TABLE OF CONTENTS

1 Introduction

Imagine being able to correctly predict those individuals who will graduate from pilot training, Harvard medical school, or even NASA's astronaut program. This would have a large impact on the world as it would reduce individuals being disappointed due to being rejected, and as it would reduce the use of resources for organizations. Also, selecting the right people would positively influence the organization to which people are selected. For example, some organizations are willing to pay millions of dollars to hire the right researcher, athlete, or CEO (e.g., Metz, 2018). However, such predictions are remarkably difficult to make. A single highly improbable event is all that is needed to significantly alter the outcome (Taleb, 2010). For example, a recruit in the selection program of the special forces can be the best in their class, but drop out due to a close relative getting sick or a wrong step in a rabbit hole. With that in mind, is it still possible to predict who will drop out and who will graduate? The field that investigates this question is called *personnel selection*.

1.1 Personnel Selection

The field of personnel selection is built on a long history of research in psychology and statistics. One of the pioneers in this field was Francis Galton in the 1880s, when he invented regression and correlation analysis, and invented the term psychometrics (Gillham, 2001). He was interested in measuring mental ability to select capable individuals, but his tests were mostly limited to tests for mental imagery, sight, hearing, and bodily strength and size (Gillham, 2001). A test that more closely resembles today's personnel selection was the Binet-Simon Intelligence scale in 1904, which was used to select French students "capable of regular schooling". This method of testing

for intelligence was soon applied in military selection during the first world war (Terman, 1918). Soon after, personality tests were also developed for selection. The Woodworth Personal Data Sheet was used to screen recruits for the US Army for shell shock (Woodworth, 1918). Subsequent personality tests were based on factor analysis, a method designed to identify underlying factors that account for the patterns in responses to a set of questions. For example, 4000 affect terms from the English dictionary were reduced to the 16 personality factor questionnaire (Cattell et al., 1970). More recently, these 16 personality factors were reduced to the Big Five personality traits, which are now considered the most important factors in personality (Costa & McCrae, 1992).

It is notably a recurring theme here that the military has been a source of innovation. Alan Turing built one of the first computers for the British military, and was one of the first to come up with the notion of Artificial Intelligence (AI) in the late 1940s (Turing, 1950). Next, the US Defense Advanced Research Projects Agency (DARPA) has funded the precursor to the internet in the 1960s (Abbate, 2000). In the 1970s, they funded the precursor to GPS (Parkinson & Gilbert, 1983) and computer chips (Miller, 2022)[1]. In the 1980s they funded the precursor for modern screens (Florida & Browdy, 1991) and more AI research (Roland & Shiman, 2002). After a lack of progress for AI in that decade, a new system called High Performance Computing (HPC) was introduced in the 1990s for DARPA (Roland & Shiman, 2002).

[1]Apart from many benefits that these and earlier mentioned inventions and inventors have provided, it should be noted that they can also be directly linked to much human suffering. For example, the first chips were used to improve the accuracy of bombs in the Vietnam war, and the intelligence tests were closely linked to the eugenics movement (see, for example, Miller, 2022 or Gillham, 2001 for more information).

A reason why many innovations have come from the military could be that they have large problems that need to be solved quickly (Housel, 2023). For example, the launch of the Sputniks satellite in 1957 by the Soviet Union led to fear in Western nations that they were falling behind in technology. In response, DARPA and NASA were founded in 1958. Currently, one large problem that needs to be solved quickly is selecting the right personnel. It is unclear how to solve this, but one step in the right direction could be to gather and analyze more data, which could lead to better predictions in personnel selection. In order to do so, we needed to develop software to efficiently gather psychological and physical measures of recruits.

1.2 Data Collection and Processing

Our data collection was carried out at and made possible by the Dutch Special Forces (*Korps Commandotroepen* in Dutch). Special forces are elite military units that are trained to perform unconventional, high-risk, and specialized missions. However, dropout rates in special forces selection programs are close to 80% (e.g., Gayton & Kehoe, 2015). The military was interested in identifying factors that could predict dropouts and use that to improve the selection process.

To do so, we collected data from recruits during the training and selection program. During this 16-week program, recruits are trained towards becoming a special forces operator and at the same time subjected to grueling physical and psychological challenges. Since there is no clear distinction between training and selection in the program, we will use the the terms *training* and *selection* interchangeably throughout this thesis. To facilitate the data collection, we developed a custom software system that allowed recruits to complete questionnaires online. Each recruit was assigned a unique login,

granting them access to the system and enabling them to complete the questionnaires. The system was designed to streamline the data collection process, allowing researchers to process and analyze pseudonimized data of the recruits. In total, this data collection resulted in about 60 000 lines of data in the period from 2019 to 2023 (this number includes missing data).

These amounts of data collection would quickly become unmanageable without the use of software. This is not unique to our research. A big part of most academic research is entirely dependent on software (McElreath, 2020a). The collection of more data also has an effect on the data processing. Manually editing data is feasible for studies with a few dozen participants and a few variables, but become impractical for hundreds of participants and a few dozen variables. For the data processing in our research, the Julia programming language (Bezanson et al., 2017) was mainly used. This had benefits, but drawbacks as well. Benefits were that the language is expressive, which means that it is easy to express complex ideas in a few lines of readable code. Julia usually sticks closer to mathematical notation than other languages, which makes it easier to translate mathematical ideas to code. Drawbacks were that that the language is still quite new[2] meaning that a lot of functionality that would have been available in other languages, such as R or Python, was not (yet) available. This forced the fixing of bugs in core Julia packages or creating new Julia packages, which both resulted in a great learning experience. For example, when adding the standard deviation to a data science package, there was a useful discussion which lead to a better implementation[3].

[2] Julia was first released in 2012, whereas Python was first released in 1991 and R in 1995.

[3] Thanks to Anthony Blaom and Okon Samuel in https://github.com/JuliaAI/MLJBase.jl/pull/766.

1.3 Statistics and Machine Learning

We based our research on earlier studies with the similar measures in similar settings, and theoretical knowledge from psychology about personality and sport science. However, much discussion exists on the validity of the typically used statistical analyses. This was challenging for our studies, which were quantitative and hence relied fully on these analyses. The core of the argument in quantitative research is the use of statistics, so researchers who question the validity of these statistics question the core of the research.

Some argue that it is best to stick to the well-known and well-established statistical tests regardless of the criticisms. If we all agree on what tests to use, then we can all use the same objective criteria to evaluate the results. This is like finding an optimial solution in a simplified world, instead of finding a good solution in a more realistic world (Simon, 1979). It is possible that the well-known approaches are not necessarily the best approaches. There could exist better approaches that would lead to more and better scientific insights.

One problem that was pointed out with a well-known approach is that decisions for hypothesis testing should be weighed carefully. For example, instead of sticking to the default p-value of 0.05, it is better to justify why the value was chosen (Lakens et al., 2018). Other researchers propose to switch to Bayesian analyses since it requires researchers to be more explicit about their assumptions (McElreath, 2020b). Even inside this Bayesian-world, there are two camps. One argues for the Bayes factors approach, that resembles frequentists hypothesis testing (e.g., Stefan et al., 2019), while another argues for visualizing the analysis and inspecting the plots (e.g., Tendeiro & Kiers, 2019, Gelman et al., 2021). This latter approach is the most computationally expensive, but does allow for greater flexibility in model definitions and is

arguably more intuitive since model assumptions are more explicit and model interpretations more visual.

With all the these statistical options, it was often not clear which model should be used for which study during our research. This is in line with the "no free lunch" theorem (Wolpert & Macready, 1997). The theory states that there is no single statistical model that is the best fit for all studies. This means that we have to choose the best model for each study, but this is not easy since we do not know which model is the best.

A solution becomes clear when considering the ancient Buddhist story of the blind men and an elephant. In this story, several blind men each touch a different part of an elephant to learn what it is. One feels the trunk and says it is like a thick tree branch, another feels a leg and declares it is like a pillar, another feels the tail and shouted it is like a rope, and so on. Seen separately, each man will come to the wrong conclusions. Hence, the solution is to combine the information from multiple blind men to get a better picture. Or in other words, we should use multiple statistical models to get a better picture of the data.

Although multiple models and variable associations do provide insights, they were not sufficient in practice. Such variable associations belong to the *inference*, or *explanation*, paradigm (Hofman et al., 2021). This paradigm is what Galton and Fisher used a century ago. For example, analyzing which variables are associated with the outcome via a *t*-test is about explaining the data. However, we wanted to apply our research to the selection, but the associations did not indicate how accurate our predictions would be. Moreover, associations can be misleading and suffer from *overfitting*. Overfitting is when a model fits the data too precisely, leading to poor predictions on new data. This is like a student who has seen the exam questions beforehand and can

answer them all, but fails when seeing new questions. There are mathematical ways to estimate overfitting, like the Bayesian Information Criterion, but these are hard to interpret and do not clearly indicate how well the model will do in practice.

To get a better understanding, we turned to the field of data science and machine learning. In this paradigm, known as the *prediction* or *algorithmic* paradigm (e.g., Hastie et al., 2009), the focus is not on explaining the model, but on predicting the outcome. In this paradigm, a model can even completely lack interpretability, i.e., a *black-box*, as long as it predicts well. For example, nobody can fully understand why neural networks, such as modern Large Language Models (LLMs), or human brains, make certain predictions. As long as the model is useful and safe, this lack of understanding is accepted in certain contexts. There is some understanding about the effectiveness required to work with such models or referees, though. For a referee, this is is tested in the form of a long process of graduating through the various levels of refereeing before being allowed to referee a World Cup final. For a machine learning model, this performance is typically tested on old data that the model has not seen before.

A common approach for this is *cross-validation*. Cross-validation works by splitting the data into a *training* and *test* set. The model sees the data in the training set and is asked to predict the data in the test set. Next, the predictions on the test set are compared to the real answers and the model receives a score. This procedure is very similar to how students are tested at universities. The student can see training questions and is then tested and evaluated on a set of test questions. However, a student may be unlucky on the choice of test set. It may be that the chosen test questions are particularly difficult (or easy) and one could thus say that the test has a bias. Since statistical models can easily be

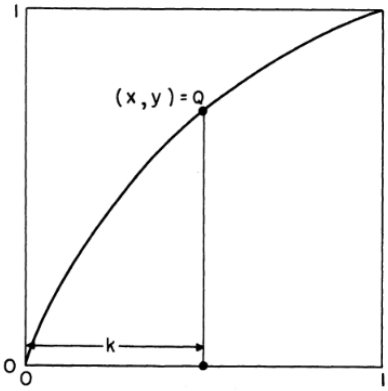
reset (forced to unlearn) and since we nowadays have fast computers, cross-validation can reduce this bias. Instead of choosing one training and test set, the technique works by splitting the data in multiple *folds*. For each fold, a subset of the data is put in the training set and a different subset is put in the test set. Then the model is evaluated in each fold and the scores from each fold are combined into a general score. The aim of this general score is to estimate how well the model (or student) will do in the future when seeing new data (or questions).

Cross-validation provides a single number indicating model performance. However, this number is not informative enough for the purpose of this thesis. For example, in the context of special forces selection, this number would not tell us how many recruits could be selected without making a mistake. Luckily, there is a way to get a clearer estimate, which was invented after the second World War. During the war, radars sent out pulses and received echos. Based on the echos, the British radar operators had to decide whether the echo was a plane or a flock of birds (Neale, 1985). The stakes for these decisions were high. Mistakes could lead to a plane getting through unnoticed or to sending pilots to a flock of birds. After the war, researchers investigated how to systematically evaluate such radar systems. Simply put, the radar systems responded with a signal of a certain strength (a continuous signal) and this had to be converted to a binary decision on whether it is a plane (a binary signal). Given two similar radar systems or two similar radar configurations, how could one decide which one was better? One way is to think about it statistically via the false alarm and detection probabilities (Kaplan & McFall, 1951). Based on these false alarm and detection probabilities, a plot can be created that shows the trade-off between the two. This was then called the

Receiver Operating Characteristic (ROC) curve (Fox, 1953), see Figure 1.1 for an example.

Figure 1.1

An example ROC curve



Note. Image from Fox (1953). Axis labels omitted for clarity.

On this plot, the probability of a detection is plotted on the y-axis and the probability of a false alarm is plotted on the x-axis. These probabilities can be estimated by varying the threshold of the radar system. In essence, this is like asking the question "If we respond only if the signal is above X, how many detections and false alarms do we expect?" For example, setting the radar to be very sensitive will lead to many detections, but also many false alarms. When plotting one line for one radar, radar designers could decide how to configure the radar for the best trade-off between detections and false alarms. When plotting multiple lines for multiple systems, designers could decide which radar system was the best. Nowadays, this last part is often done without plotting because the Area Under the Curve (AUC) can be calculated. A higher AUC means that the system is more accurate overall. Later, the ROC curve was used not only for radar systems, but for many problems where a

continuous signal is converted to a binary decision, like in this thesis where continuous model outcomes are converted to binary prediction decisions.

These were some of the tools that we used in this thesis. The aim being to investigate the data in a statistically sound way. With these tools, we aimed to answer the question of who will make it through the commando training and who will dropout.

1.4 Thesis Chapters

The core of this thesis is split into four parts. In Chapter 2, we focused on identifying personality traits that could differentiate between experienced commandos and ordinary Dutch men, and between successful graduates and dropouts in special forces training. We used the inference paradigm to compare the personality traits of these groups, providing insights into the role of personality in special forces selection.

Chapter 3 marks our transition from the inference to the inference and prediction paradigm. The chapter introduces SIRUS.jl, our implementation of the Stable and Interpretable RUle Sets (SIRUS) algorithm in Julia. SIRUS aims to combine the benefits of decision trees and random forests, offering high interpretability and stability. This chapter details the implementation, interpretability, stability, and performance of SIRUS.jl. We compare its predictive performance to similar models on various small real-world datasets.

In Chapter 4, we used various machine learning techniques, including SIRUS, to predict dropout in special forces selection programs using both physical and psychological data, such as 2800 meters running time and personality traits. This data was collected at one point in time during the first week of the selection program. We again compared the performance, explainability, and stability and showed the benefits of the SIRUS model in

this high-stakes context due to it having good predictive performance while retaining model stability and explainability.

In Chapter 5, we measured psychological and physical stress and recovery states of recruits during the training program. This aimed to find early indicators of dropout. Again, using machine learning techniques, we compared the performance, explainability, and stability of the models. We also estimated the real-world predictive performance of the most suitable model.

Then a final note on why this thesis contains some quotes and ideas from investors such as Warren Buffett, Charlie Munger, Nassim Nicholas Taleb, and Peter Lynch. It might sound like their occupation is unrelated to personnel selection. However, this is not true. All these individuals have excelled in making predictions, and often have an independently verified track record to prove it. For example, Buffett's success is based on selecting which companies will do well in the future. Just like personnel selection, this is a difficult task. Both personnel selection and company selection are about predicting those that will operate successfully in a complex environment. Just like in personnel selection, companies are also fragile: a single negative event can lead to a permanent bankruptcy. The accuracy of his selection decisions can be inspected by looking at the value of his public company Berkshire Hathaway. If you would have bought one share valued at $19 in Berkshire in 1965, then this would today be worth more than half a million (Class A shares are worth $615,591 at the time of writing). It is not possible for him to have faked these predictions. During all these decades, his financial statements have been verified by auditors, the tax authorities, and the Securities and Exchange Commission. This is why Buffett is known as the "Oracle of Omaha", and why ideas from investors seem relevant to personnel selection.

CHAPTER 2

Personality Traits of Special Forces Operators: Comparing Commandos, Candidates and Controls

This chapter is based on:

Huijzer, R., Jeronimus, B. F., Reehoorn, A., Blaauw, F. J., Baatenburg de Jong, M., De Jonge, P., & Den Hartigh, R. J. R. (2022). Personality Traits of Special Forces Operators: Comparing Commandos, Candidates, and Controls. *Sport, Exercise, And Performance Psychology*, 11(3), 369–370. https://doi.org/10.1037/spy0000296

Abstract

Dutch special forces operators, also known as *commandos*, perform in mentally and physically tough environments. An important question for recruitment and selection of commandos is whether they have particular personality traits. To answer this question, we first examined differences in personality traits between 110 experienced Dutch male commandos and a control sample of 275 men in the same age range. Second, we measured the personality traits at the start of the special forces selection program, and compared the scores of candidates who later graduated (n = 53) or dropped out (n = 138). Multilevel Bayesian models and t-tests revealed that commandos were less neurotic (d = −0.58), more conscientious (d = 0.45), and markedly less open to experiences (d = −1.13) than the matched civilian group. Furthermore, there was a tendency for graduates to be less neurotic (d = −0.27) and more conscientious (d = 0.24) than dropouts. For selection, personality traits do not appear discriminative enough for graduation success and other factors need to be accounted for as well, such as other psychological constructs and physical performance. On the other hand, these results provide interesting clues for using personality traits to recruit candidates for the special forces program.

2.1 Introduction

Dutch special forces operators, also known as commandos, perform in tough high stakes contexts that require specific physical, mental and emotional characteristics (Brailey, 2005). Commandos must remain focused and calm in combat situations despite overwhelming intense smells, sounds and images, and depend with their lives on their team's functioning. Furthermore, they work under conditions of extreme threat, isolation and complexity, and often need to interact with other cultures in politically sensitive situations (Picano et al., 2002). The individual characteristics needed to operate in such situations are typically operationalized in terms of personality dimensions; what we feel, think, need, want and do. Our research aim was to identify whether there are personality traits that are characteristic for commandos (Banks, 2006; Ones et al., 2007).

Personality of Commandos

In contemporary psychology, the highest level of the personality hierarchy is summarized in terms of five broad trait dimensions (the Big Five): neuroticism, extraversion, openness to experience, agreeableness and conscientiousness (John et al., 2010, see also Table 2.1). Since the second world war, the United States of America (U.S.) selects commandos on their emotional and interpersonal traits (emotional stability, social relations and security), intelligence processing (effective IQ, observing and reporting) and agency/surgency (motivation for assignment, energy and initiative leadership, physical ability; see Banks (2006); Picano et al. (2002). This procedure suggests that emotional stability (low neuroticism) and extraversion (activity and sociability) are key personality competencies for success in high-risk operational personnel, next to cognitive abilities. However, so far, few studies examined the personality

traits of commandos and quantified how they actually compare to civilian samples.

Table 2.1

Definition of Personality Based on the Big Five

Big Five Domain	High scoring individuals tend to be ...
Neuroticism	emotionally unstable, anxious, self-conscious, vulnerable, and experiencing negative affect.
Extraversion	sociable, assertive, energetic, excitement seeking, risk-taking, and experiencing positive affect.
Openness	perceptive, creative, reflective, flexible, curious, and appreciative of fantasy, aesthetics, and novelty.
Agreeableness	kind, cooperative, altruistic, trustworthy, trusting, generous, and empathic.
Conscientiousness	ordered, dutiful (norms/rules), self-disciplined, reliable, persistent, and achievement oriented.

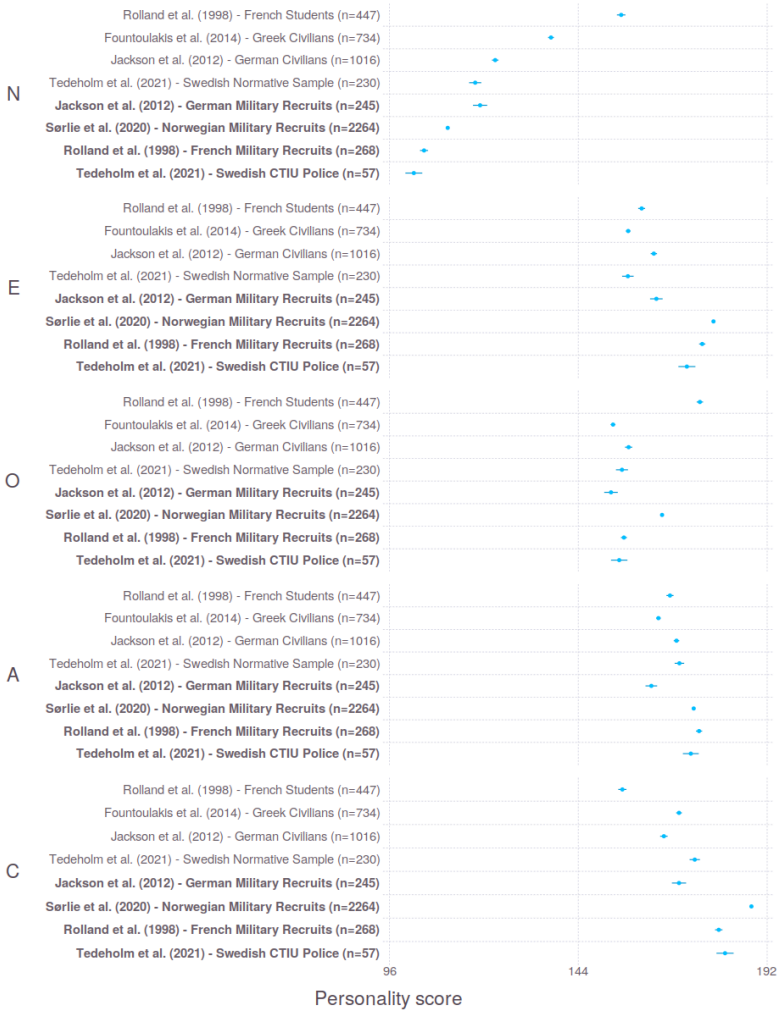
In one of the previous studies, personality trait scores among 139 U.S. Navy Sea-Air-Land (SEAL) operators were compared to scores of U.S. civilians (Braun et al., 1994). In line with the above mentioned key personality competencies, SEALs reported lower average neuroticism and agreeableness scores than civilians, but higher conscientiousness and extraversion. Additionally, more experienced SEALs reported higher conscientiousness. Although research on the personality traits of commandos is scarce, several studies examined Big Five measures of other military personnel and police officers who work in high stakes contexts. For instance, a sample of 57 Swedish counterterrorism intervention unit police officers showed lower mean neuroticism

scores (Cohen's $d = 0.70$) and more extraversion ($d = 0.70$) and conscientiousness ($d = 0.40$) than the general Swedish population (Tedeholm et al., 2021). Furthermore, a comparison of 268 French military candidates with 447 students indicated that candidates reported lower openness ($d = 2.04$) than the students (Rolland et al., 1998). Similarly, people who entered the German military were marked by lower openness ($d = 0.15$ with a propensity-score model) than people who did not enter the military (Jackson et al., 2012). The large differences between the studies in terms of effect sizes could arise from differences in sampling or methodology. For example, Jackson et al. (2012) used propensity score matching which may have increased bias and imbalance (King & Nielsen, 2019).

In Figure 2.1, we visually summarized previous studies of personality traits of workers in high stakes contexts, such as special forces units (Fountoulakis et al., 2014; Jackson et al., 2012; Rolland et al., 1998; Sørlie et al., 2020; Tedeholm et al., 2021). This shows that high stakes context workers score relatively high on conscientiousness and low on neuroticism compared to control groups. Differences in the other personality traits were less consistent. This could indicate that there is not strictly one personality trait that allows people to be proficient in high stakes contexts. However, little is known about how commandos and civilian men with a similar age and background actually differ in their personality traits. Therefore, our first research question was: how do the personality traits of experienced commandos differ from those of a matched sample of civilians in the general population?

Figure 2.1

An Informal Review of Personality Traits of Workers in High Stakes Contexts Compared to Civilians



Note. The means and standard errors (SEs) for personality scores obtained in previous research. The lower four studies focused on high stakes military contexts and the upper four on civilian populations (control groups). The means and SEs are similar to independent samples t-tests. Scores were transformed to the range $[1, 5]$, resulting in a total score in the range $[48, 240]$. For example, studies scored in the range [0,4] have lower bound $l = 0$ and upper bound $u = 196$. Any mean m in the range $[l, u]$ was transformed to a mean m' in the range $[48, 240]$ via $m' = 48 + \frac{m-l}{u-l} \cdot (240 - 48)$. Similarly, any standard deviation s was transformed to s' via $s' = \frac{s}{u-l} \cdot (240 - 48)$. The ranges for Fountoulakis et al. (2014), Jackson et al. (2012), Sørlie et al. (2020), and Tedeholm et al. (2021) were respectively $[0, 192]$, $[0, 4]$, $[0, 192]$ and $[0, 192]$, and obtained by author correspondence.

Assessment and Measures

Next to the question of how commandos differ from civilians, we examined whether personality traits of candidates, who successfully completed the selection program, differed from those who dropped out. Personality assessments are often part of the special forces selection procedure (e.g., Hartmann et al., 2003; Saxon et al., 2020), but relatively little scientific research has been conducted on this topic. Specifically focusing on the Big Five domains, a study by McDonald et al. (1990) shows that U.S. graduates scored lower on neuroticism than the dropouts. Another U.S. study on reconnaissance marines found that higher extraversion was associated with graduation (Saxon et al., 2020). Other studies focused on the Big Five personality traits on the facet level, which are more narrow personality dimensions. For example, Picano et al. (2002) studied elite military personnel screened for a non-routine military assignment and identified two facet traits that predicted success; "activity" in the extraversion domain (E4, being lively) and "straightforwardness" in the agreeable domain (A2; having high morale). Training completion in the Norwegian naval special forces was not associated with any of the Big Five domains or facets (Hartmann et al., 2003; Hartmann & Grønnerød, 2009), in discord with the findings by McDonald et al. (1990), Picano et al. (2002), and Saxon et al. (2020).

When looking at less extreme contexts, a lower neuroticism score and a higher agreeableness score were found to be related to graduation in the Canadian forces basic training (Lee et al., 2011). In the Netherlands, a large study of multiple datasets showed that successful military candidates were more likely to score high on extraversion, conscientiousness, agreeableness and openness, and low on neuroticism (Van der Linden et al., 2010). A meta-analysis on military aviators showed that lower neuroticism and higher

extraversion scores were related to training success (Campbell et al., 2010). Despite the frequent measurement of personality in special forces training programs, the degree to which the outcomes can be used for selection in such programs remains unclear. Overall, most research suggests that successful commando candidates were less likely to be neurotic and more likely to be extraverted and agreeable (e.g., Jackson et al., 2012), but not all commando studies supported these differences (e.g., Hartmann & Grønnerød, 2009). Therefore, the present study examines whether and which personality differences predict success during the commando selection procedure in the Netherlands. More specifically, we examined whether graduates and dropouts of the special forces selection program could be distinguished based on their measured personality traits.

The Current Study

The purpose of the current study was to examine whether measured personality traits differ between (1) commandos and civilians and (2) graduates and dropouts. We therefore examined the personality of a sample of Dutch male commandos, a matched control group from the Dutch population, and candidates in the special forces selection program. Our first hypothesis was that commandos reported lower neuroticism, higher conscientiousness and more extraversion than civilians (see Braun et al., 1994; Rolland et al., 1998; Tedeholm et al., 2021). No differences in agreeableness and openness were expected. Our second hypothesis was that graduates report lower neuroticism than dropouts (Campbell et al., 2010; Lee et al., 2011; McDonald et al., 1990) and more extraversion and agreeableness (Campbell et al., 2010; Hartmann et al., 2003; Lee et al., 2011; Picano et al., 2002; Saxon et al., 2020). No specific expectations were set for openness to experience and conscientiousness.

2.2 Method

Participants

Data from the, exclusively male, commandos and candidates were obtained via the Commando Corps of the Royal Netherlands Army. Commandos were approached by email, including an information letter about the study. We received active consent from 110 experienced commandos, that is, commandos who successfully finished the entire special forces training. The matched controls were derived from a large Dutch crowd-sourced dataset (Van der Krieke et al., 2016) from which 275 males aged 18 to 35 were selected (M_{age} = 27.7, SD_{age} = 4.62). New candidates were invited to participate in this study during their pre-selection training. Both candidates and instructors were informed that participation was completely voluntary and that their participation and results would not be used for selection purposes. All candidates actively consented to participate in the study and the procedure was approved by the institutional review board with code PSY-1920-S-0512. Of the 223 candidates who started the selection period, 53 graduated (M_{age} = 25.2, SD_{age} = 2.70) and 138 dropped out for non-medical reasons (M_{age} = 25.9, SD_{age} = 2.96). We excluded 32 participants who dropped out for medical reasons. The selection is based on a combination of objective and subjective measures. Candidates can drop out for non-medical reasons if they do not meet the physical requirements at any point during the selection, if they are caught in an offense such as stealing, or if the instructors unanimously agree that a person is unfit to become an operator. Furthermore, the sample sizes were limited by the number of participating operators and the number of candidates who started the selection in the period in which we collaborated

with the army. Given the sensitivity of the samples that we studied, more detailed descriptions were not provided.

Procedure

For both the commandos and candidates, participation occurred via our Your Special Forces platform (https://yourspecialforces.nl), which was specifically built for the purpose of the research project. The commandos received instructions and credentials via email, and were invited to participate in the questionnaire during their work hours. For the candidates, data collection took place at the training camp. In the first week of the selection, participants completed the assessments using tablets in a large room which was set up like a traditional classroom. Once participants entered the room, they were informed about the consent procedure, study goal, and that participation would not affect their graduation chances. We provided the participants with a pseudo-anonymous username. After logging in with their usernames, the participants accessed multiple questionnaires including the personality questionnaire, and received as much time as they needed to fill it in. Most participants finished within one hour. The matched sample of Dutch civilians completed their questionnaires online via the HowNutsAreTheDutch platform at their own time and convenience (see Van der Krieke et al., 2016 for details). Both the commandos and civilians could use a digital device of their own choosing.

Measures

The commandos and candidates completed the Dutch version of the NEO-PI-3 (Hoekstra & De Fruyt, 2014) which captures the big five personality domains with 240 items, each rated on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The questionnaire contains 48

items per domain and this is further split into 8 items per facet (6 facets per domain). The NEO-PI-3 was chosen due to its high reliability and validity and its prevalence in military personality research. The validity of the English version has been shown in multiple settings (e.g., Costa et al., 2008, De Fruyt et al., 2009, Egger et al., 2003). Furthermore, the reliability and validity of this instrument are thoroughly assessed by the Dutch Committee on Tests and Testing (COTAN), across different norm groups (including 594 male civilians and 339 civilians between 23 and 35 years of age, see Hoekstra & De Fruyt, 2014 for details). As an additional check of the validity in our sample, we conducted an Exploratory Structural Equation Modeling (ESEM) analysis, which combines the exploratory and confirmatory factor analyses (Marsh et al., 2014). The ESEM model is accepted with a reasonable fit ($p < 0.05$, CFI = 0.89, TLI = 0.83, RMSEA = 0.07) (Marsh et al., 2014, p. 785; see Table S2 in the supplemental material of Huijzer et al., 2022 for more information). The internal reliability of the scale was good, with a McDonald's omega coefficient of 0.87 and a 95% bootstrapped confidence interval ranging from 0.80 to 0.93 as calculated via the psych package (Revelle & Revelle, 2015). The Dutch civilians completed the shortened NEO-FFI (60-items) which was derived from the more comprehensive NEO-PI-3 (Van der Krieke et al., 2016).

Analyses

To examine whether commandos differed in their personality traits from matched civilians (hypothesis 1) and whether graduates differed from dropouts (hypothesis 2), we fitted a multilevel Bayesian model and *t*-tests. Latent profile analyses (LPA) were considered as well, upon request by our reviewer, but appeared less suited given the sample size. The results, which were added to Table S3 of Huijzer et al. (2022a). For LPA, one of the best

information criteria is the Bayesian information criterion (BIC) according to Nylund et al. (2007). In accordance with the results reported, the BIC metric indicated that the 2-profile model (graduates vs. dropouts or commandos vs. controls) is suitable for our data.

With 2 groups and 5 personality domains per research question, we performed Bayesian analyses to avoid the multiple comparison problem, which leads to overestimation of effect sizes or estimating them to be in the wrong direction (Gelman, 2018). We interpreted the posterior model probabilities directly (McElreath, 2020b; Tendeiro & Kiers, 2019). Bayesian techniques also allow us to conclude that there is no effect, which is an additional benefit over classical hypothesis testing (Gelman et al., 2012). We used a multilevel model with partial pooling which is a regularization technique that allows the model to combine information from different groups, and reduces the chances of detecting false-positive results (McElreath, 2020b). In our study, the Bayesian approach estimates the population parameters directly which, in our case, are the population means.

We defined and fitted the models using the Julia programming language (Bezanson et al., 2017) with the Bayesian inference package Turing.jl (Ge et al., 2018). The model is defined as

$$\begin{aligned} \alpha &\sim \text{Normal}(144, 15) \\ \sigma &\sim \text{Cauchy}(0, 2) \\ \alpha_{\text{group}[i]} &\sim \text{Normal}(\alpha, \sigma) \\ \mu_i &= \alpha_{\text{group}[i]} \\ S_i &\sim \text{Normal}(\mu_i, \sigma), \end{aligned}$$

where S_i denotes the personality score for participant i. Here, we set the prior for α to 144, which is in the middle of the lower and upper bound of the scoring range. More specifically, the number is obtained via (240 - 48) / 2 + 48 = 144. This model assumes that all groups should look similar.

Arguably, this common mean α (partial pooling) will favor solutions where differences between groups are minimized. Hence, as a validity check of our Bayesian analysis, we fitted *t*-tests. The benefit of the *t*-tests is that they can be compared to existing literature more easily and are more familiar to many readers. In this study, we considered the Bayesian results as leading and, therefore, used the *t*-tests as a backup. Note that both our Bayesian model and the *t*-test compare the means of different groups. Also note that the Bayesian model is expected to be more conservative due to the partial pooling.

For the *t*-tests, the statistical power is as follows. For hypothesis 1, the most suitable source for estimating the expected effect size compares counterterrorism police officers to civilians. The Cohen's *d* scores on the neuroticism, extraversion, openness, agreeableness, and conscientiousness (NEOAC) dimensions were −0.7, 0.7, 0, 0.2 and 0.4, respectively (Tedeholm et al., 2021). Based on this, we expect an effect size of around 0.5 which gives a power of $d \approx 0.96$ (Champely et al., 2017). For hypothesis 2, we can leverage a related

meta-analysis for an estimate of the effect size: the true validity for neuroticism and extraversion in a meta-analysis on military aviation training success is estimated to be $r = -0.25$ and $r = 0.17$ respectively (Campbell et al., 2010). In terms of Cohen's d, this is $d \approx -0.52$ and $d \approx 0.34$ respectively (Hunter & Schmidt, 2004, Eq. 7.11). Given such a medium Cohen's d effect size of 0.4, the power for the comparison of graduates and dropouts (hypothesis 2) is d ≈ 0.69.

We report Bayesian distribution estimates and credible intervals that show probabilistic uncertainty in the parameter value. This differs from the Frequentist confidence interval and the uncertainty about whether it contains the true value. Also, we provided standardized group differences by means of Cohen's d and interpreted effect sizes as very small to small (below 0.20), small to medium (0.20 to 0.50), medium to large (0.50 to 0.80), or large to very large (0.80 and higher) (Sawilowsky, 2009). As a reference, the average Pearson correlation coefficient between personality and important life outcomes is r = 0.22 (95% CI = [0.18, 0.29], Richard et al., 2003; Soto, 2019) up to r = 0.30 with other (non-test) behaviors (Caspi & Shiner, 2006; Saucier & Goldberg, 1998), thus, small to moderate effects. The code to reproduce the results has been made available at the Open Science Framework and can be accessed at https://osf.io/ysfu6.

2.3 Results

Since Bayesian samplers start at a random point, the results can vary when doing multiple runs, that is, run multiple chains. Following common practice (McElreath, 2020b), three chains were run and their results were consistent. We also checked stationarity and good mixing by visually inspecting graphs of the posterior samples. In Figure 2.2 and 2.3, the posterior distributions show

the aggregated results from the chains. The results for the *t*-tests are described in the text below; together with the results for the first and second hypotheses. The descriptives are shown in Table 2.2.

Table 2.2

Descriptive Statistics for Commandos, Graduates, Dropouts and Civilians

	Commandos	**Civilians**	**Graduates**	**Dropouts**
Number of participants	110	275	53	138
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Neuroticism	111.9 (16.7)	130.9 (37.2)	110.3 (15.5)	114.6 (15.4)
Extraversion	161.6 (12.8)	157.4 (33.1)	164.3 (13.2)	161.9 (14.9)
Openness	148.2 (14.9)	174.1 (25.2)	148.9 (13.2)	149.2 (13.9)
Agreeableness	164.2 (13.4)	160.1 (24.1)	172.5 (13.9)	171.4 (14.4)
Conscientiousness	178.3 (15.6)	166.4 (29.3)	183.9 (14.5)	180.5 (13.6)

Note. SD = Standard Deviation. Civilians refers to a male sample from the general Dutch population matched to the commandos on age and education.

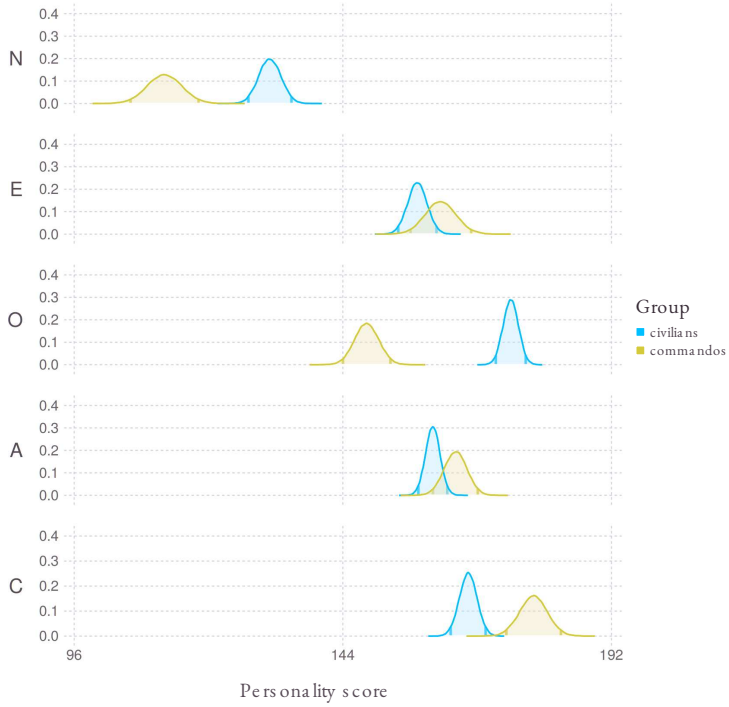
Hypothesis 1 - Commandos versus Controls

First, we examined whether commandos differed in their Big Five personality traits from civilians. We fitted Bayesian models (Figure 2.2) and performed t-tests (described in the text). In line with hypothesis 1, these models demonstrate that commandos score lower than civilians on neuroticism ($t_{(383)} = -5.15, p < 0.001, d = -0.58$) with a medium to large effect size and higher on conscientiousness ($t_{(383)} = 4.01, p < 0.001, d = 0.45$) with a small to medium effect size. Commandos also score lower on openness than civilians ($t_{(383)} = -10.1, p < 0.001, d = -1.13$) with a large to very large effect size. There were no clear differences between the groups for extraversion ($t_{(383)} = 1.30, p = 0.19$,

$d = 0.15$) and agreeableness ($t_{(383)} = 1.69$, $p = 0.09$, $d = 0.19$) with both a very small to small effect size.

Figure 2.2

Comparison of Civilians with Commandos on the Big Five Personality Domains



Note. Neuroticism (N), extraversion (E), openness (O), agreeableness (A) and conscientiousness (C). The small vertical bars in the posterior distributions show the 95% credible interval.

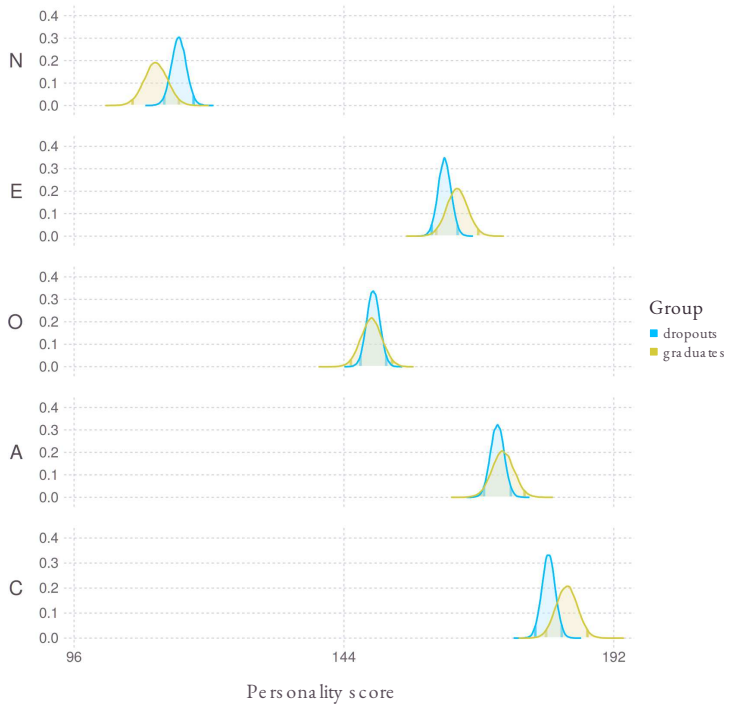
Hypothesis 2 - Graduates versus Dropouts

To examine whether commando graduates differed in their Big Five personality traits from dropouts, we again fitted a Bayesian model (Figure 2.3) and performed *t*-tests (described in the text). In contrast with hypothesis 2, none of the results were significant. Yet, the clearest effect size differences are visible for neuroticism and conscientiousness, where graduates score lower than dropouts on neuroticism ($t_{(189)} = -1.71$, $p = 0.09$, $d = -0.27$) with a small

to medium effect size. For conscientiousness, graduates score higher ($t_{(189)}$ = 1.51, p = 0.13, d = 0.24) with a small to medium effect size. Smaller effect sizes were visible for the other domains, namely openness ($t_{(189)}$ = −0.14, p = 0.89, d = 0.02) with a very small to small effect size, extraversion ($t_{(189)}$ = 1.04, p = 0.30, d = 0.17) with a very small to small effect size and agreeableness ($t_{(189)}$ = 0.49, p = 0.63, d = 0.08) with a very small to small effect size.

Figure 2.3

Comparison of Graduates with Dropouts on the Big Five Personality Domains



Note. Neuroticism (N), extraversion (E), openness (O), agreeableness (A) and conscientiousness (C). The small vertical bars in the posterior distributions show the 95% credible interval.

To derive a more nuanced insight into commando personalities we subsequently examined differences between commandos and matched controls in 30 more specific facet traits, generally thought to be informative when

predicting consequential outcomes (Stewart et al., 2022). We refrain from an interpretation of the facet differences between commandos and civilians because none was significant in our models (all d below 0.30 and p above 0.07), see Table S1 in the supplemental material of Huijzer et al. (2022a) for details. Finally, we explored whether graduates and dropouts differed in more specific facet traits, no significant differences were detected (see Table S1 of Huijzer et al., 2022 for details).

2.4 Discussion

This study was aimed to investigate (1) personality differences between experienced commandos and civilian controls and (2) whether and how personality traits distinguished graduates from dropouts during the selection period. To investigate the hypotheses, a large-scale study was conducted in collaboration with the Royal Netherlands Army. Our key observation was, first, that the group of commandos was less neurotic, more conscientious, and markedly less open to experience than civilians matched on age and education. Second, successful candidates tend to report lower neuroticism and higher conscientiousness. The other personality traits showed inconsistent results, and more nuanced facet traits did not differ between graduates and dropouts.

Hypothesis 1 - Commandos versus Controls

In line with our first hypothesis, the commandos scored lower on neuroticism and higher on conscientiousness compared to matched civilian controls. This pattern is in accordance with studies of more experienced U.S. Navy SEALs (Braun et al., 1994) and Swedish counterterrorism intervention police officers versus Swedish civilians (Tedeholm et al., 2021). For extraversion, we found no evidence to support, nor to reject, the idea that operators are more

extraverted than civilians. Although the direction of the effect that we found is in accordance with previous research, Braun et al. (1994) and Tedeholm et al. (2021) found clearer evidence that U.S. Navy SEALs score higher on extraversion than less experienced SEALs, and that counterterrorism intervention police officers score higher on extraversion than civilians, respectively. For agreeableness, we had no specific expectations, and also found no meaningful differences between commandos and controls.

Our analysis provided strong evidence for a marked difference in openness to experience between commandos and matched controls, a novel contribution to the literature on personnel selection and military psychology. This result suggests that, compared to civilians, commandos prefer routines, consistency, traditions, and familiarity, and approach new things with great caution and are less likely to be overwhelmed by emotions (Larsen et al., 2020). Openness also differed between French military candidates and general students (Rolland et al., 1998), and between German students who decided to join the military or not (Jackson et al., 2012). Contrarily, a comparison of counter-terrorism intervention unit police officers and civilians showed trivial differences in openness (Tedeholm et al., 2021). Compared to previous research, it seems that the civilians in our sample scored higher on openness than the control groups and the commandos score lower than the military groups (to see this, compare Figure 3 and Figure 4). This may be due to the nature of our matched control group, which comprised relatively young men who voluntarily participated in an online questionnaire (Marcus & Schütz, 2005). Finally, our results are partly in line with the study of multiple military datasets by Van der Linden et al. (2010) who concluded that successful military candidates in general were more likely to score low on neuroticism, and high on extraversion, conscientiousness, agreeableness, and openness.

Hypothesis 2 - Graduates versus Dropouts

For the comparison between graduates and dropouts, the results were less evident. This is likely to be caused by the homogeneity of the group in combination with the limited statistical power. Interestingly, as with the comparison between commandos and controls, the clearest patterns were found in neuroticism and conscientiousness. For neuroticism, our results suggest that graduates score lower on neuroticism than dropouts, which in the hypothesized direction. This result is also in line with the study by McDonald et al. (1990) on U.S. Navy SEAL candidates, which showed that graduates were less neurotic than those who did not graduate. Similarly, in a study on Canadian Forces basic training, it was found that lower neuroticism was associated with training success (Lee et al., 2011). Furthermore, a meta-analysis concluded that lower neuroticism predicted military aviation training success (Campbell et al., 2010). People with lower neuroticism scores tend to experience lower subjective threat, impulsivity, vulnerability to stress, and anxiety, which may be important characteristics to become a commando.

For conscientiousness, the result was in the hypothesized direction, but was not significant. A stronger pattern was found in a study on Navy SEALs who found that more experienced SEALs score higher on conscientiousness (Braun et al., 1994). We also found that graduates scored on average half a standard deviation higher on extraversion than dropouts. A clearer difference has previously been reported in a meta-analysis on military aviators (Campbell et al., 2010), a study with Navy SEALs (Hartmann et al., 2003) and a study with reconnaissance marines (Saxon et al., 2020). A likely explanation for these results is that extraverted people are more prone to seek excitement, be active, and take risks, all of which are important qualities for commandos (Keinan et al., 1984; Stewart, 2017).

Contrary to our hypothesis and previous research we did not find that graduates score higher on agreeableness (Campbell et al., 2010; Hartmann et al., 2003; Saxon et al., 2020). A possible explanation for the difference between previous findings and our outcomes is the lower power of our study or that the trait agreeableness contains facets that can be positive as well as negative for a commando. For example, having high trust and straightforwardness is important for effective teamwork (Jones & George, 1998), but being modest might not contribute to a successful mission. This observation is in line with studies of leadership that indicate that leaders tend to be extraverted and low on neuroticism, but results for agreeableness tend to be fuzzy, which suggests that a broader range of scores can be proficient strategies (Do & Minbashian, 2020; Judge et al., 2002). Finally, we did not have a hypothesis for openness to experience, and our results did not reveal a strong enough difference between the graduates and dropouts to conclude that they differ in this trait.

Limitations and Future Directions

In our study, we used the NEO-PI-3 with 240 items for the candidates and commandos, and the NEO-FFI for the civilians. This difference appeared to result in different variances in scores on personality dimensions. Indeed, upon further investigation, and comparison with other personality research, we found that the difference in variance is likely caused by the difference in length in questionnaires, and not by the group under study. In hindsight, this difference made sense because more questions imply that it is more likely that the mean score of a participant averages out, that is, that the score is less extreme. However, we do not expect that this has notably affected the conclusions. For future directions, more research is needed to investigate

individual facets. Since this increases the number of comparisons one likes to make, Bayesian analyses provide an intuitive way to handle this (Gelman et al., 2012). Also, more research is needed to investigate personality profiles instead of personality traits. Mixed models such as latent profile analysis provide an interesting avenue in this regard (Oberski, 2016; Wanders et al., 2016, see also Table S3 of Huijzer et al., 2022), assuming that model requirements such as statistical power can be met. Moreover, other factors than personality may also be important to become a commando (see introduction). Therefore, an important avenue is to discover not only the psychological but also the physical predictors of successful graduation in the special forces selection period (e.g., Saxon et al., 2020).

2.5 Conclusion

In this study, male commandos differ from a group of age-matched civilians by being less neurotic, less open to new experiences, and more conscientious. People who started the commando training showed similar differences, namely, that graduates score lower on neuroticism and higher on conscientiousness than dropouts. Our finding that the directions are the same for both comparisons adds certainty to the effects that we have found. Given the relatively small differences between the graduates and dropouts, we would argue that a personality test would not provide a strong selection instrument by itself. This is likely due to the fact that the group of people who decide to join the commandos is quite homogeneous. Hence, for selection purposes, examining additional psychological and physical measures is an important avenue. For recruitment purposes though, the use of personality tests can provide important clues as our study showed relatively strong differences between commandos and civilians.

CHAPTER 3

SIRUS.jl: Interpretable Machine Learning via Rule Extraction

This chapter is based on:

Huijzer, R., Blaauw, F. J., Den Hartigh, R. J. R. (2023). SIRUS.jl: Interpretable Machine Learning via Rule Extraction. *Journal of Open Source Software*, 8(90), 5786. https://doi.org/10.21105/joss.05786

Abstract

SIRUS.jl[4] is an implementation of the original Stable and Interpretable RUle Sets (SIRUS) algorithm in the Julia programming language (Bezanson et al., 2017). The SIRUS algorithm is a fully interpretable version of random forests, that is, it reduces thousands of trees in the forest to a much lower number of interpretable rules (e.g., 10 or 20). With our Julia implementation, we aimed to reproduce the original C++ and R implementation in a high-level language to verify the algorithm as well as making the code easier to read. We show that the model performs well on classification tasks while retaining interpretability and stability. Furthermore, we made the code available under the permissive MIT license. In turn, this allows others to research the algorithm further or easily port it to production systems.

[4]Source code available at https://github.com/rikhuijzer/SIRUS.jl.

3.1 Statement of need

Many of the modern day machine learning models are noninterpretable models, also known as *black box* models. Well-known examples of noninterpretable models are random forests (Breiman, 2001) and neural networks. Such models are available in the Julia programming language via, for example, LightGBM.jl (Ke et al., 2017), Flux.jl (Innes, 2018), and BetaML.jl (Lobianco, 2021). Although these models can obtain high predictive performance and are commonly used, they can be problematic in high stakes domains where model decisions have real-world impact on individuals, such as suggesting treatments or selecting personnel. The reason is that noninterpretable models may lead to unsafe, unfair, or unreliable predictions (Barredo Arrieta et al., 2020; Doshi-Velez & Kim, 2017). Furthermore, interpretable models may allow researchers to learn more from the model, which in turn may allow researchers to make better model decisions and achieve a higher predictive performance.

However, the set of interpretable models is often limited to ordinary and generalized regression models, decision trees, RuleFit, naive Bayes classification, and k-nearest neighbors (Molnar, 2022). For these models, however, predictive performance can be poor for certain tasks. Linear models, for instance, may perform poorly when features are correlated and can be sensitive to the choice of hyperparameters. For decision trees, predictive performance is poor compared to random forests (James et al., 2013). RuleFit is not available in Julia and is *unstable* (Bénard et al., 2021a), meaning sensitive to small changes in data. Naive Bayes, available in Julia as NaiveBayes.jl[5], is

[5]Source code available at https://github.com/dfdx/NaiveBayes.jl.

often overlooked and can be a suitable solution, but only if the features are independent (Ashari et al., 2013).

Researchers have attempted to make the random forest models more interpretable. Model interpretation techniques, such as SHAP (Lundberg & Lee, 2017) or Shapley, available via Shapley.jl[6], have been used to visualize the fitted model. However, the disadvantage of these techniques are that they convert the complex model to a simplified representation. This causes the simplified representation to be different from the complex model and may therefore hide biases and issues related to safety and reliability (Barredo Arrieta et al., 2020).

The SIRUS algorithm solves this by simplifying the complex model and by then using the simplified model for predictions. This ensures that the same model is used for interpretation and prediction. However, the original SIRUS algorithm was implemented in about 10k lines of C++ and 2k lines of R code[7] which makes it hard to inspect and extend due to the combination of two languages. Our implementation is written in about 2k lines of pure Julia code. This allows researchers to more easily verify the algorithm and investigate further improvements. Furthermore, the original algorithm was covered by the GPL-3 copyleft license meaning that copies are required to be made freely available. A more permissive license makes it easier to port the code to other languages or production systems.

3.2 Interpretability

To show that the algorithm is fully interpretable, we fit an example on the Haberman's Survival Dataset (Haberman, 1999). The dataset contains

[6]Source code available at https://gitlab.com/ExpandingMan/Shapley.jl.

[7]Source code available at https://gitlab.com/drti/sirus.

survival data on patients who had undergone surgery for breast cancer and contains three features, namely the number of axillary *nodes* that were detected, the *age* of the patient at the time of the operation, and the patient's *year* of operation. For this example, we have set the hyperparameters for the maximum number of rules to 8 since this is a reasonable trade-off between predictive performance and interpretability. Generally, a higher maximum number of rules will yield a higher predictive performance. We have also set the maximum depth hyperparameter to 2. This hyperparameter means that the random forests inside the algorithm are not allowed to have a depth higher than 2. In turn, this means that rules contain at most 2 clauses (`if A & B`). When the maximum depth is set to 1, then the rules contain at most 1 clause (`if A`). Most rule-based models, including SIRUS, are restricted to depth of 1 or 2 (Bénard et al., 2021a).

The output for the fitted model looks as follows (see Section 3.5 for the code):

```
StableRules model with 8 rules:
 if X[i, :nodes] < 7.0 then 0.238 else 0.046 +
 if X[i, :nodes] < 2.0 then 0.183 else 0.055 +
 if X[i, :age] ≥ 62.0 & X[i, :year] < 1959.0 then 0.0 else 0.001 +
 if X[i, :year] < 1959.0 & X[i, :nodes] ≥ 2.0 then 0.0 else 0.006 +
 if X[i, :nodes] ≥ 7.0 & X[i, :age] ≥ 62.0 then 0.0 else 0.008 +
 if X[i, :year] < 1959.0 & X[i, :nodes] ≥ 7.0 then 0.0 else 0.003 +
 if X[i, :year] ≥ 1966.0 & X[i, :age] < 42.0 then 0.0 else 0.008 +
 if X[i, :nodes] ≥ 7.0 & X[i, :age] ≥ 42.0 then 0.014 else 0.045
and 2 classes: [0, 1].
```

This shows that the model contains 8 rules where the first rule, for example, can be interpreted as:
If the number of detected axillary nodes is lower than 7, then take 0.238, otherwise take 0.046.

This calculation is done for all 8 rules and the score is summed to get a prediction. In essence, the first rule says that if there are less than 8 axillary

nodes detected, then the patient is more likely to survive (`class == 1`). Put differently, the model states that if there are many axillary nodes detected, then it is, unfortunately, less likely that the patient will survive. This model is fully interpretable because the model contains a few dozen rules which can all be interpreted in isolation and together.

3.3 Stability

Another problem that the SIRUS algorithm addresses is that of model stability. A stable model is defined as a model which leads to similar conclusions for small changes to data (Yu, 2020). Unstable models can be difficult to apply in practice as they might require processes to constantly change. This also makes such models appear less trustworthy. Put differently, an unstable model by definition leads to different conclusions for small changes to the data and, hence, small changes to the data could cause a sudden drop in predictive performance. One model which suffers from a low stability is a decision tree, available via DecisionTree.jl (Sadeghi et al., 2022), because it will first create the root node of the tree, so a small change in the data can cause the root, and therefore the rest, of the tree to be completely different (Molnar, 2022). Similarly, linear models can be highly sensitive to correlated data and, in the case of regularized linear models, the choice of hyperparameters. The aforementioned RuleFit algorithm also suffers from stability issues due to the unstable combination of tree fitting and rule extraction (Bénard et al., 2021a). The SIRUS algorithm solves this problem by stabilizing the trees inside the forest, and the original authors have proven the correctness of this stabilization mathematically (Bénard et al., 2021a). In the rest of this paper, we will compare the predictive performance of SIRUS.jl to the performance of decision trees (Sadeghi et al., 2022), linear models, XGBoost (Chen &

Guestrin, 2016), and the original (C++/R) SIRUS implementation (Bénard et al., 2021a). The interpretability and stability are summarized in Table 3.1.

Table 3.1

Summary of Interpretability and Stability for Various Models

	Decision Tree	Linear Model	XGBoost	SIRUS
Interpretability	High	High	Medium	High
Stability	Low	Medium	High	High

3.4 Predictive Performance

The SIRUS model is based on random forests and therefore well suited for settings where the number of variables is comparatively large to the number of datapoints (Biau & Scornet, 2016). To make the random forests interpretable, the large number of trees are converted to a small number of rules. The conversion works by converting each tree to a set of rules and then pruning the rules by removing simple duplicates and linearly dependent duplicates, see the SIRUS.jl documentation or the original paper (Bénard et al., 2021b) for details. In practice, this trade-off between between model complexity and interpretability comes at a small performance cost.

To show the performance, we compared SIRUS to a decision tree, linear model, XGBoost, and the original (C++/R) SIRUS algorithm; similar to Table 3.1. We have used Julia version 1.9.3 with SIRUS version 1.3.3 (at commit `5c87eda`), 10-fold cross-validation, and we will present variability as 1.96 $*$ standard error for all evaluations with respectively the following datasets, outcome variable type, and measures: Haberman's Survival Dataset (Haberman, 1999) binary classification dataset with AUC, Titanic (Eaton & Haas, 1995) binary classification dataset with Area Under the Curve (AUC), Breast

Cancer Wisconsin (Wolberg et al., 1995) binary classification dataset with AUC, Pima Indians Diabetes (Smith et al., 1988) binary classification dataset with AUC, Iris (Fisher, 1936) multiclass classification dataset with accuracy, and Boston Housing (Harrison & Rubinfeld, 1978) regression dataset with R^2; see Table 3.2. For full details, see `test/mlj.jl`[8]. The performance scores were taken from the SIRUS.jl test job that ran following commit `5c873da` using GitHub Actions. The result for the Iris dataset for the original SIRUS algorithm is missing because the original algorithm has not implemented multiclass classification.

Table 3.2

Predictive Performance Estimates

Dataset	Decision Tree	Linear Model	XGBoost max depth: ∞	XGBoost max depth: 2	Original SIRUS max depth: 2 max rules: 10	SIRUS.jl max depth: 2 max rules: 10
Haberman	0.54 ± 0.06	0.69 ± 0.06	0.65 ± 0.04	0.63 ± 0.04	0.66 ± 0.05	0.67 ± 0.06
Titanic	0.76 ± 0.05	0.84 ± 0.02	0.86 ± 0.03	0.87 ± 0.03	0.81 ± 0.02	0.83 ± 0.02
Cancer	0.92 ± 0.03	0.98 ± 0.01	0.99 ± 0.00	0.99 ± 0.00	0.96 ± 0.02	0.98 ± 0.01
Diabetes	0.67 ± 0.05	0.70 ± 0.06	0.80 ± 0.04	0.82 ± 0.03	0.80 ± 0.02	0.75 ± 0.05
Iris	0.95 ± 0.03	0.97 ± 0.03	0.94 ± 0.04	0.93 ± 0.04		0.77 ± 0.08
Boston	0.74 ± 0.11	0.70 ± 0.05	0.87 ± 0.05	0.86 ± 0.05	0.63 ± 0.07	0.61 ± 0.09

At the time of writing, SIRUS's predictive performance is comparable to the linear model and XGBoost on the binary classification datasets, that is, Haberman, Titanic, Breast Cancer, and Diabetes. The best performance occurs at the Diabetes dataset where both XGBoost and the SIRUS models

[8] https://github.com/rikhuijzer/SIRUS.jl/blob/5c87eda4d0c50e0b78d12d6bd2c4387f5a83f518/test/mlj.jl.

outperform the linear model. The reason for this could be that negative effects are often nonlinear for fragile systems (Taleb, 2020). For example, it could be that an increase in oral glucose tolerance increases the chance of diabetes exponentially. In such cases, the hard cutoff points chosen by tree-based models, such as XGBoost and SIRUS, may fit the data better.

For the multiclass Iris classification and the Boston Housing regression datasets, the performance was worse than the other non-SIRUS models. It could be that this is caused by a bug in the implementation or because this is a fundamental issue in the algorithm. Further work is needed to find the root cause or workarounds for these low scores. One possible solution would be to add SymbolicRegression.jl (Cranmer, 2023) as a secondary back end for regression tasks. Similar to SIRUS.jl, SymbolicRegression.jl can fit expressions of a pre-defined form to data albeit with more free parameters, which might fit better but also might cause overfitting, depending on the data. This achieves performance that is similar to XGBoost (Hanson, 2023).

In conclusion, interpretability and stability are often required in high-stakes decision making contexts such as personnel or treatment selection. In such contexts and when the task is classification, SIRUS.jl obtains a reasonable predictive performance, while retaining model stability and interpretability.

3.5 Code Example

The model can be used via the Machine Learning Julia (MLJ) (Blaom et al., 2020) interface. The following code, for example, was used to obtain the fitted model for the Haberman example at the start of this paper, and is also available in the SIRUS.jl docs[9].

[9] https://sirus.jl.huijzer.xyz/dev/basic-example/.

We first load the dependencies:

```
using CategoricalArrays: categorical
using CSV: CSV
using DataDeps: DataDeps, DataDep, @datadep_str
using DataFrames
using MLJ
using StableRNGs: StableRNG
using SIRUS: StableRulesClassifier
```

And specify the Haberman dataset via DataDeps.jl, which allows data verification via the checksum and enables caching:

```
function register_haberman()
  name = "Haberman"
  message = "Haberman's Survival Data Set"
  remote_path = "https://github.com/rikhuijzer/haberman-survival-dataset/
      releases/download/v1.0.0/haberman.csv"
  checksum =
      "a7e9aeb249e11ac17c2b8ea4fdafd5c9392219d27cb819ffaeb8a869eb727a0f"
  DataDeps.register(DataDep(name, message, remote_path, checksum))
end
```

Next, we load the data into a `DataFrame`:

```
function load_haberman()::DataFrame
  register_haberman()
  path = joinpath(datadep"Haberman", "haberman.csv")
  df = CSV.read(path, DataFrame)
  df[!, :survival] = categorical(df.survival)
  return df
end
```

We split the data into features (`X`) and outcomes (`y`):

```
data = load_haberman()
X = select(data, Not(:survival))
y = data.survival
```

We define the model that we want to use with some reasonable hyperparameters for this small dataset:

```
model = StableRulesClassifier(; rng=StableRNG(1), q=4, max_depth=2,
max_rules=8)
```

Finally, we fit the model to the data via MLJ and show the fitted model:

```
mach = let
  mach = machine(model, X, y)
  MLJ.fit!(mach)
end

mach.fitresult
```

Resulting in the fitresult that was presented in Section 3.2.

CHAPTER 4

Predicting Special Forces Dropout via Explainable Machine Learning

This chapter is based on:

Huijzer, R., De Jonge, P., Blaauw, F. J., Baatenburg de Jong, M., De Wit, A., & Den Hartigh, R. J. R. (2024). Predicting Special Forces Dropout via Explainable Machine Learning. *European Journal of Sport Science*, 24(11), 1564-1572. https://doi.org/10.1002/ejsc.12162

Abstract

Selecting the right individuals for a sports team, organization, or military unit has a large influence on the achievements of the organization. However, the approaches commonly used for selection are either not reporting predictive performance or not explainable (i.e., black box models). In the present study, we introduce a novel approach to selection research, using various machine learning models. We examined 274 special forces recruits, of whom 196 dropped out, who performed a set of physical and psychological tests. On this data, we compared four machine learning models on their predictive performance, explainability, and stability. We found that a stable rule-based (SIRUS) model was most suitable for classifying dropouts from the special forces selection program. With an averaged area under the curve score of 0.70, this model had good predictive performance, while remaining explainable and stable. Furthermore, we found that both physical and psychological variables were related to dropout. More specifically, a higher score on the 2800 meters time, connectedness, and skin folds were most strongly associated with dropping out. We discuss how researchers and practitioners can benefit from these insights in sport and performance contexts.

4.1 Introduction

The achievements of sports clubs, organizations, and military units are largely determined by the performance of the individuals in the organization. As a consequence, there is an ever increasing pressure to select the right individuals, that is, individuals who will perform successfully in the future (e.g., Den Hartigh et al., 2018). Historically, military selection has been an important breeding ground for research into selection in psychology and sports. For example, widely used instruments such as intelligence tests (Terman, 1918), personality inventories (Ellis & Conrad, 1948), and leadership measures (Fleishman, 1953) were first established and validated in military contexts. In the present study, we aimed to advance the field of selection further by applying machine learning models for the selection of elite soldiers. In doing so, we set out to investigate the predictive performance, explainability, and stability of statistical models based on relevant physical and psychological predictors. Here, predictive performance means the estimated ability of the model to predict future behaviors, explainability means how easy it is to understand the model and why certain predictions were made, and stability means the ability of the model to produce similar conclusions for small changes to the data (Yu, 2013).

Selection in High-Stakes Military Contexts

Within the military, the special forces are considered elite. Special forces operators need to be able to perform their tasks under difficult circumstances, such as continuous threat, extreme temperatures, isolation, and high task complexity, while being involved in politically sensitive situations (Picano et al., 2002). Similar to elite sports, this requires extraordinary physical and mental capabilities (Vaara et al., 2022). Special forces selection courses world-

wide simulate these circumstances in, what some countries call, *hell weeks*. During these selection weeks, recruits typically complete exercises and tasks for a large part of the day while being sleep deprived. Several studies have been conducted in the past decades to predict success versus dropout in such selection programs of the special forces. For example, a study among 800 candidates found that both physical and psychological measures, such as grit and pull-ups, significantly correlated with graduation (Farina et al., 2019). The relevance of physical and psychological factors were also found in other high-stakes military contexts. For instance, studies on 12,924 military pilots, 115 reconnaissance marines, and 57 counter terrorism intervention unit recruits found that various physical and psychological measures were associated with graduation (King et al., 2013; Saxon et al., 2020; Tedeholm et al., 2021). Furthermore, a large-scale study on 1,138 United States (U.S.) special forces candidates found that psychological hardiness significantly correlated with graduation (Bartone et al., 2008). Taken together, a multidisciplinary approach including both physical and psychological measures, is likely to perform best on the complex task of predicting dropout (Williams & Reilly, 2000).

An important note about previous research is that many studies report only model explanations, that is, the studies fit a statistical model to the data and report the fitted parameters. Interestingly, this approach is also common practice in the field of sport science. However, the outcomes produced by such models may have little ability to predict future behaviors, because of overfitting (Hofman et al., 2021; Jauhiainen et al., 2022; Yarkoni & Westfall, 2017). Also, many studies only report the results from one statistical model, such as a simple regression or the *t*-test, which largely ignores the statistical (and computational) progress made since then. Applying more

recent analytic techniques, such as model evaluation via cross-validation, could therefore improve research into the selection procedures (e.g., Abt et al., 2022).

Statistical Models from Machine Learning

Recent analytic advances can be found in the domain of machine learning, which can generally be described as computer systems that learn and adapt without following specific instructions. One example is computer vision, which contains models that can learn from visual data to automatically detect and classify sport-specific movements. In general, the field invented and re-discovered a plethora of statistical models, many of which are promising because the models are distribution-free and are able to find complex relationships in data. The distribution-free property is relevant for selection because psychometric variables are usually normally distributed while performance variables in elite performers often are not (e.g., Den Hartigh et al., 2018; O'Boyle Jr & Aguinis, 2012). Furthermore, finding complex relationships could provide new insights into the underlying processes when sufficient data is available. As an example, Jauhiainen et al. (2022) used a complex data set, containing 3-dimensional motion and physical data, to predict injuries in 791 female elite handball and soccer players. More generally, the commonly applied random forest algorithms have been very performant in different settings; especially when the number of variables is large or larger than the number of observations (Biau & Scornet, 2016).

However, machine learning is no panacea. A disadvantage of many machine learning applications in sports and the selection of military personnel is that the models are too complex to understand. Often, the complex models are then converted to a simplified form to make them interpretable,

for example by using SHAP (SHapley Additive exPlanations; for details see Molnar, 2022). Although the purpose of SHAP is to increase transparency and explainability of machine learning models, it loses information during the conversion from the complex model to the simplified representation. In other words, the simplified representation is not the same as the model that will be used for decision making. This is problematic for researchers and practitioners because the simplification could hide issues related to safety, fairness (e.g., biases), and reliability (Barredo Arrieta et al., 2020; Doshi-Velez & Kim, 2017). This is especially important in the context of selection, where wrong decisions can have a lasting impact on the individual.

Apart from predictive performance and explainability, the stability of models is also an important aspect. A stable model is defined as a model which leads to similar conclusions for small changes to data (Yu, 2013). An example of an unstable model could be a model which selects personality and sprint times to predict dropout in this year's cohort, but selects other variables for next year's cohort. In the context of selection, this variation in the prediction model is problematic. Unstable models cause various operational problems such as being deemed less trustworthy and requiring constant changes to the selection procedure (Yu, 2013).

Current Study

The purpose of the current study was to determine how well we could predict dropout of special forces recruits while retaining model explainability and stability. We used a regularized linear model as a baseline. This model is close to the linear models that are typically used for decision making in sport and psychology research. Next, we used three machine learning models, namely a decision tree, a state-of-the-art random forest, and a state-of-the-art explain-

able rule-based model. We specifically investigated how the four models compared on their predictive performance, explainability, and stability. We compared the models on their predictive performance via average area under the curve (AUC), on their explainability by comparing model interpretation techniques (e.g., linear model coefficients versus SHAP), and stability by comparing the differences between the algorithms used.

4.2 Materials and Methods

Participants

We recruited 311 participants aged between 20 and 39 (M_{age} = 26.5, SD_{age} = 3.8), who were exclusively Dutch males and all part of the selection of the Special Forces of the Royal Netherlands Army. Active consent was obtained from all participants and the procedure was approved by the ethical review board of the faculty (code: PSY-1920-S-0512). Data preprocessing, which included the removal of participants for which some data was missing, resulted in a dataset of 274 participants. Of these participants, 196 dropped out and 78 graduated. More information could not be provided due to security reasons.

Design & Procedure

Participation occurred via a platform specifically built for the research project (https://yourspecialforces.nl). The data collection was organized by researchers of the university at the training camp, and was facilitated by the staff of the Special Forces unit. Physical assessments occurred on the first day of the first week. Also in the first week of the training, participants completed the psychological assessments using tablets in a large room which was set up like a traditional classroom. Once participants entered the room for the psychological assessment, they were informed about the consent procedure,

study goal, and that participation would not affect their graduation chances. For three to four days, the participants spent roughly one hour per day on filling out the questionnaires, which were all in Dutch.

Measures

The study contained both physical and psychological measures. The physical fitness of the recruits was measured using a test battery designed to assess relevant physiological and physical characteristics that are considered to be important in military training courses (e.g., Haff & Triplett, 2015). All tests were taken in a predetermined order. First, body composition was determined by measuring length, weight, and the 4-Site Skinfold (Durnin & Womersley, 1974). Then a standardized warming up was conducted after which the recruits started in the test-circuit. Lower body power was measured with a broad jump, the best of three attempts was noted in centimeters. Next, speed and agility were tested using the Pro Agility test conducted twice with 30 seconds rest in between and both sprint times were summed. The agility test was followed by maximal grip strength of both hands with one attempt per hand using a Grip dynamometer. After this test, maximal strength of the lower body push and pull, and upper body push-kinetic chain was measured with a 3 repetition max (RM) protocol using the hex-bar deadlift and bench press exercise. Strength endurance of the upper body pull-chain was measured with pull-ups: recruits had one minute to complete as many pull-ups as possible. The penultimate test was designed to determine the anaerobic capacity of the recruits using a 60 meter sprint. It measured the time it took to sprint from one place to a place 5 meters away and back (10 meters), then 10 meters away and back (20 meters), and finally to a place 15 meters away and back (30 meters). Also here, the test was conducted twice with 30 seconds in between.

After the 60 meter sprint, the recruits had exactly 10 minutes to recover and prepare for the aerobic endurance test, a timed 2800 meter run. The recruits were instructed to complete 8 rounds on a 350 meter concrete track as fast as possible.

Regarding the psychological measures, the first day included the informed consent and a resilience questionnaire. The resilience questionnaire assessed the ability to recover or bounce back from stress via the Brief Resilience Scale (Smith et al., 2008). For example, one of the six items was "I tend to bounce back quickly after hard times". Next, goal commitment was measured via six items such as "I am strongly committed to pursuing my goals" (see Van Yperen, 2009). The next questionnaire measured self-efficacy (Bandura, 2006) with 14 items such as "How confident are you in your ability to remain calm in difficult situations?".

The second day consisted of two cognitive ability tests (Condon & Revelle, 2014). The first test contained 11 matrix reasoning items and the second test contained 24 three-dimensional rotation items. The participants were allowed to take 15 and 30 minutes respectively to finish both tests. On the third day, three questionnaires were answered. The first questionnaire was a combination of five short questionnaires, namely Mindsets (Dweck, 2000), Basic Motives (Van Yperen et al., 2014), Motivation Type (Pelletier et al., 2013), and Approach-Avoidance Temperament (Elliot & Thrash, 2010). The second measured mental toughness via the MTQ48 (Clough et al., 2002). This questionnaire contains four key components, namely Control, Commitment, Challenge, and Confidence. The third questionnaire measured Coping (Lazarus & Folkman, 1984). This questionnaire measured emotion-focused versus problem-focused coping in response to stressful events. For example, "I try to forget the whole thing by focusing on other things" which

is an example of an emotion-focused strategy. After this, the participants filled in the Dutch version of the NEO-PI-3 personality questionnaire, which measures the big five dimensions: Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness (McCrae et al., 2005).

4.3 Analyses

In order to find the best performing model, we compared four different models via MLJ.jl (Blaom et al., 2020). We calculated the models' scores on the Area Under the receiver operating characteristics Curve (AUC). The AUC is a metric that indicates how well a model predicts a binary outcome, dropout versus graduation in our case. The AUC takes into account that the threshold of the model can be chosen freely. An AUC score of 1 means that the model can perfectly predict all outcomes and a score of 0 means that the model predicts everything wrong. An AUC score of 0.5 means random guessing and AUC scores of 0.7 to 0.85 and higher are generally considered to be good to excellent in social sciences (e.g., Menaspa et al., 2010). We compared all models on their predictive performance via 12-fold cross-validation with AUC as the metric.

The first model was the baseline: a regularized linear model. Here, regularization was necessary because this study gathered relatively many variables compared to the number of observations. Without regularization, the model is likely to overfit in such situations. As regularization for the linear model, we choose Elastic Net which is a combination of Lasso and Ridge regression (e.g., Zou & Hastie, 2005) and fitted the model via MLJLinearModels.jl (Blaom et al., 2020). The strength of both regularizers was chosen automatically via hyperparameter tuning and 12 fold cross-validation. The second model was a decision tree, fitted via DecisionTree.jl (Sadeghi et al., 2022), and the third was

a state-of-the-art boosted random forest called XGBoost (Chen & Guestrin, 2016). The fourth model was a state-of-the-art Stable and Interpretable Rule Sets (SIRUS) algorithm (Bénard et al., 2021b; Huijzer et al., 2023b). The SIRUS model is essentially also a random forest algorithm, but with a small modification such that it is more stable and, therefore, explainable. Note that contrary to more continuous models such as linear models, the rules fitted by SIRUS contain hard cutpoints (e.g., *if some variable < 20, then A else B*).

Of these models, the XGBoost is the least explainable while the other three models are all explainable. That is, the XGBoost cannot easily be interpreted due the complexity of the model. For the decision tree model, despite being explainable, it has the drawback of having a low stability since the split point at the root of the tree tends to vary wildly (for details about this phenomenon, see Molnar, 2022). The stability of the logistic regression is moderate since the model is highly sensitive to the choice of regularization parameters when using ridge, lasso, or both (Hastie et al., 2009). The stability of the XGBoost is high due to the large number of trees in the model which averages out fluctuations. Finally, the stability of SIRUS is generally high too since the algorithm was designed such that the structure of the random trees is more stable (Bénard et al., 2021b). For more details about the analyses, see the code repository at osf.io[10].

4.4 Results

The summary statistics of the variables and correlations for all variables with graduation are respectively shown in Table A1 and Figure A1 and A2 of Huijzer et al. (2023). The average AUC score and standard errors are shown in 4.1. To interpret these ROC curves, note that the diagonal line represents

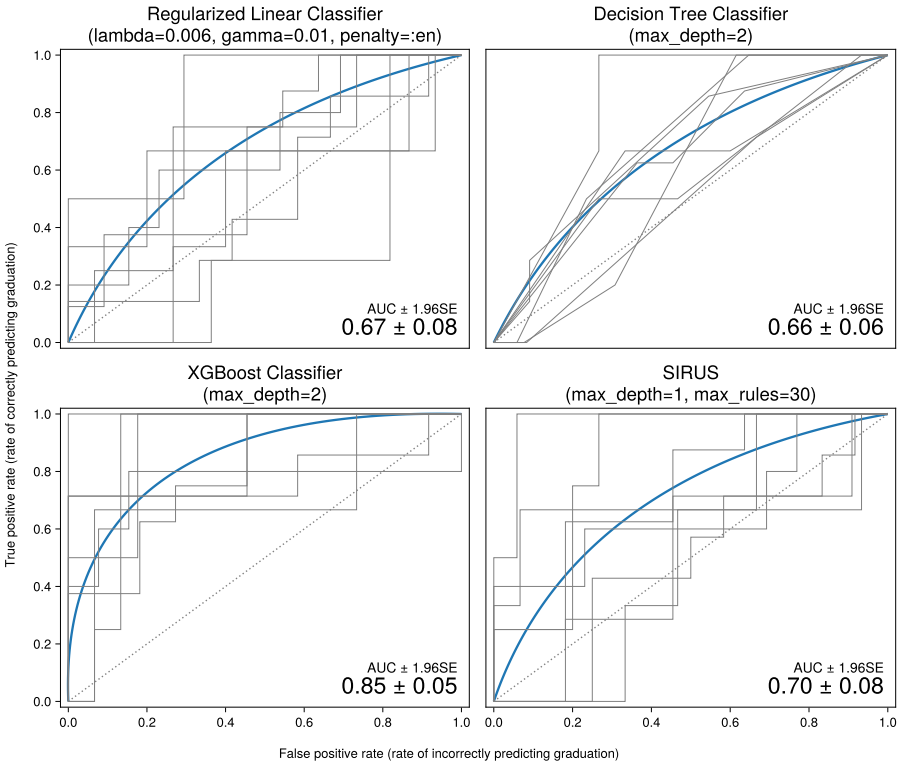
[10]https://osf.io/c8hdy/

random guessing. Next, to create the lines, a model was fitted on one of the cross-validation folds for each fold and used to predict data that the model had not seen during training. Then, note that a classification model can use different thresholds, the lower the threshold, the more likely an individual is classified as graduate. Finally, for each fold, the line is drawn by increasing the model threshold from 0 to 1 and comparing the model predictions to the true values. The AUC score is the averaged area under these curves.

The XGBoost model had the highest predictive performance, which was followed by the SIRUS model with a tree depth of 1 and at most 30 rules. Note that SIRUS with a tree depth of 2 would allow for more complex rules with two elements in the clause (e.g., *if X and Y, then A else B*) instead of only only clause (e.g., *if X, then A else B*). However, fitting a SIRUS model with a tree depth of 2 performed consistently worse, which indicated that the model overfitted the data. The logistic regression and the decision tree had slightly lower predictive performance.

Figure 4.1

Receiver Operating Characteristic (ROC) Curves



Note. The thick lines represent estimates of the average ROC curves over all folds. The smaller lines in gray display the variation on this estimate by showing the the first 8 folds in the 12-fold cross-validation. We show only 8 folds because more folds made the plot very cluttered. The average Area Under the Curve (AUC) and 1.96 ∗ standard error scores are shown in the bottom right.

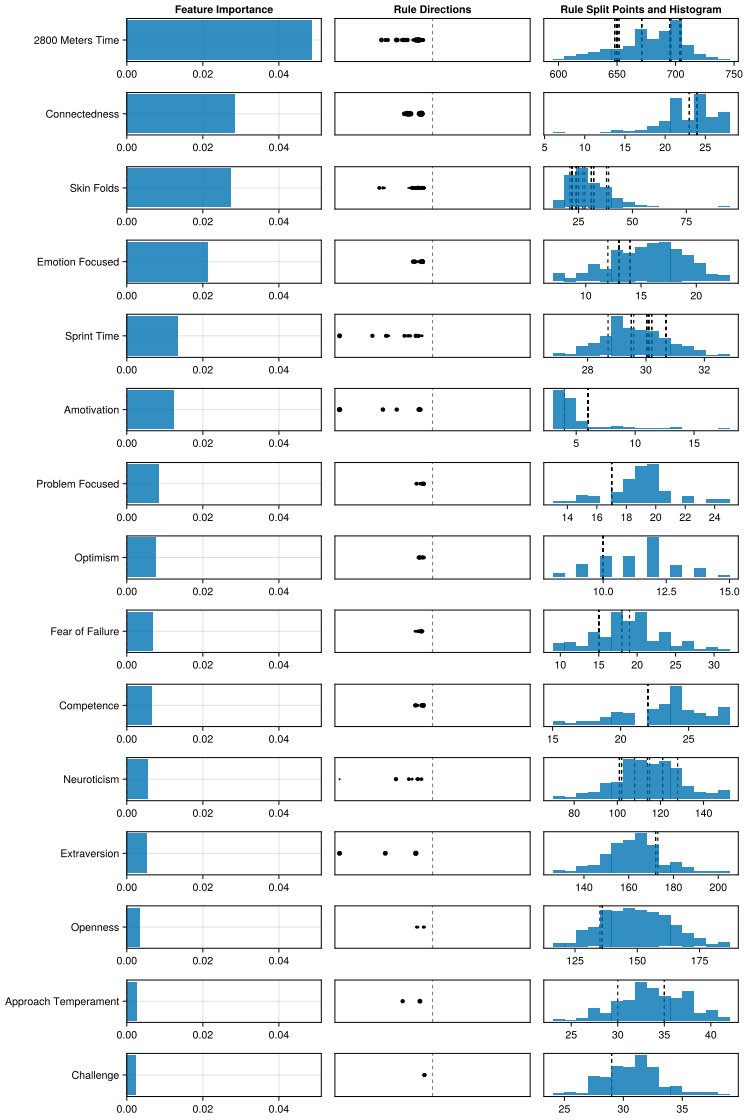
Altogether, while the XGBoost had a good predictive performance, the SIRUS model combined good predictive performance with strong stability and explainability (see Analysis section). We therefore decided to analyse the data further via this model. To do so, we have visualized the stability for different bootstrapped samples in Figure 4.2. Here, by bootstrapped samples, we mean that we took multiple random samples, via MLJ.jl (Blaom et al., 2020), of the data and fitted the model on each of these samples. The bootstrapping

allowed us to visualize the uncertainty in the model which, in turn, aids model explanations.

To inspect the model, we go through one example feature in Figure 4.2. The figure shows that the 2800 meters time had the most importance when summing the feature importances over the various bootstrapped samples. Next, we know that the rules in the SIRUS algorithm with a depth of 1 by default always point to "lower then", for example *if 2800 meters time < 650*, then *then-score* else *else-score* (Huijzer et al., 2023b). If the *then-score* is greater than the *else-score*, then the model predicts that the individual who satisfies the rule is more likely to graduate. If the *then-score* is smaller than *else-score*, then the model predicts that the individual who satisfies the rule is more likely to drop out. The plotted rule directions show the direction of this *then-score* and *else-score* via $\log\left(\frac{\text{else-scores}}{\text{then-scores}}\right)$. Thus, from the plotted rule directions, we can see that the model found that a higher 2800 meters time was associated with drop out. The exact locations of the split points (e.g., *if 2800 meters time < 650*) are shown in the right part of the plot and were different in the different bootstrapped samples. Most of the split points were at 650 seconds, and some where at 700 seconds. We plotted these split points on top of histograms of the data to show the distribution of the data.

Figure 4.2

Rules used by the Rule-Based Classifier in Different Folds



Note. This figure indicates the model uncertainty over different bootstrapped samples. The leftmost column show the feature importance, the middle column shows the directions of the rules, and the rightmost column shows the split points of the rules and a histogram of the data. Specifically, the direction shows $\log\left(\frac{\text{else-scores}}{\text{then-scores}}\right)$. The sizes of the dots indicate the weight that the rule has, so a bigger dot means that a rule plays a larger role in the final outcome. These dots are sized in such a way that a doubling in weight means a doubling in surface size.} Finally, the variables are ordered by the sum of the weights of the rules and only the first 15 are shown.

When looking at all the predictions, the running time on the 2800 meters was the most important with a clear cut-off point for all folds at about 700 seconds. This means that, for all the folds, a higher running time was found to be associated with dropping out. Furthermore, a higher score on, in particular, connectedness and skin folds were associated with dropping out.

4.5 Discussion

The purpose of the current study was to determine how well we could predict dropout of special forces recruits while retaining model explainability and stability. To do so, we compared a linear, decision tree, XGBoost, and SIRUS classifier. Of the four models, the XGBoost had the best predictive performance. This is in line with earlier research that found that XGBoost is a powerful algorithm in a wide array of tasks ranging from predicting Tweet engagements (Anelli et al., 2020) to predicting injuries in competitive runners (Lövdal et al., 2021). However, XGBoost is less explainable than SIRUS. The difference between the two is that the SIRUS algorithm simplifies the model and then uses this model for both explanations and predictions. In contrast, model explainability methods typically use a simplified representation for explanations and the complex model for predictions. This difference between explanations and predictions could hide issues related to safety, fairness (e.g., biases), and reliability which is especially problematic in the context of selection, where wrong decisions can have a lasting impact on the individual. Next, the logistic regression, which is most familiar to sport and performance scientists, was explainable, but not very stable and performed slightly poorer than the SIRUS model. The general instability of the logistic model is an issue that has been described by (Hastie et al., 2009). Furthermore, the decision tree is explainable but not stable (see Molnar, 2022). Together, the algorithm

that displayed the best combination on all aspects was the SIRUS algorithm by achieving a good predictive performance and stability, while remaining explainable.

The SIRUS algorithm appeared to be able to correctly deselect about 10% to 20% of dropouts, that is, without sending recruits home who would have graduated, depending on the fold (see the top right of the SIRUS ROC in Figure 4.1). There is still a considerable amount of variance in the ROC curves, but at least 10% would already be a meaningful number in practice. Moreover, the accuracy of the prediction will most likely improve when fitting the model on the full dataset instead of cross-validation folds and when gathering more data over time.

Since the SIRUS model performs relatively well, and is explainable and stable, we can use our domain knowledge to estimate the generalizability of the model. With this in mind, the main takeaways from the current model are that candidates who take more than roughly 700 seconds on the 2800 meters, score higher on connectedness, and have higher skin folds are more likely to drop out (see Figure 4.2).

The SIRUS algorithm appeared to be able to correctly deselect about 15% to 45% of dropouts, that is, without sending recruits home who would have graduated, depending on the fold (see the top right of the SIRUS ROC in Figure 4.1). There is still a considerable amount of variance in the ROC curves, but at least 15% would already be a meaningful number in practice. Moreover, the accuracy of the prediction will most likely improve when fitting the model on the full dataset instead of cross-validation folds and when gathering more data over time. Since the SIRUS model performs relatively well, and is explainable and stable, we can use our domain knowledge to estimate the generalizability of the model. With this in mind, the main takeaways

from the current model are that candidates who take more than roughly 700 seconds on the 2800 meters, score higher on connectedness, and have higher skin folds are more likely to drop out (see Figure 4.2).

Most of these variables are in accordance with earlier studies. For instance, a lower time for the 3-mile run also predicted graduation in 800 U.S. special forces recruits (Farina et al., 2019). Furthermore, a lower fat percentage, as measured by the skin folds, was associated with physical fitness in 140 Finnish recruits (Mattila et al., 2007). Together, this adds theoretical confidence that the predictive model will generalize to new cohorts.

Limitations and Future Research

Although the psychological measurements were well-organized and based on validated questionnaires, a limitation could be that participants faked their responses (e.g., Galić et al., 2012). To mitigate this in our study, we emphasized that data would be processed anonymously and that staff of the Special Forces unit could not access the data nor use it to make selection decisions, which has been shown to reduce the faking tendency (Kuncel & Borneman, 2007). Nevertheless, to make the transfer to real selection, the risk of faking should be accounted for. For future research, it would be interesting to investigate how selection decisions can be made on the data while new data keeps being added.

Conclusions and Practical Implications

In our attempt to predict dropout of special forces recruits by fitting machine learning models, SIRUS had a higher predictive performance than the linear classifier and decision tree, while being more explainable than the state-of-the-art XGBoost classifier. In other words, SIRUS achieves a balance between predictive performance, explainability, and stability. This

together with its ease-of-use make it particularly suitable for many research problems in science, including selection in sports, and organizational and military contexts. This better understanding of the model may outperform the accuracy of black-box models in the long run, because it allows researchers to improve the model with their domain expertise and improve their domain expertise with the model. In turn, practitioners may use this to make data-driven selection decisions. To conclude, we would encourage scientists to use SIRUS, or similar stable rule-based models. This is especially useful when working in fields, such as sports and military selection, where the number of variables often approaches the number of observations and where predictive performance, explainability, and stability are critical.

CHAPTER 5

Early Identification of Dropouts During the Special Forces Selection Program

This chapter is based on:

Huijzer, R., Blaauw, F. J., De Wit, A., De Jonge, P., & Den Hartigh, R. J. R. (2024). Early Identification of Dropouts During the Special Forces Selection Program. *PsyArXiv*. https://doi.org/10.31234/osf.io/nbs6j

Abstract

Special forces selection is a highly demanding process that involves exposure to high levels of psychological and physical stress resulting in dropout rates of up to 80%. To identify who likely drops out, we assessed a group of 249 recruits, every week of the program, on their experienced psychological and physical stress, recovery, self-efficacy, and motivation. Using both ordinary least squares regression and state-of-the-art machine learning models, we aimed to find the model that could predict dropout best. Furthermore, we inspected the best model to identify the most important predictors of dropout and to evaluate the predictive performance in practice. Via cross-validation, we found that linear regression performed best while remaining interpretable, with an Area Under the Curve (AUC) of 0.69. We also found that low levels of self-efficacy and motivation were significantly associated with dropout. Additionally, we found that dropout could often be predicted multiple weeks in advance and that the AUC score may underestimate the real-world predictive performance. Taken together, these findings offer novel insights in the use of prediction models on repeated measurements of psychological and physical processes, specifically in the context of special forces selection. This offers opportunities for early intervention and support, which could ultimately improve selection success rates.

5.1 Introduction

Special forces are often considered the most elite military units, with the potential to significantly impact strategic military outcomes. They are typically composed of highly trained and motivated individuals who are able to operate in high-stakes environments which are both psychologically and physically demanding. However, dropout rates during the selection process are close to 80% (e.g., Gayton & Kehoe, 2015). This is a concern for both the recruits and the military as it incurs a personal toll on the recruits and is costly for the military. Scientifically, a major challenge is identifying potential dropouts early in the selection period via accurate predictive models. Such models could allow for early intervention on potential future dropouts by intervening the relevant psychological and physical processes.

The relatively scarce previous research investigated dropout by comparing test scores from before the selection period with the final dropout or graduation decision. Psychological tests included, for instance, personality questionnaires and showed that a higher emotional stability and conscientiousness were associated with graduation (e.g., Jackson et al., 2012; Sørlie et al., Huijzer et al.; 2020, 2022; Rolland et al., 1998; Tedeholm et al., 2021). In other research, psychological hardiness was associated with graduation among 1,138 special forces recruits (Bartone et al., 2008) and 178 Norwegian border patrol soldiers (Johnsen et al., 2013). On the other hand, in a study including 73 South African special forces, hardiness and self-efficacy were not associated with graduation (De Beer & Van Heerden, 2014). In another study, higher self-efficacy was significantly associated with graduation among 380 special forces recruits (Gruber et al., 2009).

Physical tests typically include fitness, strength, and endurance tests. For example, in a study among 69 Finnish soldiers, baseline information of aero-

bic fitness significantly predicted graduation (Vaara et al., 2020). In a study on 160 Swedish police counterterrorism intervention units including various psychological and physical tests, the authors found that only running capacity was a significant predictor of graduation (Tedeholm et al., 2023) A study on 800 special forces recruits showed that both psychological and physical tests were significantly associated with graduation (Farina et al., 2019). Finally, a follow-up study on 117 special forces soldiers found that physical characteristics of the body, such as a lower percentage body fat and fat mass were predictors for physical performance and graduation (Farina et al., 2022).

Despite some evidence for the role of psychological and physical factors in predicting dropout, a main issue of previous studies is that they showed limited effects and different predictor combinations. For instance, when comparing agreeableness between military recruits and a civilian control group, agreeableness was found to be lower after training (Jackson et al., 2012), whereas this was not found in two recent studies (Huijzer et al., 2022a; Tedeholm et al., 2023). Such contradicting results could be due to theoretical and methodological factors. Theoretically, a commando profile could be composed of different combinations of characteristics that could allow an individual to perform in highly psychologically and physically demanding situations (e.g., Den Hartigh et al., 2016). Accordingly, and methodology-related, an important factor contributing to dropout is how recruits respond to the stress during the heavy selection program. This cannot be derived from psychological and physical measures taken at one point during the selection program. Thus, an important question is: how do recruits actually respond to, and recover from, the stress to which they are exposed? Such a question can be answered by measuring recruits during the selection period on relevant psychological and physical processes of stress and recovery.

Recent research provided initial evidence that repeated measures can be used to predict dropout. For instance, one longitudinal study on elite soldiers found that recruits who voluntarily dropped out exhibited an increase in emotional or physical pain and a decrease in self-efficacy up to three days before dropping out (Saxon et al., 2020). Similarly, in a study on 46 male and female recruits in the Australian Army basic military training course, higher stress and recovery, as measured via the Short Recovery and Stress Scale (Kellmann & Kölling, 2019), were associated with a higher risk of delayed completion (Tait et al., 2022). Similar results have been found in sports. For example, in a study on 135 adolescent elite athletes, lower recovery and higher stress states as measured by the Acute Recovery and Stress Scale (ARSS) were followed by depressive, burnout, and insomnia symptoms (Gerber et al., 2022). In a study on 74 middle and long-distance runners, recovery and exertion were considered some of the most important variables for predicting injuries (Lövdal et al., 2021, Figure 4). These findings are promising as they suggest that dropout, either voluntary or involuntary (e.g., due to injury), can be predicted in advance based on measures taken during selection or training periods.

Building upon first efforts of predicting dropout from military programs and the increasing interest in the psychological and physical stress monitoring during army training, important statistical strides can be made. Most notably, while previous studies often applied traditional statistical methods, i.e., how variables were associated with dropout or graduation, they often did not report the predictive performance. This means that associations between variables could be too small to be useful in practice or they could be wrong due to overfitting (Yarkoni & Westfall, 2017). Ideally, a study would report predictive performance for multiple models to avoid overfitting and depen-

dence on one model, and use repeated measures to allow for prediction of dropout in advance. For a recent example in the context of the marine corps, see Dijksma et al. (2022).

The current study aimed to assess the experienced psychological and physical stress and recovery of recruits during the selection weeks while improving upon the statistical methods used in previous research. In line with recommendations from previous literature, we specifically focused on the experiences of self-efficacy, motivation, and psychological and physical stress and recovery (Den Hartigh et al., 2022d). We compared various classical and state-of-the-art machine learning models via cross-validation. In addition, we explored the moment at which valid predictions of dropout could be made (e.g., one day, one week, or three weeks in advance). Such knowledge could lead to a better understanding of the dropout process, and to targeted interventions in practice.

5.2 Method

Participants

The sample for this study consisted of 249 male special forces recruits, ranging in age from 18 to 35 years. Prior to their involvement in the study, active informed consent was obtained from each recruit. The information letter informed participants about the study's purpose, procedures, and potential risks, as well as their right to withdraw from the study at any time. The participants were diverse in terms of their military experience, with some being new recruits while others had prior experience in different branches of the armed forces. Due to the sensitive nature of the data, more detailed information about the participants could not be made available.

Measures

During the selection period that lasted up to 16 weeks, we asked the following self-efficacy and motivation questions, both in Dutch: "How confident are you that you can complete the course?" (0 = not confident at all, 100 = very confident) and "How motivated are you to pass the training program?" (0 = not at all motivated, 100 = very motivated). Furthermore, we used a Dutch version of the Short Recovery and Stress Scale (SRSS), a self-report questionnaire assessing perceived stress and recovery levels (Kellmann & Kölling, 2019). The Dutch version underwent a parallel back-translation procedure (Vallerand, 1989). It was subsequently validated in a group of 385 Dutch and Flemish athletes (Brauers et al., 2024). The SRSS consists of 8 items divided into two subscales: Recovery and Stress. Items were rated on a seven-point Likert scale, with higher scores indicating greater levels of recovery or stress. The Recovery subscale evaluates an individual's current state in comparison to their best recovery state ever, with items such as "Physical performance capacity" and "Mental performance capacity". The Stress subscale assesses an individual's current state relative to their highest stress state ever, including items like "Muscle stress" and "Lack of inspiration", see Kellmann & Kölling (2019) for more information. Over the course of the study, the recruits completed the questionnaire weekly, resulting in a total of 1652 responses. On average, we received about 6 responses per person. The number of responses per participant varied due to individuals dropping out of the selection process before the end of the study. The data was collected using an electronic questionnaire, which was administered via a web-based platform that we built for this project. The collection occurred at the start of the training week, which was typically on Monday morning at 0800 hours.

Analysis

We processed the data to include the following 13 columns: *id*, *week*, *motivation*, *self-efficacy*, 8 SRSS items, and whether the individual drops out in the week after the response. Here, we truncated the data at 13 weeks, given that the data was only collected for 14 out of 16 weeks.

Next, we analyzed the model in three ways. We consider none of these ways as definitive, but instead consider each of these ways as a tool to evaluate the model (e.g., McShane et al., Hofman et al.; 2019, 2021). Firstly, we applied principles and techniques from machine learning to estimate the model's ability to predict future behaviors. We used 12-fold cross-validation and the area under the receiver operating characteristic curve (AUC) as a performance metric, both via the MLJ.jl software package (Blaom et al., 2020). The AUC is a measure of the performance of a binary classifier, where a value of 0.5 indicates random guessing and a value of 1.0 indicates perfect predictions. We used the AUC because it is a robust metric that is not sensitive to class imbalance and is a common metric in the literature. Furthermore, we used multiple different models to determine which one performed best in terms of predictive performance. We fitted a binary logistic model with no intercept as our baseline model. Next, we fitted two SIRUS models to the training data as the SIRUS model has shown to perform well in similar situations with relatively few samples and binary outcomes (Bénard et al., 2021a; Huijzer et al., 2023b). SIRUS is based on random forests, and, therefore, non-parametric meaning that it does not make assumptions about the distributions of the data. Random forest-based models are robust to outliers, do not require scaling of the data, and perform very well generally (Biau & Scornet, 2016). Finally, we fitted a modern gradient boosting model called EvoTrees.jl (Desgagne-Bouchard et al., 2024). Gradient boosting models are not fully

interpretable due to the large amounts of trees (e.g., Huijzer et al., 2023), but they are known to perform well in many situations (e.g., Chen & Guestrin, 2016; Ke et al., 2017). In the context of military selection, we prefer models with an optimal trade-off between predictive accuracy and interpretability. Therefore, to combine predictive performance and interpretation (Hofman et al., 2021), we inspected the model that scored best on this trade-off. Specifically, we fitted the model on the full dataset and inspected the fitted model.

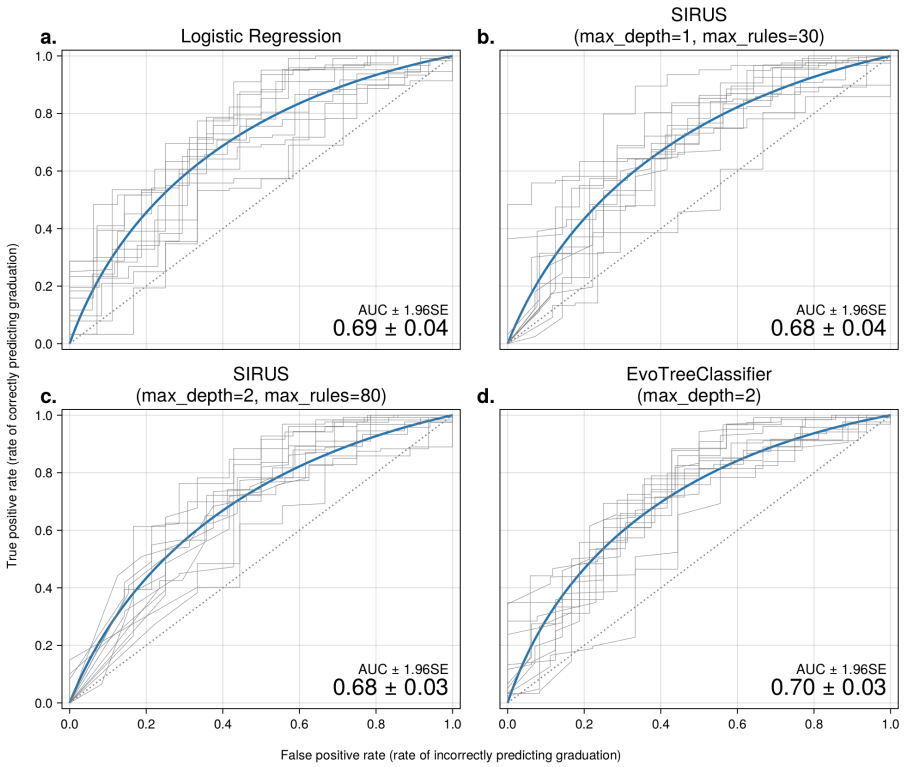
Thirdly, we evaluated the predictive performance in practice. To do so, we converted the predictions of the model in the range of 0 to 1 back to a binary outcome. We did this by choosing a threshold and using this threshold to split the outcomes in dropout and graduate groups. Next, we visualized the predictions of the model for different thresholds. This helps researchers and practitioners in selecting the right balance between the number of false positives and false negatives, and provides an indication of the predictive performance in practice.

5.3 Results

The results for the evaluation runs on the cross-validation data are shown in Figure 5.1.

Figure 5.1

Receiver Operating Characteristic (ROC) Curves



Note. The different lines show the results for all folds in the 12-fold cross-validation. The average Area Under the Curve (AUC) and 1.96 * standard error scores are shown in the bottom right of each graph.

In these results, the bottom two graphs both have a max tree depth of 2. This higher depth allows these models to capture more complex interactions between variables. However, the results show that these models do not perform markedly better than the simpler models, see 5.1. This is likely caused by more

complex models overfitting the data and could likely be solved by using more data. In general, the logistic regression model performs best since it scores best in the trade-off between predictive performance and interpretability. The interpetability is very high because the algorithm is very simple compared to the thousands of trees in gradient boosting models, and the performance is very comparable to the gradient boosting model. Therefore, we inspect the logistic regression model in more detail below.

The coefficients of the logistic model, when fitted on the full dataset, are shown in 5.1. When interpreting this model, note that there is variation in performance for the different cross-validation folds, see Figure 5.1. This is why we decided post hoc to set our alpha level conservatively to 0.001 instead of the commonly used 0.05. This lower alpha level means that we are less likely to find significant results. Setting this level post hoc seemed reasonable as we use the p-value as just one of the many tools to interpret the model (e.g., McShane et al., 2019. From Table 5.1, we can see that the variables "Self-Efficacy" and "Motivation" were significant. The positive coefficients indicate that recruits who score higher of self-efficacy and higher on motivation are less likely to drop out.

Table 5.1

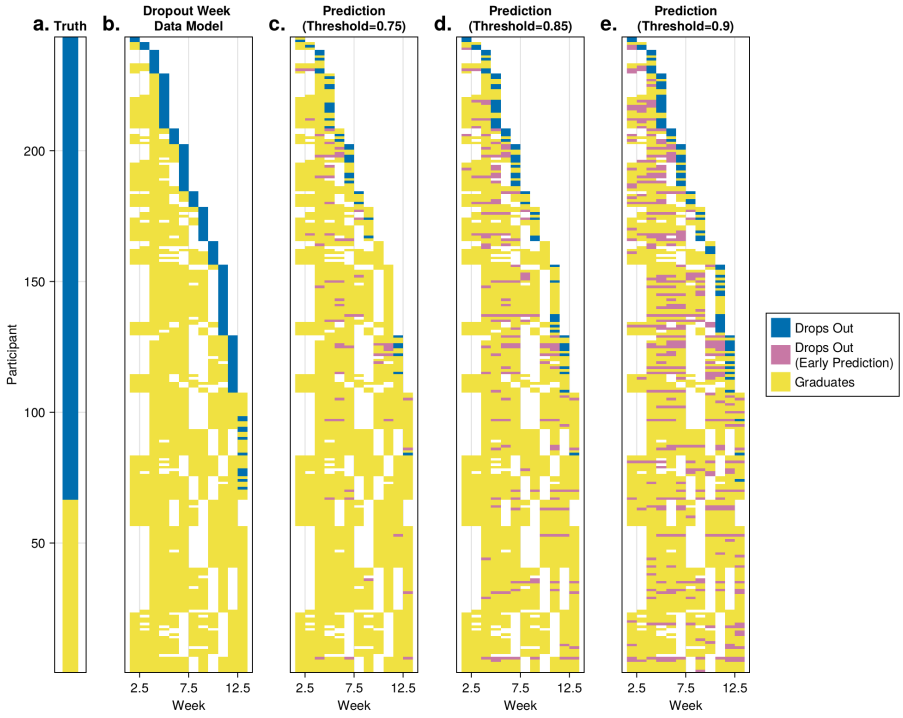
Fitted Binary Logistic Regression Statistics

Variable	Coefficient	Z-Score	*p*-Value	Lower 95%	Upper 95%
Self-Efficacy	1.734	4.78	<0.001	1.022	2.446
Motivation	1.205	3.6	<0.001	0.549	1.86
Muscle Tension	0.561	2.18	0.029	0.056	1.066
Overall Stress	−0.701	−1.95	0.051	−1.405	0.003
Lack of Enthusiasm	0.531	1.38	0.168	−0.221	1.282
Negative Emotional State	−0.493	−1.38	0.168	−1.194	0.209
Emotional Balance	0.406	1.24	0.215	−0.233	1.044
Overall Recovery	0.341	0.89	0.373	−0.413	1.096
Physical Performance	−0.343	−0.83	0.407	−1.157	0.47
Mental Performance	−0.245	−0.6	0.549	−1.049	0.56
Recovery	0.225	0.55	0.582	−0.574	1.025

Next, the predictions made by the logistic regression model are visualized in Figure 5.2. The figure shows that many of the dropouts were predicted correctly in the last week, which is in line with the AUC score as reported in Figure 5.1. Furthermore, some dropouts were predicted weeks before the actual dropout event. This suggests that the reported AUC score underestimates the actual predictive performance, since our data is modeled such that a dropout prediction is only considered correct if it is made in the week before the dropout event.

Figure 5.2

True Dropout Data and Predictions of the Model



Note. This figure shows the true points of drop out for each participant in the leftmost subfigure. The second subfigure shows how the data was modeled. The aim was to train a model that could predict dropout events. The other three subfigures show the predictions according to the model for different thresholds. Different thresholds allow practitioners to select the right balance between the number of false positives and false negatives. This, together with the AUC, provides an indication of the predictive performance in practice.

5.4 Discussion

The current study aimed to predict dropout during the special forces selection period. To that end, we assessed the recruits on psychological and physical factors related to stress and recovery during this period. We applied simple logistic models as well as more complex models on this data. Next, we used various tools to analyze the model. Specifically, we evaluated how

well each model performs, we interpreted the best model, and evaluated the predictive performance in practice. We found that a simple logistic regression model scored best on the trade-off between predictive performance and interpretability because it was interpretable and performed relatively well with an area under the curve (AUC) of 0.69. The most complex models scored only slightly better on the AUC, which suggested we had insufficient data for more complex models.

The logistic regression model's revealed that self-efficacy and motivation were significantly related to dropout. This provides support for earlier research that found that decreases in self-efficacy were related to dropout in a military context (Saxon et al., 2020). More generally, it is in accordance with the perspective that temporal measures of self-efficacy and motivation can provide important information on an individual's resilience. That is, motivation and self-efficacy are important psychological performance factors that ideally return to normal levels following psychological and physical stress. When individuals loose resilience, as reflected in their self-efficacy and motivation levels, then this could be a warning signal for negative outcomes such as psychological problems or dropout (for a review, see Den Hartigh et al., 2022). Interesting in this regard is that more direct measures of stress and recovery experiences were less predictive of dropout. One reason for this could be that the individual questions are more sensitive than items containing multiple questions. Put differently, in items with multiple questions, variations tend to average out, making it less likely that the items become significantly related to dropout. Another reason for this could be that the SRSS has, so far, only been validated in the sports context. Despite the parallels between the sport and military context, individuals are typically exposed to more extreme psychological and physical stress during the selection program. It could be

that the experience of stress and recovery are so high for everyone, that it cannot account for the variance in the outcome anymore.

Finally, we estimated the predictive performance in practice. We visualized the predictions of the model for different thresholds. This showed that the model could sometimes predict dropout multiple weeks in advance with few false positives, depending on the chosen threshold. In practice, this means that the calculated AUC scores may underestimate the predictive performance due to the way the data was modeled. Note that choosing the right threshold is important as it determines the balance between the number of false positives and false negatives. We showed multiple thresholds which could be used by practitioners to select the right balance. Since the cost of missing a dropout is high, we recommend a higher threshold, which would result in more early warnings of dropout.

Future work could improve upon the current study in several ways. First, the sample size was relatively small for machine learning models. With a higher sample size, the variation in the cross-validation folds would most likely decrease. Second, the frequency of measurements could be increased. More frequent measurements could provide more opportunities for early intervention and support. Third, this study could be complemented with quantitative measures to gain deeper insights into the personal experiences, coping strategies, and psychological states of recruits. This could help refine the predictive models and identify potential areas for intervention. Finally, intervention studies could be conducted based on the predictive models to design and test interventions aimed at reducing dropout rates. These could include psychological resilience training, targeted physical conditioning programs, or personalized support strategies.

Taken together, our study builds on previous research that has highlighted the importance of psychological and physical factors in predicting dropout in special forces selection. The longitudinal design of our study adds to this body of knowledge by demonstrating that dropout may be predicted during the selection program, offering more opportunities for early intervention and support. Even more so, by picking the right threshold, individuals at risk of dropout could sometimes be identified weeks in advance. This allows for targeted interventions and support, which could subsequently improve success rates and reduce the personal and human resource costs associated with high dropout rates.

6 General Discussion

We investigated whether it is possible to predict who will drop out from special forces selection. To do so, we gathered data to compare special forces operators with the general population, and special forces dropouts with graduates on personality traits, see Chapter 2. We found that successful recruits and operators are typically less neurotic and more conscientious when compared to respectively the general population and dropouts. These results were in line with previous research in high-stakes contexts (Braun et al., 1994; Campbell et al., 2010; Lee et al., 2011; McDonald et al., 1990). However, although we found effects in the expected directions for the dropout versus graduate comparison, they were not significant and had small to medium effect sizes.

These small effect sizes made it difficult to predict who will drop out on personality only. As other researchers put it: The traditional focus on explanations has led to intricate theories that have little (or unknown) ability to predict future behaviors (Yarkoni & Westfall, 2017). This could be like a car being judged on speed alone. For a while this could lead to better cars, but at some point, manufacturers will ignore other important factors such as comfort. With this in mind, we continued this thesis by being more focused on predictive performance. Note that solely aiming at good predictions (or comfort) does not solve all problems either. Too much focus on prediction could lead to intricate models that may lead to unsafe, unfair, or unreliable predictions (Barredo Arrieta et al., 2020; Doshi-Velez & Kim, 2017). For example, to explain how a model can be unfair or unreliable, suppose we only focus on predictive performance and the model decides to filter out all individuals born in October. The model does this because it has seen that

individuals born in October are more likely to drop out. This, so called overfitting, would likely lead to unfair and unreliable predictions. To mitigate this, we can focus on both prediction and explanation (Hofman et al., 2021) since overfitting is less likely to happen if we understand why a model makes certain predictions.

With our renewed focus on predictive performance, we found that not all prediction models were suitable for our research, because models had either a lower stability, interpretability, or predictive performance. In an attempt to improve this, we implemented our own version of the Stable and Interpretable RUle Sets (SIRUS) algorithm, and evaluated its predictive performance on multiple dataset, see Chapter 3. In the end, SIRUS scored similar in predictive performance to linear regression, see Table 3.2. This was surprising as the linear regression and SIRUS models fit the data very differently. In theory, the SIRUS model should have a strong benefit for fragile systems such as personnel selection. In such systems, negative effects are often nonlinear (Hill et al., 2024; Taleb, 2020). For example, falling from a height of 10 meters is more than 10 times as harmful as falling from a height of 1 meter. Similarly, a recruit that takes 10 seconds longer than average on a 2800 meters run could be much more likely to dropout than a recruit that takes just 1 second longer. In theory, the SIRUS model could capture these nonlinear effects better than linear regression. This is because the SIRUS model is based on random forests, which functions by finding splitpoints in the data, which do not have to be linear. However, we did not find evidence for this when testing the model on the fragile system-based datasets such as the Haberman, breast cancer, and diabetes datasets, see Table 3.2. There could be many reasons for this.

One could be that our implementation of the SIRUS model was not optimal. Another could be that the datasets were not large enough to detect the more complex patterns.

Next, we tested different models, including the SIRUS model, on a large special forces dataset in Chapter 4. This dataset consisted of physical measures (e.g., number of push-ups and 2800 meters running time) and psychological measures (e.g., the NEO-PI-3 personality questionnaire) and was gathered in the first week of the selection period. When aiming to predict who would drop out, we combined the predictions with model explanations. For the predictions, we found that XGBoost performed best in terms of predictive performance. However, the model is too complex to interpret directly. Therefore, interpretations methods, such as SHAP (Lundberg & Lee, 2017), are needed which require a simplified representation of the model. This simplification could hide biases or reliability issues. With the other models, we found that SIRUS performed slightly better than the linear model on our dataset. We also found that the SIRUS model was able to predict dropouts with a good accuracy, while retaining stability and interpretability.

Based on the ROC curves, Figure 4.1, we suspect that the SIRUS model can likely filter out about 10% of dropouts without losing graduates. In general, we expect that models are better at predicting dropouts than graduates (Hunt et al., 2011; Taleb, 2013). This could be because many things have to go right for a recruit to graduate, while only one thing has to go wrong for an individual to drop out. There are an infinite number of ways in which the recruit could be hindered from graduating, also known as *black swan* events (Taleb, 2010). For example, breaking a leg or suddenly deciding to quit could lead to immediate, and permanent, dropout. Maybe the model can predict this by finding a clue in the data, such as a poor running time or a low

motivation. Conversely, since the data that the model sees is limited, it is much harder to predict that everything will go right.

While the results of Chapter 4 are promising, they do not provide information on how recruits respond to the high levels of physical and psychological stress during the selection period. Having this information could help to predict dropout more accurately. Therefore, in Chapter 5 we used a shorter questionnaire that was filled in each week instead of a longer questionnaire that was only filled in the first week. We found that a higher self-efficacy and motivation were significantly related to dropout. This means that how participants responded at the start of the week was related to whether they would soon drop out. With these and other variables, the model achieved an average AUC of 0.69, which means it could be useful in practice. After estimating the predictive performance in practice, we found that the linear regression model could sometimes predict dropout multiple weeks in advance with few false positives.

In conclusion, to answer the question whether we can predict who will drop out from special forces selection: From Chapter 4 and 5, it looks like we can predict dropout reasonably well. The average AUC of about 0.7 generally means a reasonable predictive performance (e.g., Hosmer et al., 2013, p. 177). When following recruits over time, we can sometimes predict dropout multiple weeks in advance and use that to conduct interventions. These results are promising. Next, it is important to confirm these results in practice since that is the only way to know with certainty how well the model works.

6.1 Future Research

Future research could investigate whether improvements can be made in the predictive performance of psychological questionnaires. In our research,

traditional psychological measures performed poorer than physical measures or individual questions. More specifically, the NEO-PI-3 personality test is widely used and regarded as a good measure with high validity and reliability, but it predicted poorer than physical tests, such as 2800 meters time, in Chapter 4 and Tedeholm et al. (2023). This is in contrast to the experiences of both Dutch special forces operators and U.S. Navy SEALs, who reported that psychology played a more important role than physical fitness in their selection. They witnessed many individuals with excellent physical fitness drop out, and many individuals with poor physical fitness make it through. This could mean that there is still room for improvement in the predictive performance of psychological questionnaires. Similarly, in Chapter 5, the Short Recovery and Stress Scale (SRSS) appeared to perform poorer than the two, newly added, self-efficacy and motivation questions. This could be because the self-efficacy and motivation questions were individual questions while the SRSS consisted of multiple questions for each item. Individual questions could provide a stronger signal because more questions per item makes it more likely that the signal is averaged out. Put differently, the chance that a participant answers "extremely likely" on one question is higher than the chance that a participant answers "extremely likely" on multiple questions. It could also be that participants are less willing to participate in questions which appear to be similar. Especially in a longitudinal study, participants might lose interest when they need to answer multiple similar questions repeatedly. This could imply that a high questionnaire validity and reliability does not imply a high predictive performance, at least in the context of special forces selection. This was also found in another longitudinal study by Song et al. (2023). In this study, the authors found that single items obtained significant predictive validity, and would sometimes show a larger effect size than using multiple items. Also,

they found that multiple items would only perform moderately better than single items.

Future research could investigate whether it is possible to use individual questions instead of a subscale[11]. From an explanation perspective this suggestion might seem counter-intuitive because it would hinder reliability evaluations, but from a prediction perspective it could be useful by increasing sensitivity. In Chapter 5, the self-efficacy and motivation questions were individual questions which predicted dropout well. Fitting models on such individual questions could help to detect those questions that are most predictive in the specific context in which it is used. Put differently, instead of relying on a set of pre-determined questions, future work could use a data-driven approach to find the most suitable questions for each study. Note that this does require sufficient data, as the number of participants should typically be above 10 times as high as the number of questions to prevent overfitting (e.g., Peduzzi et al., 1996).

Relatedly, future research could investigate whether more specific questions or questionnaires could improve predictive performance for special contexts. Instead of taking a questionnaire that is widely considered valid and reliable (e.g., the NEO-PI-3) researchers could aim to find those questions or questionnaires that predict well in their specific context, such as special forces selection. For example, the single self-efficacy and motivation questions that we used in Chapter 5 predicted dropout well. With this, researchers could continuously monitor the predictive performance of questions or questionnaires and drop those that do not predict well while occasionally adding new questions or questionnaires and evaluating those. This pipeline of adding a new question or questionnaire, evaluating its performance, and adjusting

[11] A *subscale* here means a group of questions that are combined into one score.

would not be new. It is already commonly applied in, for example, social media, search engine, or self-driving car companies. These companies continuously adjust their models because the real world is continuously changing too. They do this by continuously updating their models, then testing them internally, then testing them with a small group of testers, and then finally sending them to all users. Psychology researchers could, for example, do the same by continuously updating their models, then testing them internally by interpreting the models and evaluating the predictive performance, and then sending them to the real world.

Furthermore, in this thesis, we have applied linear models and decision tree-based models. These models could be a problem for fragile systems, such as personnel selection, where negative effects are nonlinear (Taleb, 2020). For example, as pointed out earlier, falling from a height of 10 meters is more than 10 times as harmful as falling from a height of 1 meter. As another example, it looks like scoring lower on an intelligence test was exponentially related to mortality (O'Toole, 1990, Table 2). Another example could be visible in the raincloud plots in Figure 2 of Pattyn et al. (2024). In this figure on Belgian special forces, there are clear cutoff points visible in the data. Future research could investigate whether the use of models which could fit these patterns better could improve predictive performance. A perfect model for such a system could be one which combines the best of linear and tree-based models. It would cut the data into different parts, and then use a linear model (or exponential) for each part. Models like this exist (e.g., Raymaekers et al., 2023), but might need additional constraints to improve performance on small datasets.

Finally, this thesis mentioned the story of the blind men and an elepant. The blind men individually come to different (and incorrect) conclusions

about what an elephant is. When they together look at the elephant from different angles, they come to a much better conclusion. This thesis has looked into widely varying methods to predict dropout from special forces selection. We used frequentist statistics, Bayesian statistics, machine learning, single-question questionnaire items, and more. Hopefully, combining these different perspectives resulted in a more complete picture, which will allow future researchers to see the full elephant.

7 Nederlanse samenvatting (Dutch Summary)

Stelt u zich eens voor dat u vooraf kunt zeggen welke individuen een pilotenopleiding, een Harvard voorselectie, of zelfs in NASA's astronauten selectie kunnen halen. Dit zou veel teleurstelling bij individuen kunnen voorkomen en het zou organisaties veel tijd en geld kunnen besparen. Helaas zijn dit soort voorspellingen bijzonder moeilijk te maken. Een recruit kan bijvoorbeeld de beste in zijn opleiding zijn, maar toch uitvallen door problemen thuis of een misstap in een konijnenhol.

Samen met het Korps Commandotroepen hebben wij onderzocht of wij kunnen voorspellen wie gaat uitvallen van de opleiding. Om dit te doen hebben wij data van 275 recruten verzameld in de vorm van sportdata (zoals 2800 meter hardlooptijd) en psychologische data (zoals persoonlijkheidsvragenlijsten). In hoofdstuk 2 hebben wij deze persoonlijkheids data uit de eerste week van de opleiding gebruikt. Op basis van deze data hebben wij onderzocht of wij persoonlijkheidskenmerken konden gebruiken om onderscheid te maken tussen mensen die zijn uitgevallen en mensen die zijn geslaagd. Ook hebben wij onze data vergeleken met data van Nederlandse burgers uit een ander onderzoek. We vonden dat commando's typisch minder neurotisch, meer consciëntieus en minder open voor ervaringen waren dan de burgers. Voor de geslaagden vonden wij dat ze typisch minder neurotisch en meer consciëntieus leken te zijn. Voor selectie leken de persoonlijkheidskenmerken niet voldoende voorspellend te zijn.

Vanaf Hoofdstuk 3 hebben wij ons meer gericht op het voorspellen van uitval. In onze data analyses vonden wij dat veel statistische modellen niet goed presteerden op het gebied van voorspellend vermogen, stabiliteit of uitlegbaarheid. Daarom hebben wij een statistisch model wat bekend staat

als *Stable and Interpretable RUle Sets* (SIRUS) opnieuw geïmplementeerd in de programmeertaal Julia. Het doel van deze open-source implementatie was om het model beter te begrijpen en om het model beter toe te kunnen passen op onze data. Tevens zorgde deze vertaling ervoor dat het aantal regels code gereduceerd kon worden. Dit verhoogde de leesbaarheid voor onzelf en staat hopelijk in de toekomst andere onderzoekers toe om het algorithme te verbeteren of als de basis voor nieuwe algorithmes. Wij hebben het voorspellend vermogen van het model vergeleken met andere modellen en de originele implementatie. Hieruit bleek dat het voorspellend vermogen van onze implementatie vergelijkbaar was met de originele implementatie in de R programmeertaal.

In hoofdstuk 4 hebben wij deze nieuwe implementatie en enkele andere modellen weer toegepast op de data uit de eerste week van de opleiding. Dit keer hebben wij niet alleen naar persoonlijkheid gekeken, maar ook naar de sportdata. Op deze data hebben wij vervolgens vier verschillende modellen vergeleken op voorspellingsvermogen, stabiliteit en uitlegbaarheid. We vonden dat het SIRUS model het meest geschikt was voor het voorspellen van uitval. Ook vonden wij dat fysieke en psychologische data beide gerelateerd waren aan uitval. Meer specifiek, een langzamere score op de 2800 meter hardlooptijd, verbondenheid, en een huidplooimeting waren sterk gerelateerd aan uitval.

In hoofdstuk 5 hebben wij onderzocht of wij de voorspellingen konden verbeteren door de data gedurende de hele opleiding te verzamelen. Deze vragenlijsten waren korter, maar werden iedere week afgenomen in plaats van alleen in de eerste week. Wij hebben opnieuw meerdere machine learning modellen op de data getest. In dit geval was een lineaire regressie model het meest geschikt voor het voorspellen van uitval. Met dit model vonden

wij dat lagere scores op zelfeffectiviteit en motivatie geassocieerd waren met uitval. We vonden ook dat het model in veel gevallen uitval al enkele weken van tevoren kon voorspellen. Dit biedt mogelijkheden voor het vroegtijdig ingrijpen om uitval te voorkomen.

In conclusie, om de vraag te beantwoorden of wij kunnen voorspellen van de special forces opleiding: Uit hoofdstuk 4 en 5 lijkt het erop dat wij uitval redelijk goed kunnen voorspellen. Het volgende doel is om deze resultaten in de praktijk te testen, omdat dat de enige manier is om met zekerheid te weten hoe goed deze technieken werken.

8 Bibliography

Abbate, J. (2000). *Inventing the internet*. MIT press.

Abt, G., Jobson, S., Morin, J.-B., Passfield, L., Sampaio, J., Sunderland, C., & Twist, C. (2022). Raising the bar in sports performance research. *Journal of Sports Sciences*, *40*(2), 125–129. https://doi.org/10.1080/02640414.2021.2024334

Anelli, V. W., Delić, A., Sottocornola, G., Smith, J., Andrade, N., Belli, L., Bronstein, M., Gupta, A., Ira Ktena, S., Lung-Yut-Fong, A., & others. (2020). RecSys 2020 challenge workshop: engagement prediction on Twitter's home timeline. *Proceedings of the 14th ACM Conference on Recommender Systems*, 623–627. https://doi.org/10.1145/3383313.3411532

Ashari, A., Paryudi, I., & Tjoa, A. M. (2013). Performance comparison between Naïve Bayes, decision tree and k-nearest neighbor in searching alternative design in an energy simulation tool. *International Journal of Advanced Computer Science and Applications*, *4*(11). https://doi.org/10.14569/IJACSA.2013.041105

Bandura, A. (2006). *Guide for constructing self-efficacy scales*. Greenwich, CT.

Banks, L. M. (2006). *The History of Special Operations Psychological Selection.*

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bartone, P. T., Roland, R. R., Picano, J. J., & Williams, T. J. (2008). Psychological hardiness predicts success in US Army Special Forces candidates. *International Journal of Selection and Assessment*, *16*(1), 78–81. https://doi.org/10.1111/j.1468-2389.2008.00412.x

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*(1), 65–98. https://doi.org/10.1137/141000671

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

Blaom, A. D., Kiraly, F., Lienart, T., Simillides, Y., Arenas, D., & Vollmer, S. J. (2020). MLJ: A Julia package for composable machine learning. *Journal of Open Source Software*, *5*(55), 2704. https://doi.org/10.21105/joss.02704

Brailey, M. (2005). *The Transformation of Special Operations Forces in Contemporary Conflict: Strategy, Missions, Organisation and Tactics*. Land Warfare Studies Centre.

Brauers, J. J., Den Hartigh, R. J. R., Jakowski, S., Kellmann, M., Wylleman, P., Lemmink, K. A. P. M., & Brink, M. S. (2024). Monitoring the recovery-stress states of athletes: Psychometric properties of the Acute Recovery and Stress Scale and Short Recovery and Stress Scale among

Dutch and Flemish Athletes. *Journal of Sports Sciences*. https://doi.org/10.1080/02640414.2024.2325783

Braun, D. E., Prusaczyk, W. K., Goforth, H. W., & Pratt, N. C. (1994). *Personality profiles of US navy sea-air-land (SEAL) personnel*. Naval Health Research Center San Diego, CA.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Bénard, C., Biau, G., Da Veiga, S., & Scornet, E. (2021b). Interpretable random forests via rule extraction. *International Conference on Artificial Intelligence and Statistics*, 937–945.

Bénard, C., Biau, G., Da Veiga, S., & Scornet, E. (2021a). SIRUS: Stable and Interpretable RUle Set for classification. *Electronic Journal of Statistics*, *15*(1), 427–505. https://doi.org/10.1214/20-EJS1792

Campbell, J. S., Castaneda, M., & Pulos, S. (2010). Meta-analysis of personality assessments as predictors of military aviation training success. *The International Journal of Aviation Psychology*, *20*(1), 92–109. https://doi.org/10.1080/10508410903415872

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *The handbook for the Sixteen Personality Factor Questionnaire.*

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., Ford, C., Volcic, R., & De Rosario, H. (2017). *pwr: Basic functions for power analysis*.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Clough, P., Earle, K., & Sewell, D. (2002). Mental toughness: The concept and its measurement. *Solutions in Sport Psychology*, *1*, 32–45.

Condon, D. M., & Revelle, W. (2014). The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence*, *43*, 52–64. https://doi.org/10.1016/j.intell.2014.01.004

Costa, J., Paul T, McCrae, R. R., & Martin, T. A. (2008). Incipient adult personality: The NEO-PI-3 in middle-school-aged children. *British Journal of Developmental Psychology*, *26*(1), 71–89. https://doi.org/10.1348/026151007X196273

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual.*

Cranmer, M. (2023). Interpretable machine learning for science with PySR and SymbolicRegression.jl. *Arxiv Preprint Arxiv:2305.01582*. https://doi.org/10.48550/arXiv.2305.01582

De Beer, M., & Van Heerden, A. (2014). Exploring the role of motivational and coping resources in a Special Forces selection process. *SA Journal of Industrial Psychology*, *40*(1), 1–13. https://doi.org/10.4102/sajip.v43i0.1440

De Fruyt, F., De Bolle, M., McCrae, R. R., Terracciano, A., Costa, J., Paul T, & Collaborators of the Adolescent Personality Profiles of Cultures Project. (2009). Assessing the universal structure of personality in early adolescence: The NEO-PI-R and NEO-PI-3 in 24 cultures. *Assessment*, *16*(3), 301–311. https://doi.org/10.1177/1073191109333760

Den Hartigh, R. J. R., Hill, Y., & Van Geert, P. L. C. (2018b). The Development of Talent in Sports: A Dynamic Network Approach. *Complexity*, *2018*. https://doi.org/10.1155/2018/9280154

Den Hartigh, R. J. R., Meerhoff, R. A., Van Yperen, N. W., Neumann, N. D., Brauers, J. J., Frencken, W. G., Emerencia, A., Hill, Y., Platvoet, S., Atzmueller, M., Lemmink, K. A. P. M., & Brink, M. S. (2022d). Resilience in sports: a multidisciplinary, dynamic, and personalized perspective. *International Review of Sport and Exercise Psychology*, 1–23. https://doi.org/10.1080/1750984X.2022.2039749

Den Hartigh, R. J. R., Niessen, A. S. M., Frencken, W. G., & Meijer, R. R. (2018a). Selection procedures in sports: Improving predictions of athletes' future performance. *European Journal of Sport Science*, *18*(9), 1191–1198. https://doi.org/10.1080/17461391.2018.1480662

Den Hartigh, R. J. R., Van Dijk, M. W. G., Steenbeek, H. W., & Van Geert, P. L. C. (2016c). A dynamic network model to explain the development of excellent human performance. *Frontiers in Psychology*, *7*, 532. https://doi.org/10.3389/fpsyg.2016.00532

Desgagne-Bouchard, J., Blaom, A., Qin, A., Widmann, D., Lienart, T., Aluthge, D., Avital, R., Waczak, J., Ling, J., & Gennatas, S. (2024). *EvoTrees.jl*. CERN. https://doi.org/10.5281/zenodo.10569605

Dijksma, I., Hof, M. H., Lucas, C., & Stuiver, M. M. (2022). Development and validation of a dynamically updated prediction model for attrition from marine recruit training. *Journal of Strength and Conditioning Research*, *36*(9), 2523. https://doi.org/10.1519/JSC.0000000000003910

Do, M. H., & Minbashian, A. (2020). Higher-order personality factors and leadership outcomes: A meta-analysis. *Personality and Individual Differences*, *163*, 110058. https://doi.org/10.1016/j.paid.2020.110058

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *Arxiv Preprint Arxiv:1702.08608*. https://doi.org/10.48550/arXiv.1702.08608

Durnin, J. V., & Womersley, J. (1974). Body fat assessed from total body density and its estimation from skinfold thickness: measurements on 481 men and women aged from 16 to 72 years. *British Journal of Nutrition*, *32*(1), 77–97. https://doi.org/10.1079/BJN19740060

Dweck, C. S. (2000). *Self-theories: Their role in motivation, personality, and development*. Psychology press. https://doi.org/10.4324/9781315783048

Eaton, J. P., & Haas, C. (1995). *Titanic: Triumph and tragedy*. WW Norton & Company.

Egger, J. I., De Mey, H. R., Derksen, J. J., & Van der Staak, C. P. (2003). Cross-cultural replication of the five-factor model and comparison of the NEO-PI-R and MMPI-2 PSY-5 scales in a Dutch psychiatric sample. *Psychological Assessment*, *15*(1), 81. https://doi.org/10.1037/1040-3590.15.1.81

Elliot, A. J., & Thrash, T. M. (2010). Approach and Avoidance Temperament as Basic Dimensions of Personality. *Journal of Personality*, *78*(3), 865–906. https://doi.org/10.1111/j.1467-6494.2010.00636.x

Ellis, A., & Conrad, H. S. (1948). The validity of personality inventories in military practice. *Psychological Bulletin*, *45*(5), 385. https://doi.org/10.1037/h0056021

Farina, E. K., Thompson, L. A., Knapik, J. J., Pasiakos, S. M., McClung, J. P., & Lieberman, H. R. (2019). Physical performance, demographic, psychological, and physiological predictors of success in the US Army Special Forces Assessment and Selection course. *Physiology & Behavior*, *210*, 112647. https://doi.org/10.1016/j.physbeh.2019.112647

Farina, E. K., Thompson, L. A., Knapik, J. J., Pasiakos, S. M., McClung, J. P., & Lieberman, H. R. (2022). Anthropometrics and body composition predict physical performance and selection to attend special forces training in United States army soldiers. *Military Medicine*, *187*(11–12), 1381–1388. https://doi.org/10.1093/milmed/usab315

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Fleishman, E. A. (1953). Leadership climate, human relations training, and supervisory behavior. *Personnel Psychology*, *6*(2), 205–222. https://doi.org/10.1111/j.1744-6570.1953.tb01040.x

Florida, R., & Browdy, D. (1991). The invention that got away. *Journal of Technology Transfer*, *16*, 19–28. https://doi.org/10.1007/BF02371304

Fountoulakis, K. N., Siamouli, M., Moysidou, S., Pantoula, E., Moutou, K., Panagiotidis, P., Kemeridou, M., Mavridou, E., Loli, E., Batsiari, E., Preti, A., Tondo, L., Gonda, X., Mobayed, N., Akiskal, K., Akiskal, H., Costa, P., & McCrae, R. (2014). Standardization of the NEO-PI-3 in the Greek general population. *Annals of General Psychiatry*, *13*, 1–8. https://doi.org/10.1186/s12991-014-0036-9

Fox, W. C. (1953,). *Signal Detectability: A unified description of statistical methods employing fixed and sequential observation processes*. Department of Electrical Engineering. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/4946/bac2264.0001.001.pdf

Galić, Z., Jerneić, Ž., & Kovačić, M. P. (2012). Do Applicants Fake Their Personality Questionnaire Responses and how Successful are Their Attempts? A Case of Military Pilot Cadet Selection.

International Journal of Selection and Assessment, *20*(2), 229–241. https://doi.org/10.1111/j.1468-2389.2012.00595.x

Gayton, S. D., & Kehoe, E. J. (2015). A prospective study of character strengths as predictors of selection into the Australian army special force. *Military Medicine*, *180*(2), 151–157. https://doi.org/10.7205/MILMED-D-14-00181

Ge, H., Xu, K., & Ghahramani, Z. (2018). Turing: a language for flexible probabilistic inference. *International Conference on Artificial Intelligence and Statistics*, 1682–1690.

Gelman, A. (2018). The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin*, *44*(1), 16–23. https://doi.org/10.1177/0146167217729162

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2021). *Bayesian data analysis, third edition*. Chapman, Hall/CRC.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211. https://doi.org/10.1080/19345747.2011.618213

Gerber, M., Lang, C., Brand, S., Gygax, B., Ludyga, S., Müller, C., Ramseyer, S., & Jakowski, S. (2022). Perceived recovery and stress states as predictors of depressive, burnout, and insomnia symptoms among adolescent elite athletes. *Sports Psychiatry: Journal of Sports and Exercise Psychiatry*. https://doi.org/10.1024/2674-0052/a000017

Gillham, N. W. (2001). Sir Francis Galton and the birth of eugenics. *Annual Review of Genetics*, *35*(1), 83–101. https://doi.org/10.1146/annurev.genet.35.102401.090055

Gruber, K. A., Kilcullen, R., & Iso-ahola, S. (2009). Effects of Psychosocial Resources on Elite Soldiers' Completion of a Demanding Military Selection Program. *Military Psychology*, *21*, 427–444. https://doi.org/10.1080/08995600903206354

Haberman, S. (1999,). *Haberman's Survival*. https://doi.org/10.24432/C5XK51

Haff, G. G., & Triplett, N. T. (2015). *Essentials of strength training and conditioning 4th edition*. Human kinetics.

Hanson, E. (2023,). *[ANN] SIRUS.jl v1.2: Interpretable Machine Learning via Rule Extraction*. https://discourse.julialang.org/t/ann-sirus-jl-v1-2-interpretable-machine-learning-via-rule-extraction/100932/3

Harrison, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, *5*(1), 81–102. https://doi.org/10.1016/0095-0696(78)90006-2

Hartmann, E., & Grønnerød, C. (2009). Rorschach variables and Big Five scales as predictors of military training completion: a replication study of the selection of candidates to the naval special

forces in Norway. *Journal of Personality Assessment*, *91*(3), 254–264. https://doi.org/10.1080/00223890902794309

Hartmann, E., Sunde, T., Kristensen, W., & Martinussen, M. (2003). Psychological measures as predictors of military training performance. *Journal of Personality Assessment*, *80*(1), 87–98. https://doi.org/10.1207/S15327752JPA8001_17

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). Springer. https://doi.org/10.1007/978-0-387-21606-5

Hill, Y., Kiefer, A. W., Oudejans, R. R. D., Baetzner, A. S., & Den Hartigh, R. J. (2024). Adaptation to stressors: Hormesis as a framework for human performance. *New Ideas in Psychology*, *73*, 101073. https://doi.org/10.1016/j.newideapsych.2024.101073

Hoekstra, H., & De Fruyt, F. (2014). *NEO-PI-3 en NEO-FFI-3: persoonlijkheidsvragenlijsten: handleiding*. Hogrefe.

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M. J., Vazire, S., Vespignani, A., & Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature*, *595*(7866), 181–188. https://doi.org/10.1038/s41586-021-03659-0

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. John Wiley & Sons. https://doi.org/10.1002/9781118548387

Housel, M. (2023). *Same as Ever: A guide to what never changes*. Penguin Publishing Group.

Huijzer, R., Blaauw, F. J., & Den Hartigh, R. J. R. (2023b). SIRUS.jl: Interpretable Machine Learning via Rule Extraction. *Journal of Open Source Software*, *8*(90), 5786. https://doi.org/10.21105/joss.05786

Huijzer, R., De Jonge, P., Blaauw, F. J., Baatenburg de Jong, M., De Wit, A., & Den Hartigh, R. J. R. (2023). Predicting Special Forces Dropout via Explainable Machine Learning. *PsyArXiv*. https://doi.org/10.31234/osf.io/s6j3r

Huijzer, R., Jeronimus, B. F., Reehoorn, A., Blaauw, F. J., Baatenburg de Jong, M., De Jonge, P., & Den Hartigh, R. J. R. (2022a). Personality traits of special forces operators: Comparing commandos, candidates, and controls. *Sport, Exercise, And Performance Psychology*, *11*(3), 369. https://doi.org/10.1037/spy0000296

Hunt, A. P., Billing, D. C., & Orr, R. M. (2011,). Predicting success on a special forces selection course: Identifying entry tests and establishing standards. *Defence Human Sciences Symposium 2011*.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.

Innes, M. (2018). Flux: Elegant machine learning with Julia. *Journal of Open Source Software*, *3*(25), 602. https://doi.org/10.21105/joss.00602

Jackson, J. J., Thoemmes, F., Jonkmann, K., Lüdtke, O., & Trautwein, U. (2012). Military training and personality trait development: Does the military make the man, or does the man make the military?. *Psychological Science*, *23*(3), 270–277. https://doi.org/10.1177/09567976114235

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2013). *An introduction to statistical learning* (Vol. 112). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Jauhiainen, S., Kauppi, J.-P., Krosshaug, T., Bahr, R., Bartsch, J., & Äyrämö, S. (2022). Predicting ACL injury using machine learning on data from an extensive screening test battery of 880 female elite athletes. *The American Journal of Sports Medicine*, *50*(11), 2917–2924. https://doi.org/10.1177/03635465221112095

John, O. P., Robins, R. W., & Pervin, L. A. (2010). *Handbook of personality: Theory and research*. Guilford Press.

Johnsen, B. H., Bartone, P., Sandvik, A. M., Gjeldnes, R., Morken, A. M., Hystad, S. W., & Stornæs, A. V. (2013). Psychological Hardiness Predicts Success in a Norwegian Armed Forces Border Patrol Selection Course. *International Journal of Selection and Assessment*, *21*(4), 368–375. https://doi.org/10.1111/ijsa.12046

Jones, G. R., & George, J. M. (1998). The experience and evolution of trust: Implications for cooperation and teamwork. *Academy of Management Review*, *23*(3), 531–546. https://doi.org/10.5465/amr.1998.926625

Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. W. (2002). Personality and leadership: a qualitative and quantitative review. *Journal of Applied Psychology*, *87*(4), 765. https://doi.org/10.1037/0021-9010.87.4.765

Kaplan, S., & McFall, R. (1951). The statistical properties of noise applied to radar range performance. *Proceedings of the IRE*, *39*(1), 56–60. https://doi.org/10.1109/JRPROC.1951.230422

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*.

Keinan, G., Meir, E., & Gome-Nemirovsky, T. (1984). Measurement of risk takers' personality. *Psychological Reports*, *55*(1), 163–167. https://doi.org/10.2466/pr0.1984.55.1.163

Kellmann, M., & Kölling, S. (2019). *Recovery and Stress in Sport: A manual for testing and assessment*. Routledge. https://doi.org/10.4324/9780429423857

King, G., & Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis*, *27*(4), 435–454. https://doi.org/10.1017/pan.2019.11

King, R. E., Carretta, T. R., Retzlaff, P., Barto, E., Ree, M. J., & Teachout, M. S. (2013). Standard cognitive psychological tests predict military pilot training outcomes. *Aviation Psychology and Applied Human Factors*. https://doi.org/10.1027/2192-0923/a000040

Kuncel, N. R., & Borneman, M. J. (2007). Toward a new method of detecting deliberately faked personality tests: The use of idiosyncratic item responses. *International Journal of Selection and Assessment*, *15*(2), 220–231. https://doi.org/10.1111/j.1468-2389.2007.00383.x

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., & others. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171. https://doi.org/10.1038/s41562-018-0311-x

Larsen, R. J., Buss, D. M., Wismeijer, A., Song, J., & Van den Berg, S. (2020). *Personality psychology: Domains of knowledge about human nature*.

Lazarus, R. S., & Folkman, S. (1984). *Stress, appraisal, and coping*. Springer publishing company.

Lee, J. E., McCreary, D. R., & Villeneuve, M. (2011). Prospective multifactorial analysis of Canadian Forces basic training attrition. *Military Medicine*, *176*(7), 777–784. https://doi.org/10.7205/MILMED-D-10-00375

Lobianco, A. (2021). BetaML: The Beta Machine Learning Toolkit, a self-contained repository of Machine Learning algorithms in Julia. *Journal of Open Source Software*, *6*(60), 2849. https://doi.org/10.21105/joss.02849

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Lövdal, S. S., Den Hartigh, R. J. R., & Azzopardi, G. (2021). Injury prediction in competitive runners with machine learning. *International Journal of Sports Physiology and Performance*, *16*(10), 1522–1531. https://doi.org/10.1123/ijspp.2020-0518

Marcus, B., & Schütz, A. (2005). Who are the people reluctant to participate in research? Personality correlates of four different types of nonresponse as inferred from self-and observer ratings. *Journal of Personality*, *73*(4), 959–984. https://doi.org/10.1111/j.1467-6494.2005.00335.x

Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, *10*, 85–110. https://doi.org/10.1146/annurev-clinpsy-032813-153700

Mattila, V. M., Tallroth, K., Marttinen, M., & Pihlajamäki, H. (2007). Physical fitness and performance. Body composition by DEXA and its association with physical fitness in 140 conscripts. *Medicine and Science in Sports and Exercise*, *39*(12), 2242–2247. https://doi.org/10.1249/mss.0b013e318155a813

McCrae, R. R., Costa, P. T., Jr, & Martin, T. A. (2005). The NEO-PI-3: A more readable revised NEO personality inventory. *Journal of Personality Assessment*, *84*(3), 261–270. https://doi.org/10.1207/s15327752jpa8403_05

McDonald, D., Norton, J., & Hodgdon, J. (1990). Training success in US Navy special forces. *Aviation, Space, And Environmental Medicine*, *61*(6), 548–554.

McElreath, R. (2020a,). *Science as Amateur Software Development*. YouTube. https://www.youtube.com/watch?v=zwRdO9_GGhY

McElreath, R. (2020b). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman, Hall/CRC. https://doi.org/10.1201/9780429029608

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, *73*(sup1), 235–245. https://doi.org/10.1080/00031305.2018.1527253

Menaspa, P., Sassi, A., & Impellizzeri, F. M. (2010). Aerobic fitness variables do not predict the professional career of young cyclists. *Medicine and Science in Sports and Exercise*, *42*(4), 805–812. https://doi.org/10.1249/mss.0b013e3181ba99bc

Metz, C. (2018). *A.I. Researchers Are Making More Than $1 Million, Even at a Nonprofit*. https://www.nytimes.com/2018/04/19/technology/artificial-intelligence-salaries-openai.html

Miller, C. (2022). *Chip war: the fight for the world's most critical technology*. Simon, Schuster.

Molnar, C. (2022). *Interpretable machine learning*.

Neale, B. T. (1985,). *CH - The first operational radar*. The GEC Journal of Research. http://www.radarpages.co.uk/mob/ch/chainhome.htm

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 535–569. https://doi.org/10.1080/10705510701575396

O'Boyle Jr, E., & Aguinis, H. (2012). The best and the rest: Revisiting the norm of normality of individual performance. *Personnel Psychology*, *65*(1), 79–119. https://doi.org/j.1744-6570.2011.01239.x

O'Toole, B. I. (1990). Intelligence and behaviour and motor vehicle accident mortality. *Accident Analysis & Prevention*, *22*(3), 211–221. https://doi.org/10.1016/0001-4575(90)90013-B

Oberski, D. (2016). Mixture models: Latent profile and latent class analysis. *Modern Statistical Methods for HCI*, 275–287. https://doi.org/10.1007/978-3-319-26633-6_12

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology*, *60*(4), 995–1027. https://doi.org/10.1111/j.1744-6570.2007.00099.x

Parkinson, B. W., & Gilbert, S. W. (1983). NAVSTAR: Global positioning system—Ten years later. *Proceedings of the IEEE*, *71*(10), 1177–1186. https://doi.org/10.1109/PROC.1983.12745

Pattyn, N., Van Cutsem, J., Van Puyvelde, M., Van den Berg, N., Lacroix, E., Dessy, E., Verheyden, C., Huybens, W., Lo Bue, S., Tibax, V., Vliegen, R., Ceccaldi, J., Savieri, P., Stas, L., & Mairesse, O. (2024). Smart is the new strong: An investigation of the contribution of physical, cognitive, anthropometric, and personality variables to success in a Tier 1 special forces qualification course. *Sport, Exercise, And Performance Psychology*. https://doi.org/0.1037/spy0000336

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, *49*(12), 1373–1379. https://doi.org/10.1016/S0895-4356(96)00236-3

Pelletier, L. G., Rocchi, M. A., Vallerand, R. J., Deci, E. L., & Ryan, R. M. (2013). Validation of the revised sport motivation scale (SMS-II). *Psychology of Sport and Exercise*, *14*(3), 329–341. https://doi.org/10.1016/j.psychsport.2012.12.002

Picano, J. J., Roland, R. R., Rollins, K. D., & Williams, T. J. (2002). Development and validation of a sentence completion test measure of defensive responding in military personnel assessed for nonroutine missions. *Military Psychology*, *14*(4), 279–298. https://doi.org/10.1207/S15327876MP1404_4

Raymaekers, J., Rousseeuw, P. J., Verdonck, T., & Yao, R. (2023). Fast Linear Model Trees by PILOT. *Arxiv Preprint Arxiv:2302.03931*. https://doi.org/10.48550/arXiv.2302.03931

Revelle, W., & Revelle, M. W. (2015). Package 'psych'. *The Comprehensive R Archive Network*, *337*(338).

Roland, A., & Shiman, P. (2002). *Strategic computing: DARPA and the quest for machine intelligence, 1983-1993*. MIT Press.

Rolland, J. P., Parker, W. D., & Stumpf, H. (1998). A psychometric examination of the French translations of NEO-PI-R and NEO-FFI. *Journal of Personality Assessment*, *71*(2), 269–291. https://doi.org/10.1207/s15327752jpa7102_13

Sadeghi, B., Chiarawongse, P., Squire, K., Jones, D. C., Noack, A., St-Jean, C., Huijzer, R., Schätzle, R., Butterworth, I., Peng, Y.-F., & Blaom, A. (2022). *DecisionTree.jl - A Julia implementation of the CART Decision Tree and Random Forest algorithms*. Zenodo. https://doi.org/10.5281/zenodo.7359268

Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, *8*, 597–599. https://doi.org/10.22237/jmasm/1257035100

Saxon, L., DiPaula, B., Fox, G. R., Ebert, R., Duhaime, J., Nocera, L., Tran, L., & Sobhani, M. (2020). Continuous measurement of reconnaissance marines in training with custom smartphone app and watch: observational cohort study. *JMIR Mhealth and Uhealth*, *8*(6), e14116. https://doi.org/10.2196/14116

Simon, H. A. (1979). Rational decision making in business organizations. *The American Economic Review*, *69*(4), 493–513. https://www.jstor.org/stable/1808698

Smith, B. W., Dalen, J., Wiggins, K., Tooley, E., Christopher, P., & Bernard, J. (2008). The brief resilience scale: assessing the ability to bounce back. *International Journal of Behavioral Medicine*, *15*, 194–200. https://doi.org/10.1080/10705500802222972

Smith, J. W., Everhart, J. E., Dickson, W., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 261.

Song, J., Howe, E., Oltmanns, J. R., & Fisher, A. J. (2023). Examining the concurrent and predictive validity of single items in ecological momentary assessments. *Assessment*, *30*(5), 1662–1671. https://doi.org/10.1177/10731911221113563

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on Bayes Factor Design Analysis using an informed prior. *Behavior Research Methods*, *51*, 1042–1058. https://doi.org/10.3758/s13428-018-01189-8

Stewart, R. D., Mõttus, R., Seeboth, A., Soto, C. J., & Johnson, W. (2022). The finer details? The predictability of life outcomes from Big Five domains, facets, and nuances. *Journal of Personality*, *90*(2), 167–182. https://doi.org/10.1111/jopy.12660

Stewart, V. (2017). 'Commando Consciousness' and Criminality in Post-Second World War Fiction. *Journal of War & Culture Studies*, *10*(2), 165–177. https://doi.org/10.1080/17526272.2016.1215683

Sørlie, H. O., Hetland, J., Dysvik, A., Fosse, T. H., & Martinsen, Ø. L. (2020). Person-Organization Fit in a military selection context. *Military Psychology*, *32*(3), 237–246. https://doi.org/10.1080/08995605.2020.1724752

Tait, J. L., Drain, J. R., Bulmer, S., Gastin, P. B., & Main, L. C. (2022). Factors predicting training delays and attrition of recruits during basic military training. *International Journal of Environmental Research and Public Health*, *19*(12), 7271. https://doi.org/10.3390/ijerph19127271

Taleb, N. N. (2010). *The black swan: The impact of the highly improbable*. Random House Publishing Group.

Taleb, N. N. (2013). *Antifragile: Things that Gain from Disorder*. Random House Publishing Group.

Taleb, N. N. (2020). Statistical consequences of fat tails: Real world preasymptotics, epistemology, and applications. *Arxiv Preprint Arxiv:2001.10488*. https://doi.org/10.48550/arXiv.2001.10488

Tedeholm, P. G., Larsson, A. C., & Sjöberg, A. (2023). *Predictors in the Swedish Counterterrorism Intervention Unit selection Process*. *8*(1), 3. https://doi.org/10.16993/sjwop.194

Tedeholm, P. G., Sjöberg, A., & Larsson, A. C. (2021). Personality traits among Swedish counterterrorism intervention unit police officers: A comparison with the general population. *Personality and Individual Differences*, *168*, 110411. https://doi.org/10.1016/j.paid.2020.110411

Tendeiro, J. N., & Kiers, H. A. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods*, *24*(6), 774. https://doi.org/10.1037/met0000221

Terman, L. M. (1918). The use of intelligence tests in the army. *Psychological Bulletin*, *15*(6), 177. https://doi.org/10.1037/h0071532

Turing, A. M. (1950). I.—Computing Machinery and Intelligence. *Mind*, *59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Vaara, J. P., Eränen, L., Ojanen, T., Pihlainen, K., Nykänen, T., Kallinen, K., Heikkinen, R., & Kyröläinen, H. (2020). Can physiological and psychological factors predict dropout from intense 10-day winter military survival training?. *International Journal of Environmental Research and Public Health*, *17*(23), 9064. https://doi.org/10.3390/ijerph17239064

Vaara, J. P., Groeller, H., Drain, J., Kyröläinen, H., Pihlainen, K., Ojanen, T., Connaboy, C., Santtila, M., Agostinelli, P., & Nindl, B. C. (2022). Physical training considerations for optimizing performance in essential military tasks. *European Journal of Sport Science*, *22*(1), 43–57. https://doi.org/10.1080/17461391.2021.1930193

Vallerand, R. J. (1989). Vers une méthodologie de validation trans-culturelle de questionnaires psychologiques: Implications pour la recherche en langue française. *Canadian Psychology/psychologie Canadienne*, *30*(4), 662. https://doi.org/10.1037/h0079856

Van der Krieke, L., Jeronimus, B. F., Blaauw, F. J., Wanders, R. B., Emerencia, A. C., Schenk, H. M., De Vos, S., Snippe, E., Wichers, M., Wigman, J. T., Bos, E. H., Wardenaar, K. J., & De Jonge, P. (2016). HowNutsAreTheDutch (HoeGekIsNL): A crowdsourcing study of mental symptoms and strengths. *International Journal of Methods in Psychiatric Research*, *25*(2), 123–144. https://doi.org/10.1002/mpr.1495

Van der Linden, D., Nijenhuis, J., & Bakker, A. (2010). The General Factor of Personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality*, *44*, 315–327. https://doi.org/10.1016/J.JRP.2010.03.003

Van Yperen, N. W. (2009). Why some make it and others do not: Identifying psychological factors that predict career success in professional adult soccer. *The Sport Psychologist*, *23*(3), 317–329. https://doi.org/10.1123/tsp.23.3.317

Van Yperen, N. W., Rietzschel, E. F., & De Jonge, K. M. (2014). Blended working: For whom it may (not) work. *Plos One*, *9*(7), e102921. https://doi.org/10.1371/journal.pone.0102921

Wanders, R., Loo, H. M., Vermunt, J. K., Meijer, R. R., Hartman, C. A., Schoevers, R., Wardenaar, K., & Jonge, P. de. (2016). Casting wider nets for anxiety and depression: disability-driven cross-

diagnostic subtypes in a large cohort. *Psychological Medicine*, *46*, 3371–3382. https://doi.org/10.1017/S0033291716002221

Williams, A. M., & Reilly, T. (2000). Talent identification and development in soccer. *Journal of Sports Sciences*, *18*(9), 657–667. https://doi.org/10.1080/02640410050120041

Wolberg, W., Mangasarian, O., Street, N., & Street, W. (1995). *Breast Cancer Wisconsin (Diagnostic)*. https://doi.org/10.24432/C5DW2B

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82. https://doi.org/10.1109/4235.585893

Woodworth, R. (1918). *Woodworth Psychoneurotic Inventory*. https://doi.org/10.1037/t01166-000

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Yu, B. (2013). Stability. *Bernoulli*, *19*(4), 1484–1500. https://doi.org/10.3150/13-BEJSP14

Yu, B. (2020). Veridical data science. *Proceedings of the 13th International Conference on Web Search and Data Mining*, 4–5. https://doi.org/10.1073/pnas.1901326117

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x