Feature selection for biomarker discovery

Using machine learning to find a minimal set of biomarkers that relate levels of exercise to self-perceived intensity of training

Kristof Fellegi

First supervisor: Dr.ing.habil. Georg Krempl
Second supervisor: Dr. Matthieu Brinkhuis
External supervisor: Marc A. T. Teunis PhD.

A thesis presented for the degree, Master of Science

Universiteit Utrecht

Department of Information and Computing Sciences
Master of Business Informatics
The Netherlands
2018-12-12

Abstract

This research is a comparative study of feature selection methods for biomarker discovery. 10 different machine learning techniques were considered for feature selection. The main assumption behind the research was that certain biomarkers can reflect the perceived strenuousness of the different exercise levels. For measuring the perceived exercise intensity, the Borg scale was used.

Using the top 10 most expressive biomarkers selected by each model, 39 different biomarkers were selected out of the total 64. The most frequently occurred one was "factord" selected by 7 models. Biomarkers "trp" and "CORT" were both selected by 6 of the models. "ifabp", "LEUCO" and "BICARB" were selected by 5 of the models.

In general, the predictive power of the applied machine learning techniques do not vary much. The highest accuracy, 78% was achieved by Logistic Regression. Regarding the area under the ROC curve, the best result was achieved using the full logistic regression model with an AUC = 0.72.

Applying feature selection however, a better performance can be achieved compared to the models with all the predictors. Recursive feature elimination on the random forest model yielded an 81% accuracy and the Lasso on logistic regression yielded an even higher 84% accuracy.

All in all, considering the criteria for selecting candidate models, Logistic regression represents a balanced mix of model performance and interpretability.

Keywords

Feature selection, Dimension reduction, Classification, Lasso, Biomarker discov-

ery, Bioinformatics, Exercise Physiology

Declaration

I declare that the work performed in this master thesis has been done independently and in accordance with the regulations at Utrecht University. The master thesis is built upon data from the project "DiAgRaMs - Standardization of the bicycle ergometer test as stress model to assess nutritional effects on intestinal function and immune responsiveness in healthy young men" have been carried out by the Innovative Testing Research Group since 2014.

Acknowledgements

I want to express my utmost gratitude to my supervisors, Marc Teunis A. T. PhD., Dr.ing.habil. Georg Krempl and Dr. Matthieu Brinkhuis for their outstanding guidance and support during my graduation period.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

According to Hevner et. al [11], the practical relevance of the research should be equally valued with the rigor of the research performed. In this master thesis I propose an applied data science solution for a fundamental research problem in the life sciences domain.

During previous studies of the Innovative Testing Research Group [15], about 100 different biological parameters (biomarkers) were measured in 15 healthy trained volunteers, who have been exposed to different exercise protocols on cycle ergometers. The goal of this research is to understand why some exercise protocols are perceived strenuous and which biomarkers reflect it. In order to label the measured biomarkers about their expressiveness of perceived exertion, the Borg scale is used [3].

In this supervised setting, the response variable is "protocol" and the predictors are the 64 valid biomarkers with measurement values for all combinations of the variables "subject", "protocol" and "time".

For resolving this classification problem, different machine learning techniques are proposed. As a first step, features are first ranked based on their relative importance in the outcome of the different models. Then different wrapper and embedded methods are applied for feature selection.

In order to reach the goal of my research, not only the model performance measures are of interest, but also the physiological meaningfulness of the different resulting subsets of predictors. In this thesis, the ten most important features selected by each technique are presented.

1.1 Problem statement

In general, healthy humans are well adapted to physical exercise and exertion. Homeostatic balance is ill-maintained if the balance is already disturbed due to a pre-existing health problem. Hormonal responses, liver metabolism and intestinal reactions, as well as immunological responses keep the homeostatic balance during and after exercise. These physiological responses can be determined by measuring a number of relevant biological parameters (biomarkers), reflecting homeostatic balance and/or disturbances.

Measuring these biomarkers is one problem, but the real challenge lies in re-

vealing the underlying knowledge from the measurements. The main assumption behind my research is that certain measured biomarkers can reflect the perceived strenuousness of the different exercise levels.

While familiarizing with the data set, different aspects emerged that resulted in the following problem statements. There are too many measured biomarkers, which reduces the interpretability of the results. Even after domain experts tried to reduce the number of parameters based on preliminary knowledge, the resulting data set remained so vast, that analyzing it with classical statistical methods proved to be too time-consuming. Furthermore, to our best knowledge, it is unknown, which biomarkers can reflect the perceived strenuousness of different exercise levels. [15]

To account for the above points, different machine learning techniques were applied on the data in order to come up with a minimum subset of predictors reflecting the perceived strenuousness of the different exercise levels.

1.2 Motivation

The importance of solving this problem is multifaceted. Based on literature review and consultation with domain experts, no such extensive research has been conducted that relates objective measurements of physiological changes in the human body to the subjective indicators of perceived strenuousness of an exercise. [15] From a biological point of view, it will deepen our understanding of the above phenomena. The goal of my research is to understand why some exercise protocols are perceived strenuous and what are the biomarkers that reflect it. From an applied machine learning point of view, it will serve as a use-case that machine learning techniques can be used for resolving such a problem. The motive to use machine learning is the excessive amount of data gained from the preliminary experiments and the availability of techniques to conduct feature selection.

The findings of my research will contribute to the available collective knowledge in the interdisciplinary research field of bioinformatics.

1.3 Research questions

To formulate research questions, I followed the template for design problems by Wieringa [33]. Substituting the specifics of my experiment, the following statement emerges:

- Improve the comprehensibility of the measured biomarkers
- using machine learning techniques
- that satisfy this setting
- in order to come up with a minimum subset, that relate levels of exercise to self-perceived intensity of training.

Based on this statement the main research question and three related subquestions were stated.

Main RQ: “Can we devise a method applying machine learning techniques that relate levels of exercise to self-perceived intensity of training?”

SQ1: “Which machine learning techniques can be used to find a minimal set of biomarkers that relate levels of exercise to self-perceived intensity of training?”

SQ2: ”Which machine learning techniques provide the best results in the scope of accuracy, sensitivity and specificity?”

SQ3: “Does applying feature selection improve the performance measures?”

1.4 Outline

In this thesis I investigate the application of different feature selection techniques for finding the most expressive biomarkers of self-perceived intensity of training. The proposed techniques are applied on a classification-type problem to predict the probability of a biomarker level being perceived as ”heavy” or ”intermediate” based on the Borg scale. [3] The rest of the paper is organized as follows. In Chapter 2 the related theoretical concepts are discussed. Then in Chapter 3 the applied research method is detailed. Chapter 4 elaborates on the the results and Chapter 5 includes the evaluation and limitation of the results. Finally conclusions are drawn in Chapter 6 and the future works are outlined.

Chapter 2

Background and related work

This chapter briefly introduces the background theories behind both the domain specific and the machine learning related concepts of my research.

2.1 Background of the research

Homeostasis is maintained within the physiological boundaries by adaptive responses within organisms. In general, healthy humans are well adapted to physical exercise and exertion. Homeostatic balance is ill-maintained if the balance is already disturbed due to a pre-existing health problem or if the body is exposed to severe physical exercise. Hormonal responses, liver metabolism and intestinal reactions, as well as immunological responses keep the homeostatic balance during and after exercise. These physiological responses can be determined by measuring a number of relevant biological parameters (biomarkers), reflecting homeostatic balance and/or disturbances.

During previous studies, the Innovative Testing Research Group [15] measured about 100 different biological parameters in 15 healthy trained volunteers, who have been exposed to different levels of exercise on cycle ergometers. The measured parameters are derived from serum, urine and saliva samples, collected in the beginning, during and after the exercise. Sample analysis was conducted in six different laboratories. The goal of the experiment was to determine the extent of exercise and select relevant biomarkers of intestinal function and immune responsiveness in healthy young men.

Some measurements served only as meta-categories to facilitate the laboratory analysis. These were later removed from the data set. After getting rid of these meaningless biomarkers, there remained still 88 different candidates to consider. This number was further decreased to 64, due to repetition in the data set or the lack of their value added. This decision was made together with the domain experts involved in the experiment. Using the 64 valid biomarkers in every combination of subjects, protocols and time points, the resulting data set is still so overwhelming that it was very time-consuming to give meaning to it.

2.1.1 Borg scale

In order to label the measured biomarkers about their expressiveness of perceived exertion, the Borg scale was used. [3] It is a subjective quantitative measure of perceived exertion during physical activity. It was originally introduced by Gunnar Borg, who proposed a scale of 6-20 to measure the perceived exertion of individuals exposed to physical activity. Table 2.1 shows the Borg scale with the adherent verbal descriptions for each level.

Rating of Perceived Exertion	**Description**
6	No exertion
7	Extremely light
8	Very light
9	Very light
10	Very light
11	Light
12	Light
13	Somewhat hard
14	Somewhat hard
15	Hard
16	Hard
17	Very hard
18	Very hard
19	Extremely hard
20	Maximum exertion

Table 2.1: Rating of Perceived Exertion on the Borg scale

2.2 Machine Learning

There is not one generally accepted definition of Machine Learning in the literature. It serves more as an umbrella term with the aim to get computers to learn like humans do. A common definition of it is as follows. "Machine learning is the study of algorithms and mathematical models that computer systems use to progressively improve their performance on a specific task." [34] The term Machine Learning is attributed to no other than Alan Turing. In his 1950 paper, he proposed a 'learning machine' that could learn and become artificially intelligent. [21] Since then, both the field and its heuristics evolved tremendously. New techniques are being introduced every couple of years and experts are in high demand on the job market recently. [7]

Machine Learning is basically statistical learning, which refers to a vast set of tools for understanding data. Zhou interprets learning as the process of generating models from data. This is accomplished by a learning algorithm. [35]

2.2.1 Unsupervised learning

Learning tasks can be classified as supervised or unsupervised. James et al. defines supervised learning as "building a statistical model for predicting, or estimating, an output based on one or more inputs." In comparison to that, unsupervised learning focuses on learning relationships and structure from data without explicit labels being present. Unsupervised learning is the situation in which for every observation there is a vector of measurements, but there is no associated response [12] Chapter 2.

Principal Component Analysis

"The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variability present in the data set." [14] The cutoff for a data set to be considered high-dimensional is subject to opinion.

PCA does not consider the response variable when summarizing variability. With other words, it is blind to the response, which is why it is considered as an unsupervised technique. [19]

When applying PCA, the underlying assumption is that a linear combination of the predictors with high variability is probably going to be associated with the response [12] Chapter 6. This linear transformation fits a data set to a new coordinate system in such a way that the most significant variability is found on the first coordinate, and each subsequent coordinate is orthogonal to the last and has a lesser variability. In other words, PCA finds a linear projection of high-dimensional data in such a way that the variability of the projected data is maximized. This projection can be visualized in two dimensions, the first two principal components being the axes.

The first principal component is a linear combination of the original predictor variables and captures the maximum variability in the data. The second principal component is the linear combination that has the largest variability out of all linear combinations that are totally unrelated to the first principal component. It explains the second most variability in the data.

Since principal components are linear combination of the original predictor variables, capturing variability in the data, they cannot be related directly to the original predictors. Due to this fact, PCA is considered as a black-box method. [24]

The application of PCA is further discussed in Chapter 3.2.6 with graphs derived from the data set to help understand the implications.

2.2.2 Supervised learning

According to James et. al, supervised learning refers to the situation in which for every observation of the predictor measurements there is an associated response measurement [12] Chapter 2.

Within supervised learning, the current research is aimed to resolve a classification-type problem.

2.3 Candidate Machine Learning techniques

In the following section, the theoretical concepts of the applied machine learning techniques are introduced briefly. First, logistic regression, then different tree-based methods. Tree-based methods stand out from Machine Learning techniques because of their ease of use, interpretability and relatively good predicting accuracy. They can handle both categorical and continuous input variables without much data preparation. They are robust to outliers and can handle missing values antively.

When speaking of tree-based methods, the main task is splitting the predictor space into a number of simple regions. These splitting rules can be summarized in a tree [12] Chapter 8.

Although tree-based methods are widely used, they also have some disadvantages. Trees in general have high variance, which causes poor model performance. They also overfit the data easily. Trees can grow very large by which they lose their good interpretability. For this reason however, pruning can be used.

2.3.1 Logistic regression

Logistic regression is a well-known classification technique that models the log odds of an event as a linear function. [19]

It is the extension of linear regression to classification problems. The outcome variable of logistic regression is a categorical variable, while the predictors can be both categorical and continuous. In this case, the - linear relationship between variables - assumption of linear regression is violated. One way to resolve this problem is to transform the data using logarithmic transformation.

When the goal is to predict membership of only two categorical outcomes, it is called a binary logistic regression. [9]

2.3.2 Decision trees

Decision trees are hierarchical structures with nodes and directed edges. The node at the top is the root node, the nodes at the bottom are called leaf nodes and in between lie the internal nodes.

Classification trees consist of nested if-then statements. [19] The tree-building process is described as a top-down greedy approach. A split condition is used to predict class labels based on one or more input variables. The classification process starts from the root node of the tree and at each node the process will check whether the input value should recursively continue to the right or to the left sub-branch according to the split condition. The process stops when it meets any leaf nodes. This is called recursive binary splitting [12] Chapter 8.

The goal is to split the data into sub-sets where each sub-set is as pure as possible.

Mathematically it is more feasible to measure impurity than purity. Different approaches were proposed for this purpose. Two of the most popular impurity measures are the Gini-index and Entropy/Information gain. The lower the index, the higher the purity of the split. The decision tree selects the split that minimizes the applied impurity measure. [20]

2.3.3 Random Forest

Random Forest is an ensemble of trees trained on bootstrapped samples of the training data and then combined to yield an improved prediction accuracy. Although, there is no free lunch, the improved prediction accuracy implies decreased interpretability of the model [12] Chapter 8.

For each tree, variable importance is calculated based on the prediction accuracy on the out-of-bag portion of the data. This process is done iterating through each randomly selected predictor variable. This extra randomness leads to a collection of trees that are decorrelated from each other. Random forest is an improvement compared to bagging in a sense that correlation between the sampled trees is reduced. This results in a reduced variance when averaging the trees.

2.3.4 Boosting

Boosting is another tree-based ensemble method. Similarly to the random forest, the process starts out with training on bootstrap samples from the data. However,in boosting trees are grown sequentially, using information from previously grown trees. This results in an improvement in prediction power.

Even though boosting offers a lot of flexibility in hyperparameter tuning, it is reluctant to overfitting [12] Chapter 8. However, this high flexibility makes it computationally expensive as it requires a large grid search during tuning. Nevertheless, boosting is considered as one of the most accurate machine learning technique.

2.4 Feature selection

This section introduces the feature selection concepts used in this paper.

Due to the high-dimensionality of the data set, feature selection is suggested to avoid the curse of dimensionality. [8]

Feature selection is the process of finding a subset of the original feature-set, such that an induction algorithm that is run on data - containing only the subsetted features - generates a learner with the highest possible accuracy. [17]

One way to categorize feature selection methods is as follows:

- Filters,
- Wrappers,
- Embedded methods.

In the following sections, these categories of feature selection methods are detailed.

2.4.1 Filters

According to Kohavi and John [17], filtering is considered as a pre-processing step in feature selection. It is also called feature ranking, as features are ranked based

on their relevance for predicting the desired outcome. Then only those features will be included in the outcome that pass some criterion. [19]

As stated by Guyon and Elisseeff [10], filters are sometimes preferred to other variable subset selection methods because of their computational and statistical scalability. Computationally, they only require computing and sorting n scores. Statistically, they are robust against overfitting, as they increase bias, but decrease variance in the model.

2.4.2 Wrappers

Kuhn defines wrappers as search algorithms that treat the predictors as the inputs and utilize model performance as the output to be optimized. [19]

Wrappers evaluate multiple models by adding or removing predictors to find the optimal combination that maximizes model performance.

Guyon et al. [10] defines the application of wrappers as a 3-step process. First a method for searching the feature space has to be selected. Then the model performance measure needs to be defined that will halt the search. Finally, based on these, the predictors will be selected as the outcome.

Kohavi and John [17] consider wrappers a simple and powerful way to address the problem of feature selection, regardless of the chosen machine learning technique.

One group of examples of wrappers are stepwise methods.

Stepwise methods

According to Guyon et. al [10], stepwise methods are two types of greedy search strategies: forward selection and backward elimination.

They are computationally advantageous and robust against overfitting.

The evaluation of the information derived from removing predictors is based on different relative quality criterion.

In the case of forward selection, initially only the constant is included in the model. Then predictors are added one-by-one, iteratively. At each iteration one predictor is added that has the highest simple correlation with the outcome. If it makes a contribution to the predictive power of the model, it is retained and another predictor is considered. This process continues until a cut-off value for the relative quality criterion has reached [9] Chapter 7.

In the case of backward elimination, initially all the predictors are present in the model. Then they are removed one-by-one, iteratively. At each iteration one predictor is removed that has the smallest effect on the relative quality criterion. This process continues until removing one more predictor would cause the relative quality criterion to change direction [9] Chapter 7.

According to Field et. al [9] Chapter 7, backward elimination is preferred to the forward stepwise method, because of suppressor effects. These occur when a predictor has an effect but only when another predictor is held constant. In other words, forward selection is more inclined to make a Type II error.

2.4.3 Embedded methods

Guyon et. al defines embedded methods as algorithms that perform feature selection in the process of model training. [10]

The idea is to combine the advantages of both filter and wrapper methods. To proceed more efficiently, embedded methods directly optimize a two-part objective function with a goodness-of-fit term and a penalty for a large number of variables. [10] This results in a learning algorithm that performs feature selection and classification simultaneously.

One advantage of embedded methods is that they do not split the training data into a training and validation set, hence they use the available data better. Besides, they do not need to search the whole feature space at every iteration, so they reach a solution faster. [10]

Embedded methods are usually specific to given learning algorithms. The most important example is the Lasso, which will be described in the following section.

Lasso and Ridge Regression

Least Absolute Shrinkage and Selection Operator (Lasso) is a powerful regularization-type feature selection method. It estimates the outcome while automatically selecting significant features by shrinking the coefficients of unimportant predictors to zero. [27]

Ridge regression and Lasso do not use least squares to fit, but a different criterion that has a penalty that will shrink the coefficients toward 0, or exactly 0 in the case of the Lasso.

The tuning parameter lambda controls the overall strength of the penalty. The best value of lambda can be found using cross-validation.

When alpha=1, the Lasso will be used, while setting alpha to 0 will result in using Ridge Regression. The main difference between the two models is that Ridge Regression minimizes the residual sum of squares of the coefficients, while Lasso minimize their absolute value.

2.5 Validation

When the modeling is done, the next and final step is to test the predictability of the algorithms on a new unseen data set [12] Chapter 6.

A common and acclaimed way of evaluating any machine learning technique is to split up the initial data set into a training and a test set. The common approach is to train the models on the - usually larger chunk - training set and test them on an unseen test set. This way the performance of the selected method can be tested immediately [12] Chapter 2.

The goal of the validation is to gain performance measures of models based on which conclusions can be drawn regarding their relative performance. The most important evaluation metric for binary classification problems is the confusion matrix. It is a convenient way to display Type I. and Type II. errors [12] Chapter 4. Elements on the diagonal of the matrix represent correctly classified cases, while off-diagonal elements represent misclassified ones.

Based on the confusion matrix, different performance measures can be calculated. Presenting them all is out of the scope of this thesis.

2.5.1 Cross-Validation

Cross-validation is a technique for assessing the accuracy of a model across different samples [9] Chapter 7. It is one way to reduce the variance of an estimate by averaging multiple estimates together. Cross-validation can help determining the optimum number of features in a model.

In K-fold cross-validation, the data is partitioned into K subsets of equal sizes. In each iteration, one of the subsets is held out as a test set and the rest is used for training.

Chapter 3

Methods

3.1 Design Science Methodology

The logical structure of my research is based on the design science methodology, proposed by Hevner et. al. [11]. This methodology carefully balances between the theoretical and practical soundness of research projects. This approach is particularly important for my graduation project, being a fundamental research with an applied data science solution.

According to Wieringa [33], Design science is the design and investigation of artifacts in context. Hevner [11] splitted up the required factors for a design science research project to theoretical knowledge (Knowledge Base), application domain (Environment) and the proposed artifacts (Design Science). Furthermore, he introduced three cycles to iterate through and connect the three factors. The relevance cycle aims to find input and requirements from the application context to the designed artifact. The rigor cycle on the other hand aims to establish solid theoretical foundations for the research to help evaluate that later. The internal design cycle iterates through the activities of building, evaluating and refining the designed artifacts. The three cycles together create a solid methodology, widely used in the engineering and computer science fields.

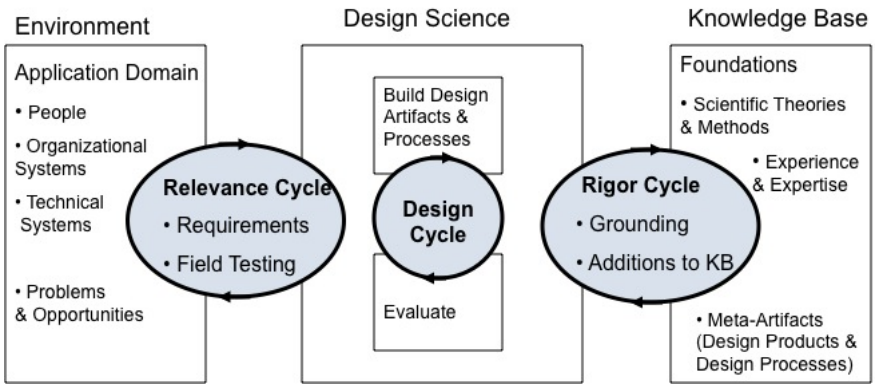

Figure 3.1: Design Science Research Methodology by Hevner 2007 [11]

3.1.1 Environment

The application domain of my research is physiology and human biology. The initial experiment was carried out on healthy young men. For the set-up, execution and evaluation of the experiment, experts from various fields contributed with their domain specific knowledge. These are biologists, immunologists, toxicologists, physiologists and statisticians [15, 13].

3.1.2 Design Science

Within the Design Science phase, the design cycle iterates through the activities of building, evaluating and refining the designed artifacts.

The artifacts in this research are the applied machine learning techniques for feature selection.

For evaluating the artifacts, the evaluation metrics were used as outlined in 2.5.1.

3.1.3 Knowledge Base

The reason I use both the Design Science Methodology and the CRISP-DM process model is that they can be nicely aligned with each other. Both methodologies follow cyclic iterations of activities. I consider the CRISP-DM model as part of the Knowledge Base [11] of my research.

3.2 CRISP-DM

In order to increase the reproducibility of my research, I follow the Cross-Industry Standard Process for Data Mining (CRISP-DM) [5]. It is an open standard process model, developed by a consortium of over 200 interested organizations, funded by the European Union. According to Meta S. Brown, it is by far the most widely used analytics process model [4]. It has six major phases as depicted in Figure 3.2.

It provides a framework for the timely logical structure of the applied data science workflow. In the following paragraphs all six phases are detailed tailored to the application in this thesis.

3.2.1 Business Understanding

As the first step of any data science project, business understanding is the process of familiarizing with the application domain and outlining data science solutions that adhere to the desired goals.

During the business understanding phase, I started out with a systematic literature review. The aim of it was to become familiar with the state-of-the-art machine learning techniques and their application in the life sciences domain. The most frequently used keywords for search were the following: Feature selection, Classification, Dimension reduction, Lasso, Bioinformatics, Exercise Physiology, Biomarker discovery.

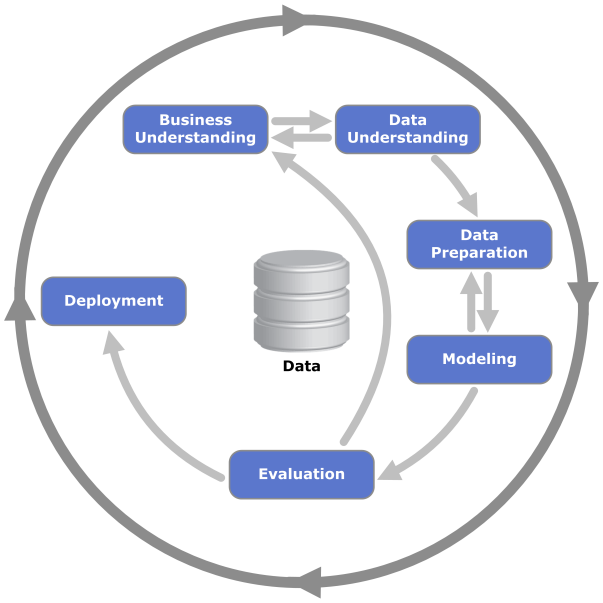

Figure 3.2: The CRISP-DM Process Model

My first encounter with the project was at the Avicenna hackathon [1] in February 2018. That was the first time I learnt about the preliminary experiments and the resulted data set that also became the initial data set I used for this research.

Next to the systematic literature review, I have also regularly consulted with domain experts of the Innovative Testing Research Group. Since my knowledge of the application domain (life sciences) is limited, I strongly rely on their input when discussing the physiological relevance of the biomarkers in question.

3.2.2 Data Understanding

Data understanding is about collecting data and developing a clear understanding of it. Different statistical methods are broadly used for this purpose, [9] from which I will propose the ones used in this research.

Data was collected during previous studies of the Innovative Testing Research Group [15] from serum, urine and saliva samples, in the beginning, during and after the exercise. These biological measurements comprise the major part of the data set. I received the data as an RData extension. R [23] is a language and environment for statistical computing. It has been widely used by scientists for data analysis and visualization in recent years. [16] The R environment was used for the entire workflow, from data acquisition through analysis and visualization to reporting. For reproducibility purposes, Appendix A includes the complete list of necessary information about the different versions and packages used during this research poject. Measurements for the Borg scale were received as an xlsx extension. After some initial data preparation in Microsoft Excel, this data was also loaded into R for further analysis.

The received data was in a so-called tidy format. This heuristic was introduced by Hadley Wickham [29] and follows the following principles:

- Each variable forms a column.

- Each observation forms a row.
- Each type of observational unit forms a table.

Presenting the data this way highly enhances its comprehensibility and manageability. In the words of Hadley "Tidy data is a standard way of mapping the meaning of a data set to its structure." [29]

Although the data was in a tidy format, it was a combination of two data sets. This results from two different studies carried out at two different locations with the same aim. The reason for the second study was to increase the validity and reproducibility of the experiments. The related paper about the experiments is still being written, which explains the lack of a citation here. The originally received data set contained almost 60000 observations and 13 variables in a stacked (long) format. Regarding the relative size of the data sets, the first one accounts for 76%, while the second accounts for 24%.

3.2.3 Data Preparation

Data preparation includes different methods for transforming data in order to prepare it for the machine learning techniques in question.

Exploratory Data Analysis

After checking the dimensions of the data sets and the classes of the variables, some questions occurred. To answer them, data manipulation and visualization techniques were used. This process is also called Exploratory Data Analysis (EDA), which is an iterative process to explore data in a systematic way. [31]

The first step was to split up the data by study and explore the resulting data sets separately.

Data cleaning

As an initial step of data preparation, the features present in the model were examined. This was carried out in consultation with the researchers involved in the preliminary experiments. [15] Some of the measured parameters were grouping features with no directly meaningful effect. After getting rid of these meaningless parameters, there remained still 88 different candidates to consider. This number was further decreased to 64, due to repetition in the database or the lack of their value added. The set of 64 valid parameters formed the basis of the further analysis.

Missing values

The next step in the data preparation phase was to deal with possible missing values.

Hadley differentiates explicitly and implicitly missing values. [31] Explicitly missing values are denoted by "NA", while implicitly missing values are simply not present in the data. "One way to think about the difference is with this Zen-like koan: An explicit missing value is the presence of an absence; an implicit missing value is the absence of a presence." [31] To investigate the missing values, different data manipulation and visualization techniques were applied. For visualization

purposes, the R package "ggplot2" was used. [28] This package is a collection of tools for visualization. To manipulate the data, the R package "dplyr" was used. [32] Both packages are part of the so-called "tidyverse" [30], which is a collection of R packages for data exploration, manipulation and visualization. They all share a common design philosophy with interoperability and reproducibility in mind.

Initially, there were no explicitly missing values in either of the data sets due to preliminary cleaning carried out by the participating researchers. However, there were 1201 and 1354 implicitly missing values in each of the data sets. Proportionally these represent 3% and 11% of the data respectively, so it is justifiable to further explore them. After some data manipulation and visualization, the following findings were revealed:

- The separate data sets contained different protocol levels.
- Most of the missing values are derived from the aggregated factor levels of the variables "protocol", "subject" and "time" in the combined data set.
- There is systematic missingness in the case of subject 8, so his results are excluded from further analysis.

Based on these findings I wrote a function for each data set that cleans up the irrelevant factor levels and filters out the values for subject 8. The cleaned data sets now contained only 10 and 30 actual missing cases respectively. These result from measurement errors during the experiment. Their relative proportion in the data sets are so tiny (0,03% and 0,26% respectively) that they were excluded from further analysis.

Since the second study contained different "protocol" levels than the first one, I excluded the values of the second study from further analysis. This way, the robustness and reliability of the research highly increases.

Intra- and inter-subject variability

One of the questions stated during exploratory data analysis was "How do the concentrations of the different biomarkers change in each subject by time and protocol?" To answer this question, data was transformed and visualized. Figure 3.3 gives an indication to answer this question based on the example of one of the measured biomarkers.

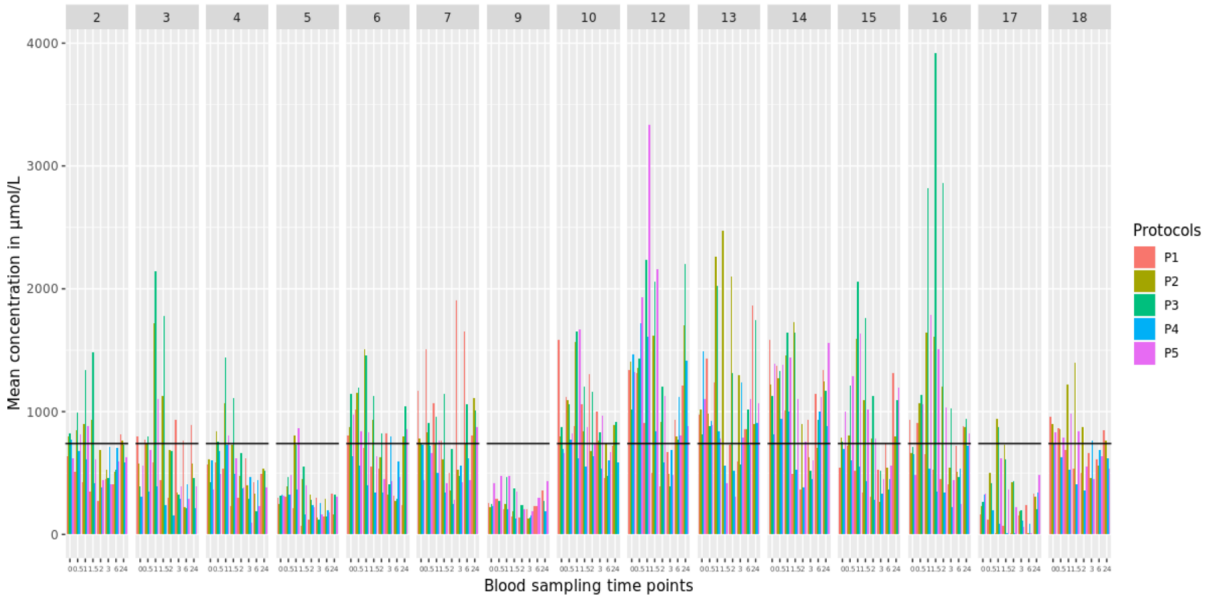

Figure 3.3: Change of "ifabp" concentration in each subject by time and protocol

As it can be seen from Figure 3.3 both intra- and inter-subject variability is present in the data. To account for intra-subject variability, the preliminary experiment was set up with strict rules in place regarding the diet and exercise habits of the participating subjects. [15]

To account for inter-subject variability, concentration levels of the measured biomarkers were normalized to the baseline (P1 - rest) protocol.

Normalization

Even though the preliminary experiments were set up in such a way, to minimize the effect of uncontrolled variables as much as possible, the data set still contains some noise.

Considering the goal of this research, some noise can be disregarded, but some need to be accounted for. That is why the measured concentration levels of the biomarkers were normalized to a baseline level, the P1/rest protocol.

Normalized concentration is calculated by dividing the concentration of the exercise protocols (P2, P3, P4, P5) by the corresponding baseline (P1) concentration for every combination of subject and time points. This way the measurements become more meaningful, expressing their relative value compared to the rest condition.

Transformation to wide format

Originally, the received data was in a long (stacked) format. This is advantageous for storing and presenting high-dimensional data, but not appropriate for many machine learning techniques. That is why as a next step of data preparation, the data set was transformed to a so-called wide format.

The first step to achieve this was to generate unique identifiers for row names. For this purpose, values of the columns "subject", "protocol" and "time" were extracted and combined to one string per row.

Then the stacked data set was spread according to the measured biomarkers. This way, the new transformed data set contained a column for each measured biomarker and for every one of them the measurement values were included in rows for every combination of "subject", "protocol" and "time".

As a last step, the dependent variable, "protocol" was added to the transformed data set as a separate column.

Correlation Analysis

To get a picture of the underlying structure in the data, correlation analysis was performed. According to Field et. al [9], bivariate correlation is a measure of the strength of relationship between two variables. It can also be a measure of the strength of an experimental effect (effect size). To express correlation, different techniques can be used. In this research, Pearson's correlation coefficient r was used. It is the standardized covariance between variables on a scale between -1 and +1. A coefficient of +1 indicates a perfect positive relationship, while a coefficient of -1 indicates a perfect negative relationship. A coefficient of 0 indicates no linear relationship.

Cohen [6] proposed the following effect sizes based on Pearson's r:

- r = .10 (small effect): The effect accounts for 1% of the total variance.
- r = .30 (medium effect): The effect accounts for 9% of the total variance.
- r = .50 (large effect): The effect accounts for 25% of the total variance.

This provides a useful benchmark for assessing effect sizes, however interpreting effect sizes is highly domain dependent. [2]

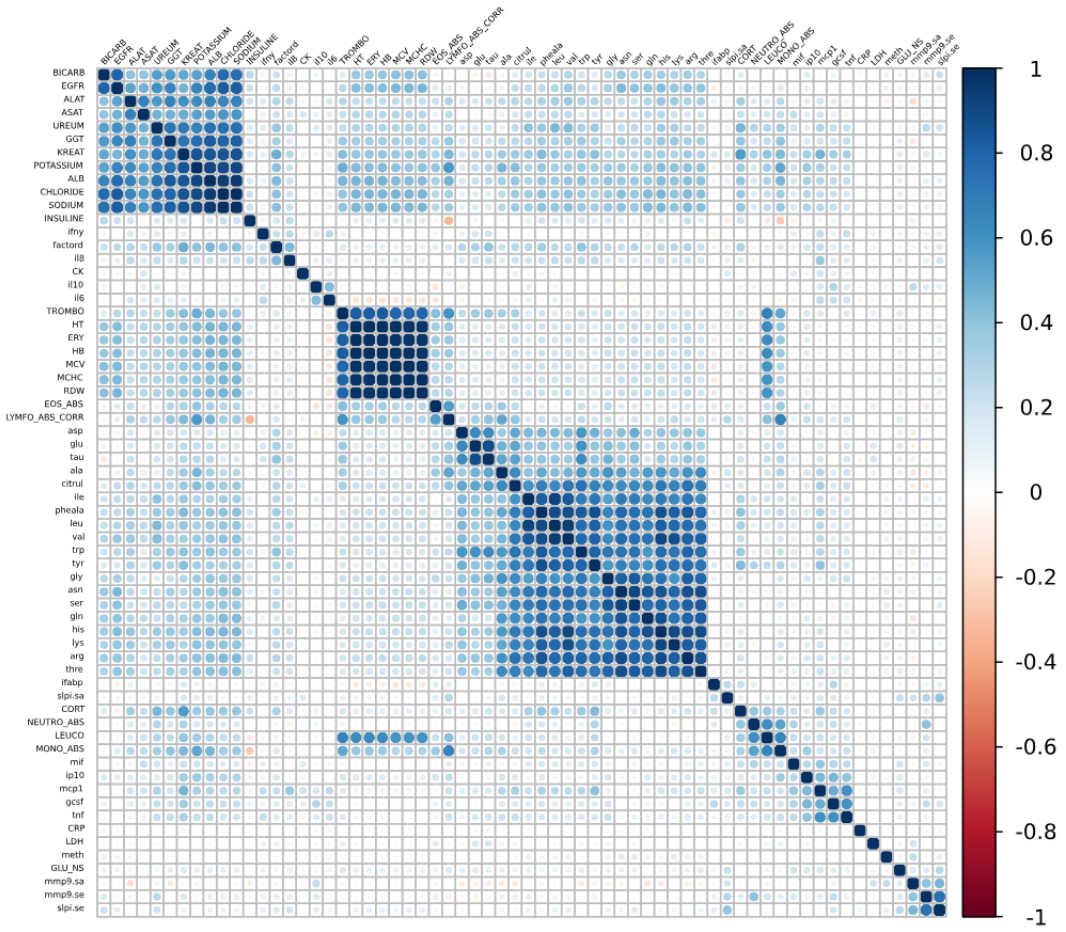

Figure 3.4: Correlations between the 64 valid biomarkers

Figure 3.4 shows the underlying correlations in the data. As it can be seen, based on Cohen's proposed effect sizes, certain biomarkers highly correlate with each other.

Principal Component Analysis

Initially the data set in this research did not include true, meaningful labels for the measured parameters, so I approached it in an unsupervised way.

To explore the data set in an unsupervised way and try to reduce the dimensions, Principal Component Analysis (PCA) was performed.

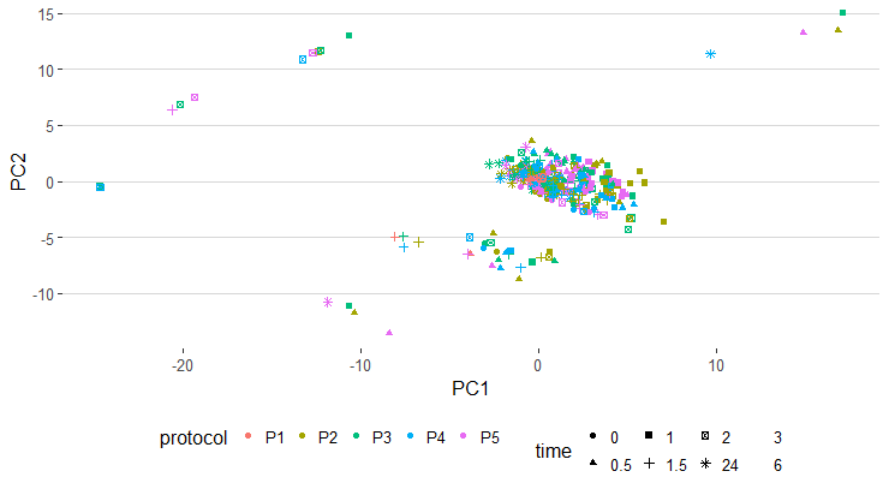

Figure 3.5: 2D visualization of the data points along the 1st and 2nd principal component

Figure 3.5 is a projection of the linear combinations of the data points along the first two principal components that explain the most variance in the data. This is a proven way of visualizing high-dimensional data sets in two dimensions. As it can be seen from Figure 3.5 data points are segregated into three bigger clusters. However, these clusters do not align with neither the protocol levels, nor the time points. It only shows the structure of the data after a linear transformation. That is why further techniques are needed to give meaning to the underlying knowledge in the data.

Introduction of new labels for the dependent variable "Protocol"

The main assumption behind my research was that the measured biomarkers can reflect the perceived strenuousness of the exercise. During the preliminary experiments, the self-perceived intensity of each exercise protocol was measured with the so-called Borg scale. [3]

Values extracted from the Borg scale were used in order to label the measured biomarkers about their expressiveness of perceived exertion. Labelling the measured biomarkers allows to differentiate clear categories and to apply supervised machine learning algorithms for selecting the most expressive predictors. Figure 3.6 shows the original protocol levels applied during the preliminary experiment.

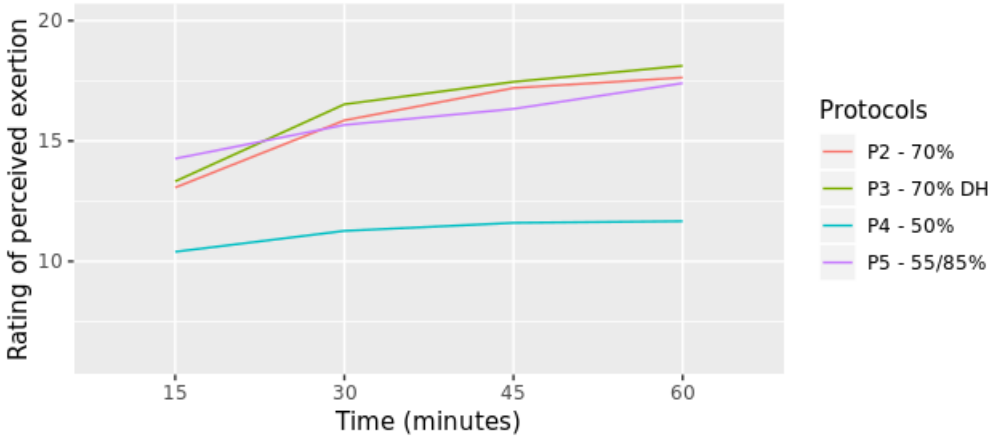

Figure 3.6: Average Perceived Exertion levels by the original protocol levels

As it can be seen from Figure 3.6, protocols P2, P3 and P5 converge and show similar values. That is why all three are coded as "heavy" in Figure 3.7. Furthermore protocol P4 deviates from the rest. In accordance with the Borg scale, it is coded as "intermediate". The "baseline" category is P1, the rest protocol, for which there were no meaningful measurement values understandably. In the experimental setting, the perceived exertion during P1 (the rest protocol) was considered 6 (the minimum value on the Borg scale) for all subjects. This served as a baseline, relative to which other protocol levels could be considered. Figure 3.7 shows the relabelled protocol levels.

Figure 3.7: The original and the new protocol levels based on the Borg scale

3.2.4 Modeling

Modeling is about applying the previously selected models on the prepared data set. It can be decomposed to training the models, tuning their parameters and testing their performance.

As a first step of the modeling phase, the data set was split into a training and a test set. Choosing a cut-off value regarding the fraction of training and test set is an open problem. [10] In this research I used 80% of the data for training the models and the remaining 20% for testing them.

Different machine learning techniques were used as candidate models for feature selection. Table 6.1 shows them in order of appearance in this thesis.

Logistic regression
Random forest
Boosting
Decision tree with the Gini-index
Decision tree with Entropy
Recursive feature elimination on RF
Lasso using minimum lambda
Lasso using 1 standard error lambda
Ridge regression using minimum lambda
Ridge regression using 1 standard error lambda

Table 3.1: Candidate Machine Learning techniques for feature selection

The rationale behind the chosen techniques is explained in section 5.2.5.

As a first approach, a so-called forced entry method was used. This means introducing all the predictors at once. Since some of the candidate techniques do not do automatic feature selection, features were first ranked based on their relative importance in the outcome of the model. Then, the top 10 most important features were selected, that reflect the perceived intensity of different exercise levels. The number 10 as a "minimum" number was chosen based on results of the PCA, considering the comprehensibility of the underlying biomarkers and the formal requirements of this thesis. A detailed discussion about it can be found in section **??**

After that, different wrapper and embedded methods were used for feature selection.

As a last step, candidate models were tested on the unseen part of the data set. For modeling purposes the R package Caret [18] was used.

3.2.5 Evaluation

After the modeling phase, the next step is to test the predictability of the candidate models on a new unseen data set. [12]

The goal of the evaluation is to gain performance measures of models, based on which conclusions can be drawn with certain confidence about the performance of the used algorithms.

As evaluation criteria, the following approaches were taken into consideration:

- Model performance
- Model interpretability
- Meaningfulness of the resulted subset

Regarding model performance, different measures were calculated based on the confusion matrices of the candidate techniques. These were then compared and conclusions were drawn. In my thesis I present the following performance measures for the applied models.

- Accuracy

- Sensitivity
- Specificity
- Positive Predicted Value
- Negative Predicted Value
- Prevalence
- Arean Under the ROC Curve

Next to the objective measures of performance, the interpretability of the different techniques were evaluated. For this purpose, the principle of Occam's razor was used, i.e. the simpler the model the better.

Being an interdisciplinary study, aspects from other domains need to be taken into consideration for evaluating the results. To evaluate the biological meaningfulness of the resulted subsets, I regularly consulted with a domain expert from the field.

3.2.6 Deployment

The deployment of a data science project depends on the goals stated previously and the degree to which the project fulfills them.

The main deliverable of this research project is the code I wrote in R. Functions will be extracted from it and included in the R package 'gramlyr'. [26] This package is being built with the aim to demonstrate reproducible data analysis in the life sciences through different real-life examples.

Chapter 4

Results

This chapter presents the results of my research. The sections describe the top 10 most important Biomarkers selected by the different Machine Learning techniques. Furthermore confusion matrices and the resulting performance measures are included, where it is relevant.

4.1 Logistic Regression

		Actual		
		Positive	**Negative**	**True/Total**
Predicted	**Positive**	66	16	0.80
	Negative	4	7	0.36
	True/Total	0.94	0.70	0.78

Table 4.1: Confusion Matrix of the Logistic Regression model

4.2 Decision Trees

		Actual		
		Positive	**Negative**	**True/Total**
Predicted	**Positive**	58	15	0.79
	Negative	12	8	0.60
	True/Total	0.83	0.65	0.71

Table 4.2: Confusion Matrix of the Decision Tree with the Gini-index

		Actual		
		Positive	**Negative**	**True/Total**
Predicted	**Positive**	60	16	0.79
	Negative	10	7	0.59
	True/Total	0.86	0.70	0.72

Table 4.3: Confusion Matrix of the Decision Tree with Information gain

The plots of the created decision trees can be found in Appendix B and Appendix C.

4.3 Random Forest

		Actual Positive	Negative	True/Total
Predicted	**Positive**	68	19	0.78
	Negative	2	4	0.33
	True/Total	0.97	0.83	0.77

Table 4.4: Confusion Matrix of the Random Forest model

4.4 Boosting

		Actual Positive	Negative	True/Total
Predicted	**Positive**	65	17	0.79
	Negative	5	6	0.45
	True/Total	0.93	0.74	0.76

Table 4.5: Confusion Matrix of the XGBoost model

4.5 Random Forest with Recursive feature elimination

From stepwise methods, recursive feature elimination was used. The underlying algorithm in the model was a Random Forest, used with 5-fold cross-validation. The best subset size was estimated to be 5 predictors. To stay consequent, the top 10 selected biomarkers are presented in this thesis in table 4.11.

		Actual Positive	Negative	True/Total
Predicted	**Positive**	70	18	0.80
	Negative	0	5	0.00
	True/Total	1.00	0.78	0.81

Table 4.6: Confusion Matrix of the Random Forest model applying Recursive Feature Elimination

4.6 Lasso

In the case of the Lasso model, 5-fold cross-validation was used to select the optimal value of lambda. Both the the models with the minimum value of lambda and

the value 1 standard error from the minimum were used and their performance is presented here.

		Actual		
		Positive	**Negative**	**True/Total**
Predicted	**Positive**	69	14	0.83
	Negative	1	9	0.10
	True/Total	0.99	0.61	0.84

Table 4.7: Confusion Matrix of the Lasso model using the minimum value of lambda

		Actual		
		Positive	**Negative**	**True/Total**
Predicted	**Positive**	69	20	0.78
	Negative	1	3	0.25
	True/Total	0.99	0.87	0.77

Table 4.8: Confusion Matrix of the Lasso model using the lambda 1 standard error away from the minimum

4.7 Ridge Regression

The same way as with the Lasso, in the case of the Ridge Regression model, 5-fold cross-validation was used to select the optimal value of lambda. Both the the models with the minimum value of lambda and the value 1 standard error from the minimum were used and their performance is presented here.

		Actual		
		Positive	**Negative**	**True/Total**
Predicted	**Positive**	68	19	0.78
	Negative	2	4	0.33
	True/Total	0.97	0.83	0.77

Table 4.9: Confusion Matrix of the Ridge Regression model using the minimum value of lambda

		Actual		
		Positive	**Negative**	**True/Total**
Predicted	**Positive**	69	20	0.78
	Negative	1	3	0.25
	True/Total	0.99	0.87	0.77

Table 4.10: Confusion Matrix of the Ridge Regression model using the lambda 1 standard error away from the minimum

4.8 All models

Table 4.11 shows the 10 most expressive biomarkers, that reflect the intensity of exercise protocols, selected by each candidate model.

Rank	GLM	RF	XGB	DT_gini	DT_info	RFE	Lasso_min	Lasso_1se	Ridge_min	Ridge_1se
1	mif	LEUCO	LEUCO	tau	LEUCO	LEUCO	HT	trp	pheala	citrul
2	ip10	CORT	factord	glu	CORT	CORT	ERY	citrul	citrul	pheala
3	GLU_NS	ip10	CORT	CORT	trp	EGFR	RDW	pheala	trp	BICARB
4	meth	ifabp	LYMFO	LEUCO	NEUTRO	ifabp	MCV	ala	gln	trp
5	ala	mcp1	ifabp	ifabp	ifabp	KREAT	leu	BICARB	ser	gln
6	GGT	LYMFO	il8	NEUTRO	tyr	NEUTRO	HB	ile	BICARB	ser
7	trp	il8	ile	asp	TROMBO	BICARB	val	factord	factord	factord
8	factord	factord	glu	leu	mcp1	LYMFO	trp	ser	leu	ile
9	ifapb	KREAT	BICARB	ALAT	GLU_NS	ip10	meth	CORT	ala	ala
10	tau	NEUTRO	UREUM	ile	LYMFO	factord	POTAS	KREAT	GGT	KREAT

Table 4.11: Top 10 most important Biomarkers selected by the different applied machine learning techniques

Table 4.8 shows the different performance measures provided by the different applied machine learning techniques.

	GLM	RF	XGB	DT_gini	DT_info	RFE	Lasso_min	Lasso_1se	Ridge_min	Ridge_1se
Accur.	0.78	0.77	0.76	0.71	0.72	0.81	**0.84**	0.77	0.77	0.77
Sensit.	0.94	0.97	0.93	0.83	0.86	1.00	0.98	0.99	0.97	0.99
Specif.	0.30	0.17	0.26	0.35	0.30	0.22	0.39	0.13	0.17	0.13
PPV	0.80	0.78	0.79	0.79	0.79	0.80	0.83	0.78	0.78	0.78
NPV	0.64	0.67	0.55	0.40	0.41	1.00	0.90	0.75	0.67	0.75
Prev.	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75
AUC	**0.72**	0.63	0.69	0.59	0.58	0.61	0.69	0.56	0.57	0.56

Table 4.12: Performance measures of the different applied machine learning techniques

Chapter 5

Evaluation

5.1 Discussion

In this section, different discussion points are detailed about the included machine learning techniques in the experiment. The order of presenting them simply reflects the logical order in which they occurred during the workflow process.

5.1.1 High-dimensional data

The cut-off for a data set to be considered high-dimensional is subject to opinion. Initially, the data set contained 13 variables and almost 60000 observations in a stacked (long) format. After cleaning and transforming it to a (wide) format, the resulting data set contained 64 parameters and 590 observations. Based on these dimensions, the transformed data set is considered to be a high-dimensional one.

5.1.2 Correlations

When evaluating the subsets chosen by the candidate machine learning techniques, correlations between the biomarkers were not taken into consideration. This certainly limits the validity of the results as certain biomarkers highly correlate with each other. If multiple of those biomarkers are selected by a model, the effect of individual biomarkers are biased. To assess the found correlations, I consulted with a domain expert in the research field of microbiology. Based on his opinion, the found correlations align with our existing physiological knowledge. To increase the validity of the research, it is advised to carefully consider correlations present in the data. According to De Silva et. al [8], if highly correlated features are present, individual features may exhibit similar performance to the collective feature subset. If there are perfectly correlated features in the data, they can be considered redundant. Including them in the model results in no additional information. Therefore, using only the non-redundant features will improve performance. [10]

5.1.3 Binomial classification

Normalizing the concentration values to the baseline level allowed to handle the problem as a binomial classification, instead of a multinomial one. In addition,

normalization also increased the meaningfulness of the results, as absolute values of the rest protocol by themselves do not have much value added within the current experimental setting.

5.1.4 Feature selection

De Silva et. al [8] propose some valid arguments about the relevance of features in a data set. "An irrelevant feature carries no useful information in describing the relationships of the underlying data. However, a feature that is irrelevant by itself may become useful when considered in combination with some other features." This is why features were not introduced and analyzed individually, but as subsets with different feature selection methods, considering their combined effect.

Field et. al [9] warn about the bias-variance trade-off. "There is also the danger of over-fitting (having too many variables in the model that essentially make little contribution to predicting the outcome) and underfitting (leaving out important predictors) the model."

From the presented machine learning techniques, the Lasso provides a good example for this. Figure 5.1 shows the change of model performance, measured with the Area Under the Curve (AUC), as a function of change in the number of predictors included in the model.

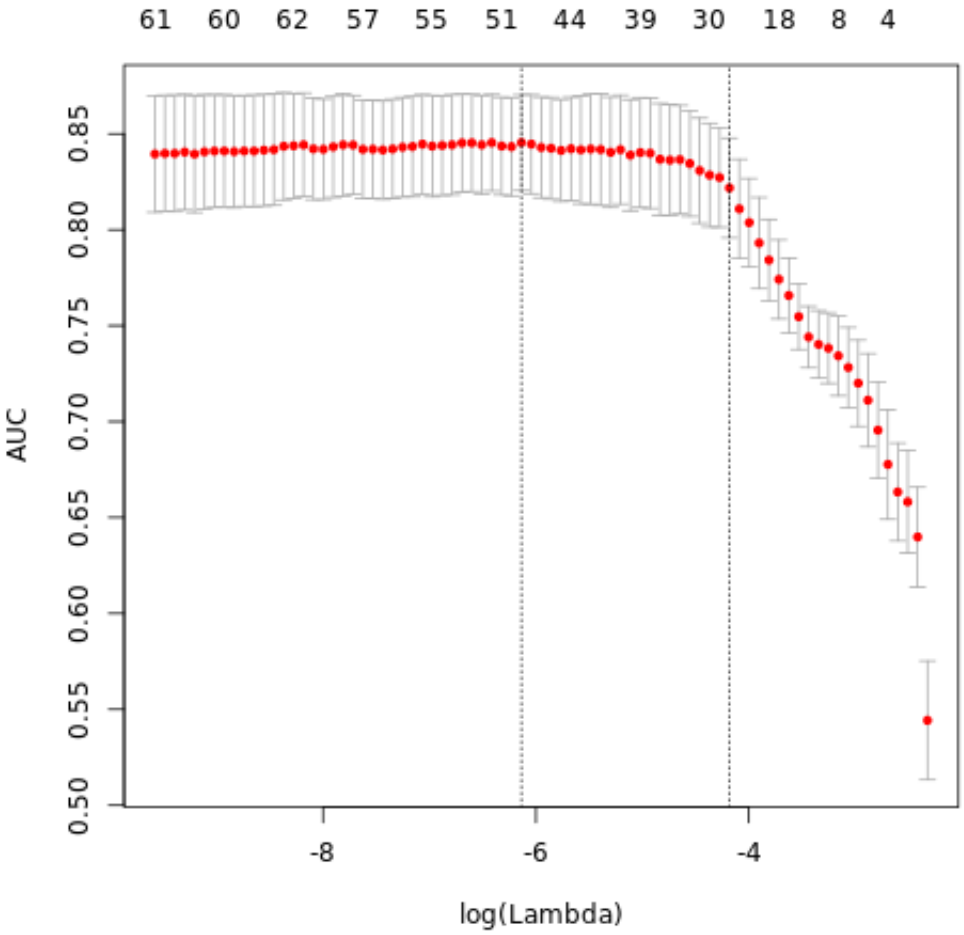

Figure 5.1: Change of model performance (AUC) by change in the number of predictors included in the model

As it can be seen from the graph, by decreasing the number of predictors in the model, the predictive power of the model is decreasing. This could be due to the fact, that there are only few important features in the data set and we can benefit from greedily select from all the features. The two dashed lines represent the "best" model, using the minimum value of the lambda parameter, and the model 1 standard deviation away from that. In Chapter 4.8 the top 10 selected predictors by both cut-off values are presented.

Another example is provided by the Random Forest model with Recursive Feature Elimination applied on it. As it can be seen from Figure 5.2, the highest accuracy was reached with the model with 40 variables. However, using only 3 variables, a 75% accuracy can be achieved.

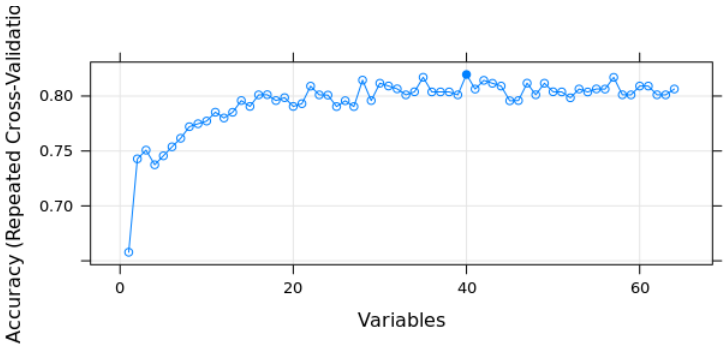

Figure 5.2: Change of Accuracy by the number of predictors in a Random Forest model, using Recursive Feature Elimination

Filters and Wrappers

As a first approach, all the predictors were presented at once for the model. According to Field et. al [9] forced entry is a common approach to present predictors for a model. Unlike stepwise methods, it is not influenced by random variation in the data, which means it has an increased repeatability compared to stepwise methods.

Then different filter and wrapper methods were applied for variable selection. Kuhn [19] argues about the advantages and disadvantages of both methods as follows. "Filter methods are usually more computationally efficient than wrapper methods, but the selection criterion is not directly related to the effectiveness of the model." Furthermore, most filter methods evaluate each predictor separately which results in a higher chance of selecting redundant (i.e. highly-correlated) predictors. Wrapper methods are more computationally intensive as they evaluate many different subsets. As a result, they also tend to overfit more easily.

Lasso and Ridge Regression

As stated by Melkumova et. al, [22], too small lambda values can lead to overfitting, while too large lambda values can lead to underfitting. To find the optimal value of lambda, cross-validation can be used.

The purpose of regularization is to balance between accuracy and simplicity. Using the lambda value 1 standard error away from the minimum results in a simpler model compared to using the minimum lambda value, in the sense that it has less predictors in the model. However this also means, that this simpler model is also less accurate than the one obtained with using the minimum value of lambda.

In theory, the Lasso is capable of performing effective variable selection. However, in practice, it creates excessive biases when selecting significant variables. Furthermore, it is not consistent in terms of variable selection. [25] This was observed in this study as well. Using different lambda values resulted in entirely different coefficient estimates and hence variables to be selected as best predictors. See Table **??** and Table **??**.

5.1.5 Variance explained by principal components

Within Principal Component Analysis, a so-called Scree plot was created. A Scree plot shows the distribution of total variance in the data explained by each principal component. The principal components are presented by decreasing order of contribution to total variance. Figure 5.3 is the resulting Scree plot of the underlying principal component analysis performed on the data.

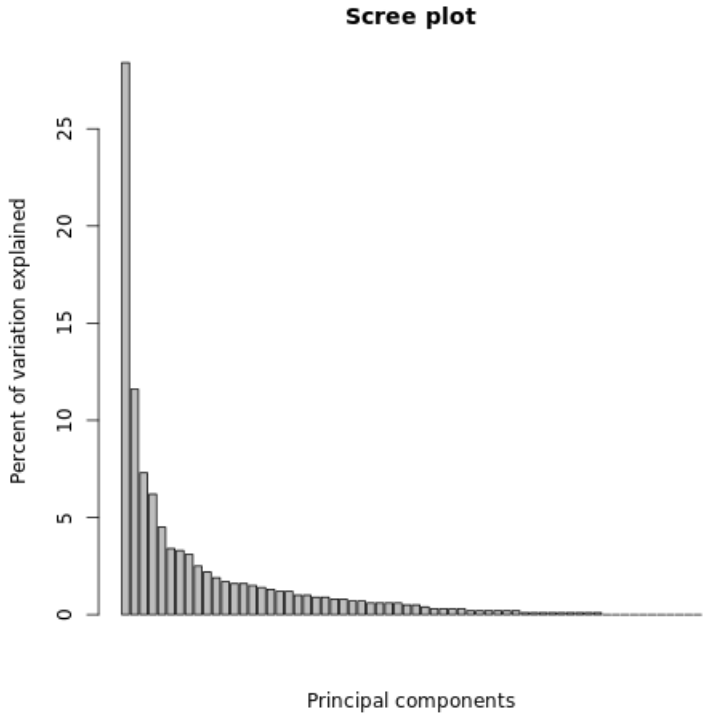

Figure 5.3: Proportion of explained variance by principal components

The Scree-plot shows a steep decreasing trend in the data. As it can be seen, the first principal component explains significantly higher variance in the data than the rest of the principal components. Furthermore, the marginal contribution of every further principal component in the model to explain variance in the data is steadily decreasing. The first 10 principal components explain around 75% of the total variance in the data. The remaining 54 principal components account for 1% or less of the explained variance.

5.1.6 Evaluation of the results

According to Feelders et. al, model interpretation is an important aspect of evaluating models. Often there is a trade-off between model performance and ease of model interpretation. The goal of the modeling task determines which quality measure is considered more important. In the case of this research, the goal of modeling was to find a balance between the two measures. When selecting the candidate machine learning techniques, this perspective was kept in sight.

Guyon et al. [10] have a valid discussion point regarding the evaluation of model performances. When comparing several models, robustness and simplicity should be the guiding principles. The authors suggest that simple, but accurate models should provide stability and good generalization. Even though, more complex models tend to have a better performance, it is often worth resigning from high performance in the favor of less accurate but more stable or simpler models.

5.2 Limitations

5.2.1 Biased population

A remarkable limitation of this research derives from the experimental design of the preliminary experiment. When selecting participating subjects for the study, strict requirements were applied. Among these were the health and general fitness condition of the subjects. For this study, only healthy, trained men were selected from the age group 21-35. Although this was a conscious decision from the researchers, it highly biases the results. When drawing conclusions from this research, one has to interpret them within the scope of the mentioned population. [15]

5.2.2 Noise in the data

Based on initial analysis, it is known that there is both systematic and random noise in the data, which decreases the comprehensibility and interpretability of the results. Therefore, when generalizing based on conclusions drawn from the experiment one has to be careful.

5.2.3 The Borg scale

For measuring the perceived strenuousness of the training, the Borg scale was used. However, it is a widely applied measure for such physiological purposes, the validity of it is argued, since it tries to interpret subjective measurements in an objective way.

In this study I attempt to find further possible connections between objective measurements and the perceived subjective effect of them. Using the Borg scale as class label implies a threat to the validity of the results.

5.2.4 Correlations

As previously discussed, correlations between biomarkers are present in the data. They were investigated from a data-driven perspective, however, their physiological meaningfulness would require further consultation with domain experts. Since the effect of the found correlations were not taken into consideration for selecting and presenting the top 10 biomarkers selected by each model, it is a limitation of my research and the results.

5.2.5 Candidate Machine Learning techniques

One of the limitations of this research is the choice of the applied machine learning techniques. The entire field of machine learning is relatively young, newer and newer techniques still emerge every day. For resolving the problem of this research, many different machine learning techniques could have been used with countless tuning possibilities. Presenting them all is infeasible and not very meaningful. Therefore, I had to make a selection from the available techniques. The criteria for model selection was a balance between performance and interpretability. To establish the balance, both more flexible and more complex models were selected.

Chapter 6

Conclusion

6.1 Conclusions

This chapter presents the conclusions drawn based on the evaluated results. First, some general conclusions are drawn, then the Main Research Question and the related sub-questions are answered.

6.1.1 General conclusions

As a general conclusion, it can be said that different machine learning techniques ranked the importance of biomarkers -reflecting the perceived strenuousness of different exercise levels- differently.

Using the top 10 most expressive biomarkers selected by each model, 39 different biomarkers were selected out of the total 64. The most frequently occurred one was "factord" selected by 7 models. Biomarkers "trp" and "CORT" were both selected by 6 of the models. "ifabp", "LEUCO" and "BICARB" were selected by 5 of the models.

Tree-based models showed some similarity in both the selected biomarkers and their relative importance. The most frequently occurring biomarkers were: "CORT", "ifabp" and "LEUCO" selected by all tree-based models. "LYMFO" and "NEUTRO" were selected by 4 of them and "factord" was selected by 3 of the models.

Logistic regression selected slightly different biomarkers, than the rest of the models, however it provided the best accuracy.

Lasso selected entirely different subsets of biomarkers using two different values of the lambda parameter. The subsets selected by ridge resgression, using two different values of the lambda parameter, show more similarities with 7 out of 10 biomarkers selected by both versions.

In general, the predictive power of the applied machine learning techniques do not vary much. Table 4.8 shows the different performance measures of the applied machine learning techniques.

As it can be seen, Accuracy values vary within the range 71% - 84%. Sensitivity measures show higher variance including values between 83% and 100%. The specificity measures of the models are significantly lower with values between 13% and 39%. This can be caused by the class imbalance, as roughly 3/4 of the cases belonged to the majority class, that was considered "positive". This is also expressed

by the "Prevalence" measure.

6.1.2 Answers to the Research Questions

The Main Research Question of this research was:

Main RQ: "Can we devise a method applying machine learning techniques that relate levels of exercise to self-perceived intensity of training?"

The answer to the main research question is yes, we can devise a method applying machine learning techniques to resolve this problem.

As described in this thesis, measures of the Borg scale were used as labels to express the self-perceived intensity of different exercise levels. Using a label enabled to approach the problem in a supervised way. This resulted in applying different machine learning techniques for selecting subsets of the biomarkers.

SQ1: "Which machine learning techniques can be used to find a minimal set of biomarkers that relate levels of exercise to self-perceived intensity of training?"

To give an exhaustive answer to this question is difficult. The available spectrum of machine learning techniques for such classification problems is vast.

The criteria for choosing the techniques outlined in this thesis was a balanced mix of model performance and model interpretability. To establish the balance, both more flexible, and more complex models were selected.

Furthermore I paid attention to apply different approaches for feature selection, i.e. filters, wrappers and embedded methods. Table 6.1 shows the applied machine learning techniques in this research.

Logistic regression
Random forest
Boosting
Decision tree with the Gini-index
Decision tree with Entropy
Recursive feature elimination on RF
Lasso using minimum lambda
Lasso using 1 standard error lambda
Ridge regression using minimum lambda
Ridge regression using 1 standard error lambda

Table 6.1: Candidate Machine Learning techniques for feature selection

SQ2: "Which machine learning techniques provide the best results in the scope of accuracy, sensitivity and specificity?"

From the applied machine learning techniques, different ones provided the best

results regarding the different performance measures. In the scope of accuracy, the logistic regression model performed best with 78% of the cases predicted correctly. The highest sensitivity, 97% was provided by the Random Forest model. The highest specificity, 35% was achieved by the Decision Tree using the Gini-index.

SQ3: "Does applying feature selection improve the performance measures?"

The answer to this question is yes, applying feature selection improves the performance measures. Recursive feature elimination on the random forest model yielded a 81% accuracy. It was even outperformed by the Lasso on logistic regression with its 84% accuracy. The sensitivity of the Random Forest model was outperformed by the same model when Recursive Feature Elimination was applied on it. It gives a 100% sensitivity, which is certainly overfitting (the model still uses 40 predictors, see Figure 5.2). Regarding specificity, the Lasso with the minimum value of lambda achieves 39%.

Regarding the area under the ROC curve, the best result was achieved using the full logistic regression model with an AUC=0.72. This value is higher, than any of the AUC measures provided by the applied feature selection techniques.

6.2 Future work

6.2.1 Domain specific problem

A logical follow-up step would be to independently evaluate the results with different domain experts. The main aspects of this evaluation would be existing domain-specific knowledge and the presented correlations between the measured biomarkers.

If different domain experts would interpret the results based on previously agreed confidence intervals and significance levels, then their conclusions could be compared. This way, selection bias of the machine learning techniques could be accounted for. Final conclusions could be drawn with a higher confidence based on these compared individual conclusions, which would increase the overall validity of the experiment.

The aim of of this research was to understand why some exercise protocols are perceived strenuous and what are the biomarkers that reflect it.

The subjects of the preliminary experiments were restricted to healthy young men. This means, the capability to make generalized conclusions is limited. A counter-pole is recommended to establish more solid benchmark levels of biomarkers, based on which, further conclusions can be drown.

To this end, in follow-up experiment, individuals with a known disease problem would go through the same protocols, and the same biomarkers would be measured. This way, both the reproducibility and the validity of the research could be improved.

6.2.2 Data Science related problem

The available spectrum of machine learning techniques for resolving such classification problems is vast. Depending on different goals in mind, different techniques

can be proposed for the same problem.

The criteria for choosing the techniques outlined in this thesis was a balanced mix of model performance and model interpretability.

A follow-up study is recommended, putting more weight to either model performance or model interpretability. This way, the odds of finding the most expressive biomarkers that relate self-perceived intensity of different exercise levels can be increased.

To try to further improve the results, K-fold cross-validation could be performed with different folds.

References

[1] *Avicenna hackathon*. URL: https://www.avicennahackathon.nl/ (visited on 07/12/2018).

[2] Thom Baguley. "Understanding statistical power in the context of applied research". In: *Applied ergonomics* 35.2 (2004), pp. 73–80.

[3] Gunnar A Borg. "Psychophysical bases of perceived exertion". In: *Med sci sports exerc* 14.5 (1982), pp. 377–381.

[4] S. Meta Brown. *What IT Needs To Know About The Data Mining Process*. URL: https://www.forbes.com/sites/metabrown/2015/07/29/what-it-needs-to-know-about-the-data-mining-process/#1fc3daa1515f (visited on 07/12/2018).

[5] Pete Chapman et al. "CRISP-DM 1.0 Step-by-step data mining guide". In: (2000).

[6] Jacob Cohen. "A power primer." In: *Psychological bulletin* 112.1 (1992), p. 155.

[7] Thomas H Davenport and DJ Patil. "Data scientist". In: *Harvard business review* 90.5 (2012), pp. 70–76.

[8] Anthony Mihirana De Silva and Philip HW Leong. *Grammar-based feature generation for time-series prediction*. Springer, 2015.

[9] Andy Field, Jeremy Miles, and Zoë Field. *Discovering statistics using R*. Sage publications, 2012.

[10] Isabelle Guyon and André Elisseeff. "An introduction to variable and feature selection". In: *Journal of machine learning research* 3.Mar (2003), pp. 1157–1182.

[11] Alan R Hevner. "A three cycle view of design science research". In: *Scandinavian journal of information systems* 19.2 (2007), p. 4.

[12] Gareth James et al. *An introduction to statistical learning*. Vol. 112. Springer, 2013.

[13] Lonneke M JanssenDuijghuijsen et al. "Adaptation of exercise-induced stress in well-trained healthy young men". In: *Experimental physiology* 102.1 (2017), pp. 86–99.

[14] Ian Jolliffe. "Principal component analysis". In: *International encyclopedia of statistical science*. Springer, 2011, pp. 1094–1096.

[15] Shirley Kartaram et al. "Plasma citrulline concentration, a marker for intestinal functionality, reflects exercise intensity in healthy young men". In: *Clinical Nutrition* (2018).

[16] *R popularity*. URL: `https://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html` (visited on 12/05/2018).

[17] Ron Kohavi and George H John. "The wrapper approach". In: *Feature extraction, construction and selection*. Springer, 1998, pp. 33–50.

[18] Max Kuhn. *caret: Classification and Regression Training*. R package version 6.0-78. 2017. URL: `https://CRAN.R-project.org/package=caret`.

[19] Max Kuhn and Kjell Johnson. *Applied predictive modeling*. Vol. 26. Springer, 2013.

[20] Breiman Leo et al. "Classification and regression trees". In: *Wadsworth International Group* (1984).

[21] Computing Machinery. "Computing machinery and intelligence-AM Turing". In: *Mind* 59.236 (1950), p. 433.

[22] LE Melkumova and S Ya Shatskikh. "Comparing Ridge and LASSO estimators for data analysis". In: *Procedia Engineering* 201 (2017), pp. 746–755.

[23] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017. URL: `https://www.R-project.org/`.

[24] Jonathon Shlens. "A tutorial on principal component analysis". In: *arXiv preprint arXiv:1404.1100* (2014).

[25] Jonathan Taylor and Robert J Tibshirani. "Statistical learning and selective inference". In: *Proceedings of the National Academy of Sciences* 112.25 (2015), pp. 7629–7634.

[26] Marc Teunis and Jan-Willem Lankhaar. *gramlyr: Demonstration of Reproducible Data Analysis in Life Sciences*. R package version 0.1.0. 2017.

[27] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), pp. 267–288.

[28] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: `http://ggplot2.org`.

[29] Hadley Wickham et al. "Tidy data". In: *Journal of Statistical Software* 59.10 (2014), pp. 1–23.

[30] Hadley Wickham. *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.2.1. 2017. URL: `https://CRAN.R-project.org/package=tidyverse`.

[31] Hadley Wickham and Garrett Grolemund. *R for data science: import, tidy, transform, visualize, and model data*. " O'Reilly Media, Inc.", 2016.

[32] Hadley Wickham et al. *dplyr: A Grammar of Data Manipulation*. http://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

[33] Roel J Wieringa. *Design science methodology for information systems and software engineering*. Springer, 2014.

[34] *Machine Learning*. URL: `https://en.wikipedia.org/wiki/Machine_learning` (visited on 11/27/2018).

[35] Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC, 2012.

Appendices

Appendix A

Session Info

- R version 3.4.2 (2017-09-28), `x86_64-pc-linux-gnu`
- Running under: `Ubuntu 16.04.3 LTS`
- Matrix products: default
- BLAS: `/usr/lib/libblas/libblas.so.3.6.0`
- LAPACK: `/usr/lib/lapack/liblapack.so.3.6.0`
- Base packages: base, datasets, graphics, grDevices, grid, methods, stats, utils
- Other packages: caret 6.0-78, clusterCrit 1.2.8, corrr 0.3.0, DMwR 0.4.1, dplyr 0.7.99.9000, factoextra 1.0.5, forcats 0.3.0, foreach 1.4.4, Formula 1.2-3, gdtools 0.1.7, ggplot2 3.1.0, ggthemes 3.4.0, glmnet 2.0-16, Hmisc 4.1-1, lattice 0.20-38, Matrix 1.2-15, pROC 1.13.0, purrr 0.2.5, readr 1.1.1, reshape2 1.4.3, rpart 4.1-13, rpart.plot 3.0.4, stringr 1.3.1, survival 2.43-1, svglite 1.2.1, tibble 1.4.2, tidyr 0.8.2, tidyverse 1.2.1, xgboost 0.71.2
- Loaded via a namespace (and not attached): abind 1.4-5, acepack 1.4.1, assertthat 0.2.0, backports 1.1.2, base64enc 0.1-3, bitops 1.0-6, broom 0.5.0, caTools 1.17.1.1, cellranger 1.1.0, checkmate 1.8.5, class 7.3-14, cli 1.0.1, cluster 2.0.7-1, codetools 0.2-15, colorspace 1.3-2, compiler 3.4.2, crayon 1.3.4, curl 3.2, CVST 0.2-1, data.table 1.11.8, ddalpha 1.3.1.1, DEoptimR 1.0-8, digest 0.6.18, dimRed 0.1.0, DRR 0.0.3, foreign 0.8-71, gbm 2.1.4, gdata 2.18.0, ggrepel 0.8.0, glue 1.3.0, gower 0.1.2, gplots 3.0.1, gridExtra 2.3, gtable 0.2.0, gtools 3.8.1, haven 1.1.2, hms 0.4.2, htmlTable 1.12, htmltools 0.3.6, htmlwidgets 1.3, httr 1.3.1, ipred 0.9-6, iterators 1.0.10, jsonlite 1.5, kernlab 0.9-27, KernSmooth 2.23-15, knitr 1.20, latticeExtra 0.6-28, lava 1.6, lazyeval 0.2.1, lubridate 1.7.4, magrittr 1.5, MASS 7.3-51.1, ModelMetrics 1.1.0, modelr 0.1.2, munsell 0.5.0, nlme 3.1-137, nnet 7.3-12, pillar 1.3.0, pkgconfig 2.0.2, plyr 1.8.4, prodlim 1.6.1, quantmod 0.4-13, R6 2.3.0, RColorBrewer 1.1-2, Rcpp 0.12.18, RcppRoll 0.2.2, readxl 1.1.0, recipes 0.1.2, rlang 0.2.2.9001, robustbase 0.93-2, ROCR 1.0-7, rstudioapi 0.8, rvest 0.3.2, scales 1.0.0, sfsmisc 1.1-1, splines 3.4.2, stats4 3.4.2, stringi 1.2.4, tidyselect 0.2.5, timeDate 3042.101, tools 3.4.2, TTR 0.23-4, withr 2.1.2, xml2 1.2.0, xts 0.11-2, yaml 2.2.0, zoo 1.8-4

Appendix B

Decision Tree based on the Gini-index

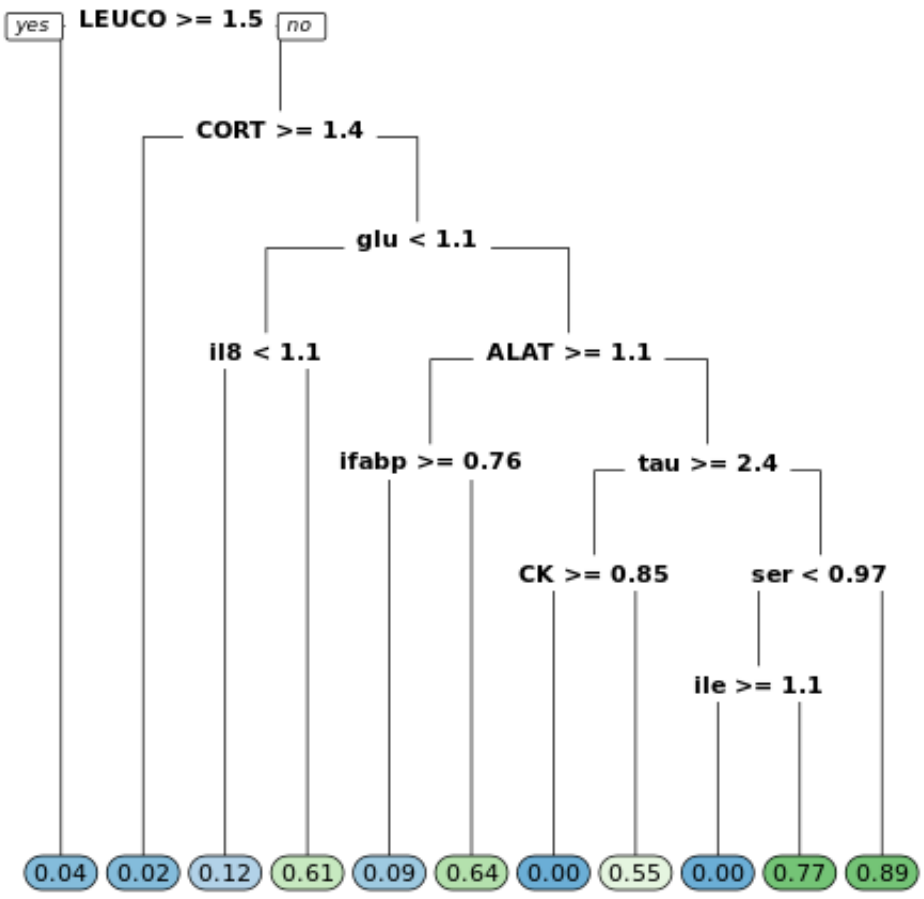

Figure B.1: Decision Tree based on the Gini-index

Appendix C

Decision Tree based on Information gain

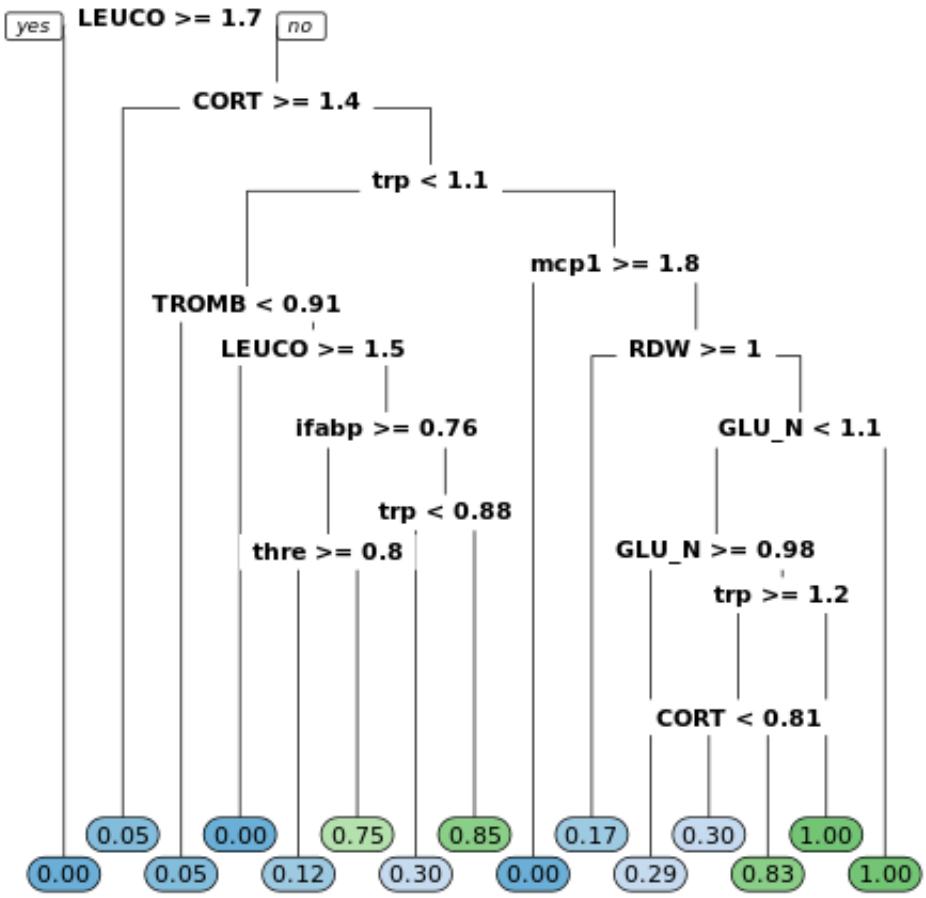

Figure C.1: Decision Tree based on Information gain