Although trained on a distinctively small dataset, the HALE system produced prosthesis longevity outcomes that were reasonably good. For example, when used on the complete dataset 20 of the 49 predictions were within the range of the actual longevity ±1 year. Indications of model overfitting is present through the high adjusted R^2 scores, an issue that could be resolved by adding dataset variables and samples. Machine learning techniques have proven prone to overfitting on lesser datasets and the data provided for this project can only be described as small.

Performance of different regression models was compared to IBMs SPSS statistical package. Results show that the multiple linear regression model was of comparable performance while the regression based decision tree could not really measure up. Both models produce a highly similar set of coefficients for each regressor in the model as well as a similar intercept value. The greatest disparity between the linear regression models were the resulting statistical significance p-values. The p-values HALE produced through scikit-learn were approximately double the p-values reported in the model summary produced by SPSS. These results are rather satisfying as they can inform the future development.

- **RQ3** Are there any guidelines regarding machine learning that could be suggested to software developers that use scikit-learn, an open-source machine learning framework?

Yes. When developing a system that utilizes data mining and machine learning techniques, the developer should always explore the problem space and applicable theories. Consulting experts in whichever field the data comes from, as well as seeking knowledge about the methods, should be one of the first steps in development. In this case it was important to understand what is the impact of applying regression analysis, and what can cause poor outcomes such as poor model fit.

A general guideline that was observed in many tutorials, guides and explanations is that data visualization can be a key tool in understanding the data. A developer working on a machine learning tool should take their time to plot the data into graphs while developing the system. Automating the process of generating various graphs using the data at hand can prove a valuable tool that can help understand the data better, and display trends and relationships with a single glance rather than working through a large set of values.

User-friendly interface as well as elements providing help and documentation of results are important for the system to appeal to users.

Chapter 9

Conclusions and Recommendations for Future Work

9.1 Conclusions

This thesis has explored possibilities of using machine learning to answer important clinical questions such as longevity of total hip arthroplasty implants. This is the question that the Biomatlab Research Group at Haukeland University Hospitals orthopedic clinic and orthopedic surgeons are approaching using patient and biomedical data. Both these experts can be seen as the user groups that could utilize the developed *HALE* system for their routine work. User requirements suggested two different prediction cases: for individual patients and for the collective groups with a final outcome expressed as years of longevity.

User assessments has indicated that the system was appealing in the terms of functionality and easiness of navigation. In its appearance the system is rather technical and provides individual predictions or tables, both commonly seen in any other statistical package.

The open-source system scikit-learn was used to implement the machine learning components. Two regression models were applied, multiple linear regression and C&RT decision trees. They proved to be highly similar to that of IBMs SPSS software which was used to validate scikit-learns machine learning modules. The performance of the models were comparably good and similar in structure. The advantage of scikit-learn is that it was manageable and easy to use even by a novice developer. Another long term advantage is that additional machine learning procedures can easily be added to the system.

Design science has proven to be a good framework for development and has given a functional artifact that could be evaluated. This has given the potential users a hand-on experience and gained a trust for the future use. The finding suggests that there is an actual need for this kind of machine learning in the clinical practice and research.

9.2 Future work

This section details work that would follow in the future. In the first place we would improve certain features such as the general usability of the system with addition of more helpful elements for the user such as toggleable descriptions for each required element the user is exposed to. The database could be expanded towards the clinical side, for which there is interest from surgeons. This would require work with adding variables and different kinds of machine learning procedures, requiring creation of new user interfaces in some places. With growing data there might be a need for discriminant analysis and cluster analysis which are available in the scikit-learn framework. User involvement would be important to define new tasks and new evaluations would need to be conducted. However this should not a difficult extension of work since data science framework and the tools enable feasibility.

Novice users would need to be trained and made aware of risks connected to development using scikit-learn. The future development should include tips and help functions that would inform the users of overfitting or underfitting of the models, lack of significance and meaningful outcomes. This way easiness of development could be fully utilized whilst minimizing risks of obtaining potentially misleading outcomes. We have given some idea of that in the guidelines for the future users as detailed in the answer to the third research question 8.5.

9.2.1 Machine Learning, Data Handling and Improvements

For the machine learning models and the data processing components used in HALE room for improvement is present. Many available regression models can be implemented and tested, as detailed below.

Regression Models

Due to the modularized nature of scikit-learn and Python usage of a good many regression models can be implemented in the future. Several machine learning methods from both supervised and unsupervised techniques, such as multi-layer perceptron, naive Bayes, gaussian mixture models and support vector machines can be implemented and explored. Comparative evaluation can be conducted on the different models to determine which model suits the data currently available best, or be made available to users to choose from should they deem it necessary.

Data Processing

Scikit-learn offers an expansive set of tools for data preprocessing that can be used in the HALE system. While standardization of regressors was tested and yielded no discernible differences in the results, more preprocessing tools such as min-max-scaling could prove beneficial to the regression models. Additionally, the provided PARETO dataset contained many missing values that were replaced with the mean value of each specific column, a common approach for machine learning. In the future these can instead be predicted with a machine learning model based on all samples whose variables are complete.

Expansion of Data

The current dataset is as previously mentioned very small. Additionally the regressors than can be utilized for predicting the longevity of a prosthesis before the primary surgery are extremely limited and do not contribute sufficiently to the estimated longevity - the regression model leans too heavily towards the mean longevity present in the dataset. For future work an effort to expand this dataset in both variables and samples would be highly beneficial to the HALE system in its current state as well as any future iterations.

Bibliography

Abu-Amer, Y., Darwech, I., and Clohisy, J. C. (2007). Aseptic loosening of total joint replacements: mechanisms underlying osteolysis and potential therapies.

Aldrich, J. (1995). Correlations Genuine and Spurious in Pearson and Yule. *Statistical Science*, 10(4):364–376.

Bangor, A., Kortum, P., and Miller, J. (2009). Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. Technical report.

Bermingham, M. L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., Wright, A. F., Wilson, J. F., Agakov, F., Navarro, P., and Haley, C. S. (2015). Application of high-dimensional feature selection: Evaluation for genomic prediction in man. *Scientific Reports*, 5(1):10312.

Bremer, M. (2012). Math 261A -Spring 2012 Multiple Linear Regression.

Brien, W. W., Clinical Professor of Orthopaedic Surgery, A., Rod Davey, J., Di Cesare, P., and Malchau, H. Surgical technique completed in conjunction with Hip system. Technical report.

Brooke, J. (1996). SUS - A quick and dirty usability scale. In *Usability Evaluation in Industry*, pages 4–7.

Cameron, A. C. and Windmeijer, F. A. G. An R-squared measure of goodness of fit for some common nonlinear regression models. Technical report.

Carone, G. and Costello, D. (2006). Can Europe afford to grow old? *Finance and Development*, 43(3):28–31.

Chakrabarti, S., Ester, M., Fayyad, U., and Gehrke, J. (2006). Data mining curriculum: a proposal. In *ACM SIGKDD*, pages 1–10.

Chatterjee, S. and Hadi, A. S. (2006). *Regression analysis by example.*

Claesen, M. and De Moor, B. (2015). Hyperparameter Search in Machine Learning. Technical report.

Clifton, C. (2010). Data mining | computer science | Britannica.com.

Dzhurov, Y., Krasteva, I., and Ilieva, S. (2009). Personal Extreme Programming–An Agile Process for Autonomous Developers.

Faggella, D. (2018). 7 Applications of Machine Learning in Pharma and Medicine.

Fargon, J. R. H. M. and Fischer, S. J. M. (2015). Total Hip Replacement - OrthoInfo - AAOS.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AAAI AI Magazine*, 17(3).

Flanagan, D. (2011). *JavaScript: The Definitive Guide 6th Edition.*

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques.*

Hevner, A. R., March, S. T., Park, J., and Ram, S. (2004). DESIGN SCIENCE IN INFORMATION SYSTEMS RESEARCH 1. *Design Science in IS Research MIS Quarterly*, 28(1):75.

Hewett, Baecker, Card, Carey, Gasen, Mantei, Perlman, Strong, and Verplank (2009). ACM SIGCHI Curricula for Human-Computer Interaction : 2. Definition and Overview of Human-Computer Interaction.

Hilbert, M. and López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025):60.

Humphrey, W. S. (2000). The Personal Software Process SM (PSP SM). Technical report.

jQuery Foundation. jQuery. https://jquery.com/.

Kallio, H., Pietilä, A. M., Johnson, M., and Kangasniemi, M. (2016). Systematic methodological review: developing a framework for a qualitative semi-structured interview guide.

Koehrsen, W. (2018). Automated Machine Learning Hyperparameter Tuning in Python.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Technical report.

Kuhlman, D. (2009). *A Python Book: Beginning Python, Advanced Python, and Python Exercises.*

Kurtz, S. M., Ong, K. L., Schmier, J., Mowat, F., Saleh, K., Dybvik, E., Kärrholm, J., Garellick, G., Havelin, L. I., Furnes, O., Malchau, H., and Lau, E. (2007). Future clinical and economic impact of revision total hip and knee arthroplasty. In *Journal of Bone and Joint Surgery - Series A*, volume 89, pages 144–151.

Longo, L. and Dondio, P. (2016). On the relationship between perception of usability and subjective mental workload of web interfaces. In *Proceedings - 2015 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2015*, volume 1, pages 345–352.

Ma, Y., Richards, M., Ghanem, M., Guo, Y., and Hassard, J. (2008). Air pollution monitoring and mining based on sensor Grid in London. *Sensors*, 8(6):3601–3623.

Malchau, H., Herberts, P., Eisler, T., Garellick, G., and Söderman, P. (2002). The Swedish Total Hip Replacement Register. In *Journal of Bone and Joint Surgery - Series A*, volume 84, pages 2–20.

Minitab (2018). Multiple Regression Analysis: Use Adjusted R-Squared and Predicted R-Squared to Include the Correct Number of Variables.

Mitchell, T. M. (1997). *Machine Learning*.

Nielsen, J. (1994). Usability inspection methods. In *Conference companion on Human factors in computing systems - CHI '94*, pages 413–414, Boston.

Nielsen, J. and Molich, R. (1990). Heuristic Evaluation of User Interfaces. In *Proc. CHI 1990*, pages 249–256.

Nieuwenhuijse, M. J., Nelissen, R. G., Schoones, J. W., and Sedrakyan, A. (2014). Appraisal of evidence base for introduction of new implants in hip and knee replacement: A Systematic review of five widely used device technologies.

Obermeyer, Z. and Emanuel, E. J. (2016). Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *New England Journal of Medicine*, 375(13):1216–1219.

Park, Y. S., Shin, W. C., Lee, S. M., Kwak, S. H., Bae, J. Y., and Suh, K. T. (2018). The best method for evaluating anteversion of the acetabular component after total hip arthroplasty on plain radiographs. *Journal of Orthopaedic Surgery and Research*, 13(1):66.

Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot, M., and Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Technical report.

Persson, P.-V. and Rietz, H. (2017). Predicting and Analyzing Osteoarthritis Patient Outcomes with Machine Learning.

Peters, T. PEP 20 – The Zen of Python | Python.org. https://www.python.org/dev/peps/pep-0020/#id4.

Piatetsky, G. (2013). KDnuggets Annual Software Poll:RapidMiner and R vie for first place. https://www.kdnuggets.com/2013/06/kdnuggets-annual-software-poll-rapidminer-r-vie-for-first-place.html.

Preece, J., Rogers, Y., and Sharp, H. (2015). *Interaction design*.

PyData. Python Data Analysis Library, Version 0.14.1.

Python Software Foundation (2012). The Python Standard Library — Python v3.3.0 documentation.

Quintero, D., Ancel, T., Cassie, G., Ceron, R., Darwish, A., Felix, G. G., He, J. J., Keshavamurthy, B., Makineedi, S., Nikalje, G., Pal, S., Salie, Z., and Tiwary, A. (2012). Workload Optimized Systems Tuning POWER7 for Analytics Strengthens.

Rao, C. R. C. R. (1973). *Linear Statistical Inference and its Applications*.

Reitermanová, Z. (2010). Data Splitting. *Week of Doctoral Students 2010 – Proceedings of Contributed Papers*, pages 31–36.

Rokach, L. and Maimon, O. (2008). *Data mining with decision trees : theory and applications*.

Ronacher, A. Welcome | Flask (A Python Microframework). http://flask.pocoo.org/.

Russell, S. J. and Norvig, P. (1995). Artificial Intelligence: A Modern Approach. *Neurocomputing*.

Scikit-learn. 1.1. Generalized Linear Models — scikit-learn 0.20.0 documentation. https://scikit-learn.org/stable/modules/linear_model.html#ordinary-least-squares.

Scikit-learn. 1.10. Decision Trees — scikit-learn 0.20.0 documentation. https://scikit-learn.org/stable/modules/tree.html.

Scikit-learn. sklearn.feature_selection.f_regression — scikit-learn 0.20.0 documentation.

Scikit-learn. sklearn.linear_model.LinearRegression — scikit-learn 0.20.1 documentation.

Scikit-learn. sklearn.model_selection.train_test_split — scikit-learn 0.20.0 documentation.

Seif, G. (2018). Selecting the best Machine Learning algorithm for your regression problem.

Song, Y. Y. and Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, 27(2):130–135.

Stone, B. K., Scibilia, B., Pammer, C., Steele, C., and Keller, D. (2013). Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit? http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit.

Tighe, P., Laduzenski, S., Edwards, D., Ellis, N., Boezaart, A. P., and Aygtug, H. (2011). Use of Machine Learning Theory to Predict the Need for Femoral Nerve Block Following ACL Repair. *Pain Medicine*, 12(10):1566–1575.

Torvalds, L. (2017). About - Git.

Ulrich, S. D., Seyler, T. M., Bennett, D., Delanois, R. E., Saleh, K. J., Thongtrangan, I., Kuskowski, M., Cheng, E. Y., Sharkey, P. F., Parvizi, J., Stiehl, J. B., and Mont, M. A. (2008). Total hip arthroplasties: What are the reasons for revision? *International Orthopaedics*, 32(5):597–604.

Vanrusselt, J., Vansevenant, M., Vanderschueren, G., and Vanhoenacker, F. (2015). Postoperative radiograph of the hip arthroplasty: what the radiologist should know.

World Health Organization (2005). WHO World Alliance for Patient Safety. Draft Guidelines for adverse Events Reporting and Learning Systems.

Appendix A

Informed Consent for Semi-Structured Interviews.

Do you want to participate in the research project "A System for Hip Arthroplasty Implant Longevity Estimation"?

This is a question for you to participate in a research project where the purpose is to develop a machine learning-based, user-friendly supportive system for orthopedic surgeons who can help estimate how long a hip implant will last in a given patient. In this letter we give you information about the goals of the project and what participation will involve for you.

Purpose

The project is carried out in connection with the completion of a master's thesis. The purpose is to develop an easy-to-use system that can estimate the lifetime of a hip implant by surgeons (possibly other healthcare professionals) entering patient information in the system and then presented with an estimate and additional information about the estimate. This estimate is calculated by an underlying machine learning technique where a regression model has been refined against anonymous data from previous patients. This system will offer, through a very user-friendly experience, the ability to gain insight into the patient's future and the ability to adjust variables that can lead to a longer life of the implant.

Who is responsible for the research project?
Department of Information and Media Studies at the Faculty of Social Sciences, University of Bergen

Why do you get questions about participating?
You have been chosen as a potential participant because your position as a doctor or surgeon for total hip arthroplasty is highly relevant to the use of the above system - you are the target audience for users of this system.

What does it mean for you to participate?

If you choose to participate in this project, it means that you want to interview where you will test the above system and provide feedback on the user experience. The interview is partially structured.

The interview will last for about 45 minutes. Written notes will be posted along the way. Audio from the interview will be recorded.

Volunteering is optional

It is voluntary to participate in the project. If you choose to participate, you can withdraw your consent at any time without giving any reason. All information about you will then be anonymized. It will not have any negative consequences for you if you do not want to attend or later choose to withdraw.

Your privacy - how we store and use your information

We will only use the information about you for the purposes we have described in this letter. We treat the information confidentially and in accordance with the privacy policy.

- The parties who want access to the Department of Information and Media Studies are Per-Niklas Longberg (student) and Ankica Babic (supervisor)
- All personal information about you will be stored on an encrypted USB flash drive separate from other data. This includes name list where your name will be replaced with a reference, the link between name and reference will be stored on the above-mentioned USB flash drive. Recording of interview will be saved on the same piece. Transcription of recordings is anonymized by reference.

No participants will be recognized in the publication unless they have approved the use of names in the assignment. All personal information is replaced by references.

What happens to your information when we finish the research project?

The project is scheduled to end on 01.12.2018. Personal data and audio recordings stored in connection with the studies will be deleted from the USB flash drive, which will then be destroyed.

Your rights

As long as you can be identified in the data material, you are entitled to:

- an overview of what personal data is registered about you,
- to get personal information about you,
- Get deleted personal information about you,
- Get a copy of your personal information (data portability), and
- to send a complaint to your privacy representative or data protection agency regarding the processing of your personal information.

What gives us the right to process personal information about you?

We process information about you based on your consent.

On behalf of the Department of Information and Media Studies, NSD - Norwegian Center for Research Data AS has considered that processing of personal data in this project is in accordance with the privacy policy.

Where can I find out more?

If you have questions about the study or wish to avail yourself of your rights, please contact:

Department of Information and Media Studies, University of Bergen

- Per-Niklas Longberg (Student)
    - (47) 47 37 97 53
    - plo002@uib.no
- Associate Professor Ankica Babic (supervisor)
    - (47) 55 58 91 39
    - Ankica.Babic@uib.no

NSD - Norwegian Center for Research Data AS, by email (personvernombudet@nsd.no) or phone: 55 58 21 17.

With best regards

Project Manager (Researcher / tutor)	Student
Ankica Babic	Per-Niklas Longberg

--- ---

Consent Statement

I have received and understood information about the project Hip Arthroplasty Implant Longevity Prediction, and have had the opportunity to ask questions. I agree to:

To test the above system
Participate in a part-organized interview
That my name can be published in the completed master thesis

I agree that my information will be processed until the project is completed, approx. 01/12/2018

--- ---
(Signed by project participant, date)

Appendix B

Interview Guide for Semi-Structured Interviews.

Intervjuguide

Format:	Ansikt til ansikt
Svarregistrering:	Lydopptak, notater

Hovedmålet ved intervjuet er å først etablere den selvoppfattede tekniske forståelse hos deltaker, la deltaker utforske systemet satt til evaluering og deretter utforske deltakers oppfattelse av systemet relatert til brukervennlighet, arbeidsflyt og brukbarhet.

Innledning

Varighet: ca 5 minutter

Deltaker informeres om hva prosjektet går ut på og hva jeg vil oppnå med dette intervjuet. Deltakers egenvurdering på teknologisk ferdighet blir utforsket.

Spørsmål:

- På en skala fra 1 til 10, hvor teknologisk kompetent føler du deg?

- Bruker du ofte å beregne hvor lenge et implantat vil holde?
    - Er dette noe pasienter ofte ønsker å vite?
    - Beregner du en estimering selv, eller har dere et system for dette?
        - Hvor lang tid bruker du på dette? (dag/uke/måned)

- Har du erfaring med lignende systemer?
    - I så fall hvilke?

- Hvor mange systemer bruker du i gjennomsnitt i løpet av en arbeidsdag?
    - I forhold til vanskelighetsgrad, hvordan vil du beskrive disse?

Systemtest/utforsking

Varighet: 10-20 minutter

Deltaker får full frihet til å utforske systemet og teste dets funksjoner.

Oppgaver:

1. **Start systemet og plott inn vilkårlig pasientinformasjon i skjemaet. Lagre dette.**

2. **Sjekk at pasientinformasjonen er korrekt (i forhold til hva du plottet inn.)**

3. **Utfør en prediksjon (basert på resultatet av utførelsen av oppgave 1).**

4. **Begynn prosessen på nytt, og plott inn ny vilkårlig pasientdata. Lagre dette.**

5. **Endre på hvilke *features* (kolonner i datasettet) som brukes i prediksjonen før du utfører en ny prediksjon.**

6. **Utforsk statistikken bak prediksjonen.**

Deltaker vil få bistand til eventuelle tekniske problemer.

Hoveddel

Varighet: 10-20 minutter

Deltaker vil bli spurt spørsmål relatert til gjennomførelsen av oppgavene

Spørsmål:

- Kan du beskrive hvordan du opplever systemet?
  - Føles systemet enkelt å bruke?

- Er det noe i arbeidsflyten for å få systemet til å fungere som du føler er unødvending, overflødig eller vanskelig?
  - I så fall hvilke elementer? Hvordan er det vanskelig?

- Føles noen deler av systemet mer komplisert enn andre?
  - I så fall hvilke?
  - Hvordan skiller disse delene seg ut?
  - Har du noen tanker om hva som kunne gjøres for å forenkle disse deler?

- Var stegene i arbeidsflyten godt forklart?

    - Følte du at du hadde kontroll på systemet?

- Hvilke ytterligere parametere vil du ta inn (?)
    - Noe fra (helse-vest post-op skjema) du vil se i systemet?
        - Alder
        - Vekt
        - Annen sykdom
        - Implantat-type

- Utledende spørsmål
    - Har du noe mer å tilføye?

Appendix C

System Usability Scale Questionnaire.

System Usability Scale

Strongly disagree — Strongly agree

1. I think that I would like to use this system frequently

□ □ □ □ □
1 2 3 4 5

2. I found the system unnecessarily complex

□ □ □ □ □
1 2 3 4 5

3. I thought the system was easy to use

□ □ □ □ □
1 2 3 4 5

4. I think that I would need the support of a technical person to be able to use this system

□ □ □ □ □
1 2 3 4 5

5. I found the various functions in this system were well integrated

□ □ □ □ □
1 2 3 4 5

6. I thought there was too much inconsistency in this system

□ □ □ □ □
1 2 3 4 5

7. I would imagine that most people would learn to use this system very quickly

□ □ □ □ □
1 2 3 4 5

8. I found the system very cumbersome to use

□ □ □ □ □
1 2 3 4 5

9. I felt very confident using the system

□ □ □ □ □
1 2 3 4 5

10. I needed to learn a lot of things before I could get going with this system

□ □ □ □ □
1 2 3 4 5

Appendix D

NSD Approval for Processing Personal Data.

Det innsendte meldeskjemaet med referansekode 110728 er nå vurdert av NSD.

Følgende vurdering er gitt:
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 med vedlegg 06.11.2018, samt i meldingsdialogen mellom innmelder og NSD. Behandlingen kan starte.

MELD ENDRINGER
Dersom behandlingen av personopplysninger endrer seg, kan det være nødvendig å melde dette til NSD ved å oppdatere meldeskjemaet. På våre nettsider informerer vi om hvilke endringer som må meldes. Vent på svar før endringer gjennomføres.

TYPE OPPLYSNINGER OG VARIGHET
Prosjektet vil behandle alminnelige kategorier av personopplysninger frem til 01.12.2018.

LOVLIG GRUNNLAG
Prosjektet vil innhente samtykke fra de registrerte til behandlingen av personopplysninger. Vår vurdering er at prosjektet legger opp til et samtykke i samsvar med kravene i art. 4 og 7, ved at det er en frivillig, spesifikk, informert og utvetydig bekreftelse som kan dokumenteres, og som den registrerte kan trekke tilbake. Lovlig grunnlag for behandlingen vil dermed være den registrertes samtykke, jf. personvernforordningen art. 6 nr. 1 bokstav a.

PERSONVERNPRINSIPPER
NSD vurderer at den planlagte behandlingen av personopplysninger vil følge prinsippene i personvernforordningen om:

- lovlighet, rettferdighet og åpenhet (art. 5.1 a), ved at de registrerte får tilfredsstillende informasjon om og samtykker til behandlingen
- formålsbegrensning (art. 5.1 b), ved at personopplysninger samles inn for spesifikke, uttrykkelig angitte og berettigede formål, og ikke behandles til nye, uforenlige formål
- dataminimering (art. 5.1 c), ved at det kun behandles opplysninger som er adekvate, relevante og nødvendige for formålet med prosjektet

- lagringsbegrensning (art. 5.1 e), ved at personopplysningene ikke lagres lengre enn nødvendig for å oppfylle formålet

DE REGISTRERTES RETTIGHETER

Så lenge de registrerte kan identifiseres i datamaterialet vil de ha følgende rettigheter: åpenhet (art. 12), informasjon (art. 13), innsyn (art. 15), retting (art. 16), sletting (art. 17), begrensning (art. 18), underretning (art. 19), dataportabilitet (art. 20).

NSD vurderer at informasjonen om behandlingen som de registrerte vil motta oppfyller lovens krav til form og innhold, jf. art. 12.1 og art. 13.

Vi minner om at hvis en registrert tar kontakt om sine rettigheter, har behandlingsansvarlig institusjon plikt til å svare innen en måned.

FØLG DIN INSTITUSJONS RETNINGSLINJER

NSD legger til grunn at behandlingen oppfyller kravene i personvernforordningen om riktighet (art. 5.1 d), integritet og konfidensialitet (art. 5.1. f) og sikkerhet (art. 32).

For å forsikre dere om at kravene oppfylles, må dere følge interne retningslinjer og/eller rådføre dere med behandlingsansvarlig institusjon.

OPPFØLGING AV PROSJEKTET

NSD vil følge opp ved planlagt avslutning for å avklare om behandlingen av personopplysningene er avsluttet.

Lykke til med prosjektet!

Kontaktperson hos NSD: Belinda Gloppen Helle
Tlf. Personverntjenester: 55 58 21 17 (tast 1)