The Capability of Machine Learning for Predicting Disability Probabilities

Based on long-term absenteeism

Master thesis
Industrial Engineering and Management
Sandra Motamedi Nia
July 2021

UNIVERSITY OF TWENTE.

Predicting disability probabilities

based on long-term absenteeism

Author
Sandra Motamedi Nia
Student number: s1738968
Master: Industrial Engineering and Management
Specialization: Financial Engineering
Additional specialization: Business Administration

Educational Institution
University of Twente
Drienerlolaan 5
7522 NB Enschede
The Netherlands

Host company
Achmea
Laan van Malkenschoten 20
7333NP Apeldoorn
The Netherlands

First supervisor
Dr. A. Abhishta
Faculty Behavioral Management and Social Sciences

Second supervisor
Dr. ir. G.C. van Capelleveen
Faculty Behavioral Management and Social Sciences

External supervisor
Dr. R. Germs
Actuary Income Insurance
Business Finance Income Team

achmea

UNIVERSITY OF TWENTE.

Acknowledgments

This report finalizes my master Industrial Engineering and Management at the University of Twente. Despite the fact of performing the research from home due to the COVID-19 restrictions, I have gained valuable knowledge and experiences.

Looking back on the past years, I feel extremely grateful for all the personal growth, and most of all, the inspiring people I have met along the way. I deeply realize that I could not have achieved this significant milestone in my life without the support of the people around me. Therefore I want to use this opportunity to express my gratitude to everyone who supported me throughout this process.

First of all, my sincere acknowledgments go to my supervisors of the University of Twente, Dr. Abhishta, and Dr. ir. van Capelleveen. I want to thank my supervisors for the critical comments necessary on multiple occasions to make my research and thesis fulfill the academic standards. In face of the current pandemic, it was still possible to help me through the process by providing me constructive feedback via pleasant video calls. I want to thank them for their patience, guidance, and valuable remarks throughout my entire master trajectory.

I am very grateful for the opportunity to perform the research at Achmea, especially in these extraordinary times. During my time at Achmea, many people have helped me improve my understanding of the insurance industry and the possibilities and difficulties of using data within a large organization. I would therefore like to thank my colleagues at Achmea, in particular the Business Finance Team Income department, for the pleasant internship period. From the start, I have been regarded as a fully-fledged colleague and have been actively involved in the organization.

In particular, I want to sincerely thank my in-company supervisor, Dr. Germs, for the support and guidance during my internship at Achmea. During our meetings, I received essential information which I needed to execute my research. I am grateful for the professional guidance throughout the entire research at Achmea. Dr. Germs introduced me to the field of actuarial sciences and devoted so much of his time to this study. I also want to thank Mr. Trimp and Mr. Hubers for their involvement and support in the research. Their valuable feedback brought this research to a higher level.

Finishing my studies required more than academic support alone. I would therefore like to sincerely thank my mother and my fiancé for their unconditional support and encouragement. They helped me through ups and downs during my studies.

Management Summary

Machine learning (ML) refers to a group of techniques used by data scientists that allow computers to learn from data [1]. It is applied all around us, for example in segmenting customers by supermarket chains, placing targeted online advertisements by Facebook, and determining energy needs by suppliers. In the coming years, a stronger focus is expected on the application of machine learning within the insurance industry since the insurance business is experiencing exponential growth of data. Examples include applications for improving fraud detection, predicting customer questions, or estimating the extent of the damage. Another application that is still in development is the application of machine learning within the insurance industry to predict the probability of a claim occurrence.

In the Netherlands, employers insure their employees against the risk of incapacity for work with the WIA insurance. WIA is the Dutch abbreviation for the Work and Income (Capacity for Work) Act. Like any insurance, a premium is paid by the insured of the WIA. One of the key factors of determining the premium is the expected probability that a policyholder will become incapacitated for work for at least two years, also known as a WIA-influx (disability). To be eligible for the WIA benefit in case of disability, one of the policy conditions is the 42nd-week sickness report. If an employee is incapacitated for work for 42 weeks, an employer must report, according to the policy conditions, to the insurer. However, the data of the 42nd-week (long-term absenteeism) sickness report is not yet used within the determination of the disability probability.

In this research, we examine the use of machine learning models together with the 42nd-week data within the disability prediction of the insurance company Achmea, which is one of the largest suppliers of financial services (mainly insurance) in the Netherlands. Improving the disability prediction will provide valuable information for both the pricing process and the determination of provisions within the WIA insurance product.

As the number of not disabled individuals versus disabled individuals (claims) is disproportionate, the data is extremely imbalanced which might pose a problem in the accuracy of the disability probability prediction since machine learning models assume balanced datasets. Hence, due to the highly imbalanced dataset between the number of disabled versus not disabled policyholders, we examine re-balancing data methods and compare the results of the evaluation metrics. In this way, we examine whether re-balancing methods can improve class separation. However, both the accuracy and the predictive probability became less accurate. Hence, despite the fact both the recall and precision increased, we conclude that re-balancing the data results in inaccurate probabilities since the distribution of the data is changed.

In general, Generalized Models (GLMs) are used by pricing actuaries for predicting claim probabilities [2], also for the determination of the disability probability within the DII. However, in the last twenty years, data analytics have been developed, including machine learning models, which has risen the interest in the use of machine learning techniques to predict claim frequency. In contrast to GLM, these models do not assume a linear relationship but consider the data structure as unknown.

We explore selected algorithms from literature research. The models are selected based on certain requirements within the insurance industry such as transparency and interpretability which are key requirements to ensure explainability to all stakeholders.

Within the research, we merge and use three datasets of WIA policyholders of Achmea: WIA Policy-, Claim- and 42nd week reports data. The 42nd-week report data (long-term absenteeism) is available on an employee and contract basis, for the years 2017 and 2018. The dataset contains approximately 7000 records with 73 features. Since large numbers of features can cause poor performance for machine learning algorithms, we performed factor analysis. In this way, we reduce a large number of features into fewer factors.

To avoid overfitting during the development and validation of the models, 30% of the unique insured parties were randomly assigned to a test set for this study. The remaining 70% of the insured is the training set. All selected algorithms are trained exclusively using this training set. When developed each model within the training set with hyperparameter tuning and 5-fold stratified cross-validation. We compare the machine learning models with evaluation metrics.

By assessing the evaluation metrics accuracy, area under the ROC curve, and the brier score, we conclude that the machine learning models Logistic Regression (LR), Gradient Boosting (GB), and the Extreme Gradient Boosting (XGB) models achieve the highest results. The performance of these three ML models is comparable. When taking into account the applicability, LR is suggested since it is easy to implement and has a low computing time. Moreover, the Logistic Regression is familiar in the insurance industry since it is simply the GLM when describing it in terms of the logit link function. Moreover, according to Occam's Razor, we should prefer simpler models over complex models [3]. However, both the GB and XGB have a lot of flexibility due to the various hyperparameters which can be tuned and the XGB requires less exhaustive data pre-processing.

Concluding, in this research, we demonstrate by using experiments that machine learning models can make accurate predictions for claim occurrence. Hence, we conclude that machine learning techniques have potential added value within the insurance industry.

However, completely automating the prediction process does not seem sensible due to several factors. Therefore, we propose to develop a hybrid system in which actuaries are helped by suggestions from an ML model. This seems to be the most efficient way to combine the speed and quality benefits of an ML model with the actuaries' experience.

It has to be taken into account that the predictive model created within this research does not explicitly consider rare events such as the COVID-19 pandemic. Hence, in the case of rare events, the predictive model cannot be used directly. To indicate the trends and relationships which can be noticed based on the 42nd-week notifications, we have performed research on the impact of COVID-19.

Besides, the current disability probability depends on Age of the employee, Gender, Type of employment relationship, Salary, the sector of the company where the employee is employed. The additional features from the 42nd-week reports demonstrated a high predictive power in the feature importance analysis. In particular the features of the disability rate at the moment of being 42nd weeks ill and the expectation of becoming disabled indicated by the employer reveal an immense predictive power in comparison with the currently used features of the employee characteristics. This illustrates the added value of predicting disability based on long-term absenteeism data.

Contents

List of Tables

List of Figures

List of Acronyms

Abbreviation	Full form
AI	Artificial Intelligence
AOV	“Arbeidsongeschiktheidsverzekering”
AUC	Area Under the Curve
BFTI	Business Finance Team Inkomen
BI	Business Intelligence
CAO	“Collective ArbeidsOvereenkomst”
CV	Cross-Validation
DII	Disability Income Insurance
dS&I	“Divisie Schade & Inkomen”
DT	Decision Tree
GB	Gradient Boosting
GLM	Generalized Linear Model
IBNR	Incurred But Not Reported
IIF	“Instroom Inschalingsfactor”
IVA	“Inkomensvoorziening Volledig Arbeidsongeschikten”
KNN	K-Nearest Neighbour
LR	Linear Regression
ML	Machine Learning
ROC	Receiver Operating Characteristics
SMOTE	Synthetic Minority Over-Sampling Technique
UWV	“Uitvoeringsinstituut Werknemersverzekeringen”
WGA	“Werkhervatting Gedeeltelijk Arbeidsgeschikten”
WIA	“Wet Werk en Inkomen naar Arbeidsvermogen”
WULBZ	“Wet Uitbreiding Loondoorbetalingsverplichting Bij Ziekte”
XGB	Extreme Gradient Boosting

Terms and Definitions

Term	Definition
AOV	Insurance that ensures income for self-employed entrepreneurs in the event of incapacity of work.
Claimyear	The year in which the damage occurs or has taken place.
Client/Customer	In this research, the customer refers to the employer.
Exit product	In case of incapacity for the work of employees, the employer is obliged to continue to pay the wages of these employees for 2 years. After these 2 years, these employees receive a benefit from the UWV, for which the employer pays a premium to the tax authorities. Instead, the employer can choose to bear the risk itself. The premium paid to the Tax Authorities stops and from then on the employer is responsible for the payment of the wages to the employees and the recovery process. To cover this risk, the employer can take out an Employer's Exist Insurance at Achmea.
Group (Collective) Insurance	This is insurance for groups. For example, an employer takes out an insurance policy whereby several employees are ensured of continued payment of wages.
IBNR damage provision	These are provisions for claims that have occurred but have not yet been reported to the insurer.
Incapacity of work	An employee is incapacitated for work if he/she is unable or not fully able to perform his/her work. A distinction is made here between full and partial incapacity for work. This can be the result of illness or an accident.
IVA	Employees who are incapacitated for work for a long time and therefore have a very small chance of recovery. This risk is covered by the UWV.
Obligation to continue to pay wages	An employer is obliged to continue to pay wages for 2 years during the period that an employee is ill. In the years of the obligation to continue to pay wages, the employer is obliged to continue to pay at least 70% of the employee's gross wages.
Policy	Notarial deed in which the insurance contract is recorded in writing or digitally.
Rehabilitation opportunity	The probability that the insured person who is incapacitated for work recovers or that benefits will have to be paid for a long time.
Residual earning capacity	A percentage that indicates how much a sick employee can still earn after becoming ill.
UWV	An institute that ensures the professional and efficient implementation of employee insurance schemes and offers labor market and data services.
WGA	For employees who are between 35% and 80% incapacitated for work, but also for employees who are completely, but not long-term incapacitated for work.
WIA	Collective insurance that insures employers to be able to comply with the obligation to continue to pay wages after 2 years of illness and (partial) disability of employees.
WIA-Influx	The chance of disability, which is the chance that the insured person becomes incapacitated for work, and is not fully recovered 104 weeks after the start date of illness.

1. Introduction

1.1 Research objective

The main purpose of insurance companies is to offer protection against financial losses in exchange for a predetermined fixed premium which is set before the future costs are exposed [4]. Since the insurance market is highly competitive, the insurer must charge both a fair premium and at the same time, keep enough facilities to cover the expected loss of the policyholder [5]. The expected future losses are not the same for every policyholder. Therefore, to successfully predict both the premiums and provisions, insurers aim to cluster policyholders as optimal as possible per risk profile, also known as *risk classification* [6]. With risk classification, an insurance company is capable of inquiring distinctive prices to particular classes, which is crucial for both the solvency of the insurer and overcoming the high competition between insurance companies [2]. If one insurer applies risk classification to a particular variable and the other insurer does not, there may be adverse selection.

Generalized Linear Models (GLM) are the state-of-art analytic insurance model for both the claim frequency and claim amount model. The developments in the field of algorithmic models in the last twenty years and the increased computing power are causing an increasing interest in machine learning techniques in the actuarial world. However, only a few papers in the insurance literature go beyond the actuarial comfort zone of GLMs [4].

Machine learning is the computer science industry that studies algorithms performing a particular task without being explicitly programmed. ML makes use of statistical models and a data set also called the training data. On account of the better availability of powerful hardware and large data sets, we are experiencing explosive growth in applications of machine learning [7] [8].

With the upswing of data analytics, in this study, we focus on machine learning techniques to develop predictive models for claim probability, specific for the “Wet Werk en Inkomen naar Arbeidsvermogen” (WIA). The WIA is the collective insurance in the Netherlands that insured employers be able to comply with the obligation paying wages after two years of illness to their employees. Since the WIA claim probability of an employee is very low, it is a complex process to set a premium for this insurance product.

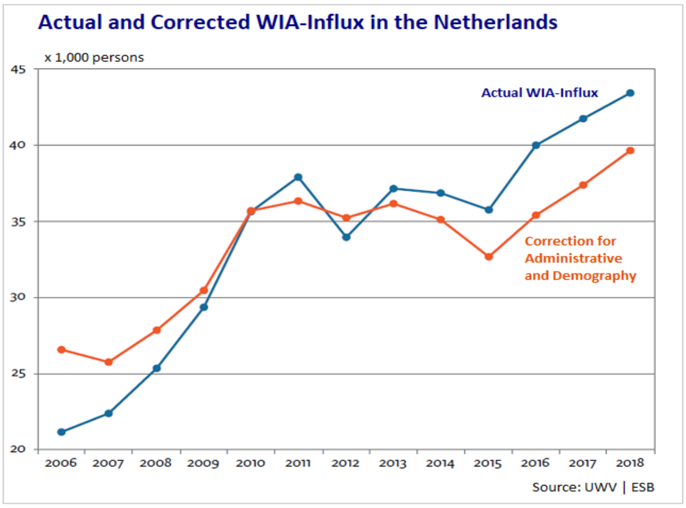
1.2 Company introduction

We conduct this research at Achmea in Apeldoorn on behalf of the Business Finance Team Income (BFTI) department of the Non-Life and Income division. Achmea is the largest insurance company in the Netherlands. Within the department Non-Life and Income, insurance products are managed and further developed for various brands: Centraal Beheer, Interpolis, Avéro, and Zilveren Kruis. The purpose of this department is to protect customers as well as possible against the risk of disability with Income insurance policies.

1.3 Problem Identification

Every year an average of 27,000 people become incapacitated for work in the Netherlands [9]. The chance is approximately 1 in 12 to become incapacitated for work [10]. To be eligible for coverage in case of incapacity for work, certain policy conditions must be met. One of the conditions is submitting a 42nd-week report. As soon as an employee is incapacitated for work for 42 weeks (long-term absenteeism), the employer, who took out the WIA insurance for the employee, must report the long-term absenteeism to both the "Uitvoeringsinstituut Werknemersverzekeringen (UWV) and the Disability Income Insurance (DII). The registration of the 42nd weekly reports is not primarily intended to predict the transition to WIA-influx (sickness for at least 104 weeks) and therefore not yet used for this purpose.

Recent studies from the national institute for health and disability Insurance inform that the number of disabled employees (more than 2 years incapacitated for work due to illness) is rising [11], Figure 1-1: Actual and Corrected WIA influx (9 May 2019). The increase in the number of disabled employees gave rise to the use of the 42nd weekly reports for the earlier and more accurate prediction of disability probabilities.



Source: UWV | ESB

Figure 1-1: Actual and Corrected WIA influx (9 May 2019)

Currently, the prediction of disability probabilities is not as accurate as desired by Achmea and the 42nd-week report's data are not yet used for this prediction. For the WIA insurance products, there are almost 2 years (104 weeks) between the moment an insured event (illness) occurs and the moment a disability occurs with possible entitlement to benefits.

- For pricing purposes, this means that Achmea has to wait at least two years before they can evaluate their prediction (expected WIA inflow probability) with the realized inflow.
- For provision purposes, this means Achmea has to wait at least 2 years before they can base the provisions for a specific claim year on the actual income WIA occasions.

An as short as possible period of accurate prediction of the WIA influx is essential since it has an impact on the expected damage burden which is important for:

- *Pricing*: WIA premium that is asked from customers.
  Customer premium = Expected claims expense + Costs + Commercial margins and discounts
- *Provisions* of Achmea: Determination of money that Achmea needs to reserve now to be able to pay future benefits (claims)

The consequence of not having accurate predictions of the WIA influx after 104 weeks is that for Pricing, the price will be too high or too low. Also, for the facilities, too much or too little money will be reserved. An uncertain premium and provision have a (negative) financial impact on both the insurer and the insured. Therefore, the problem is that within Achmea the period of accurate predictions of the WIA influx is too long. This problem might be reduced if *Achmea has an accurate prediction model of disability probabilities based on existing data and the added data of long-term absenteeism events.*

Within the insurance business, Generalized Linear Models are used as pricing models but more advanced techniques are not yet widely used [12]. Using the 42nd-week reports together with the implementation of machine learning techniques might improve the accuracy of the disability prediction. Therefore, the aim of this research is a broad exploration of the value of 42nd-week reports data and machine learning applications for the WIA influx. We put focus on statistical performance, interpretation, and business implications.

1.4 Academic relevance

The interest of Achmea to introduce machine learning techniques to improve the prediction of the disability probability might provide the necessary motivation to start with more extensive and differentiated use of advanced models within the insurance sector. The application of machine learning techniques in general within the insurance sector is not new. However, the developments are still in the initial phase and there is still a need for a lot of experimentation and evaluation, in particular to WIA influx prediction, of which fewer is known.

1.5 Research Method

We carry this research out by using the Design Science Research Methodology (DSRM) for Information Systems (IS) [13]. DSRM is a research methodology that focuses on creating and evaluating innovative artifacts. These artifacts apply further knowledge to the production of information systems for management and organizational purposes [14]. The objective of design-science research is to develop technology-based solutions to important and relevant business problems.

Design Science is a set of synthetic and analytical techniques specified for information systems that have developed the DSRM methodology [13]. Design Science involves two primary activities:

1. Creating new knowledge by designing new and innovative things and processes (artifacts).
2. The analysis of the use and performance of artifacts.

An artifact must solve a relevant and important problem. Besides, during Design Science research an attempt is made to develop something new at all times [15]. Design Science creates and evaluates IT artifacts designed to solve business problems.

In this research, the planned deliverable to determine the added value of the 42nd-week reports in the prediction of WIA influx is developing a predictive model based on machine learning techniques. Many research methodologies rely on explanatory or descriptive research. However, Design Science Research tries to solve a problem by designing an artifact.

Developing an IT artifact is also a relatively new subject, on which most of the research methodologies have not yet been adopted. For these reasons, Design Science Research is considered a suitable methodology for this study [16]. Figure 1-2 below illustrates the process model of the DSRM framework.

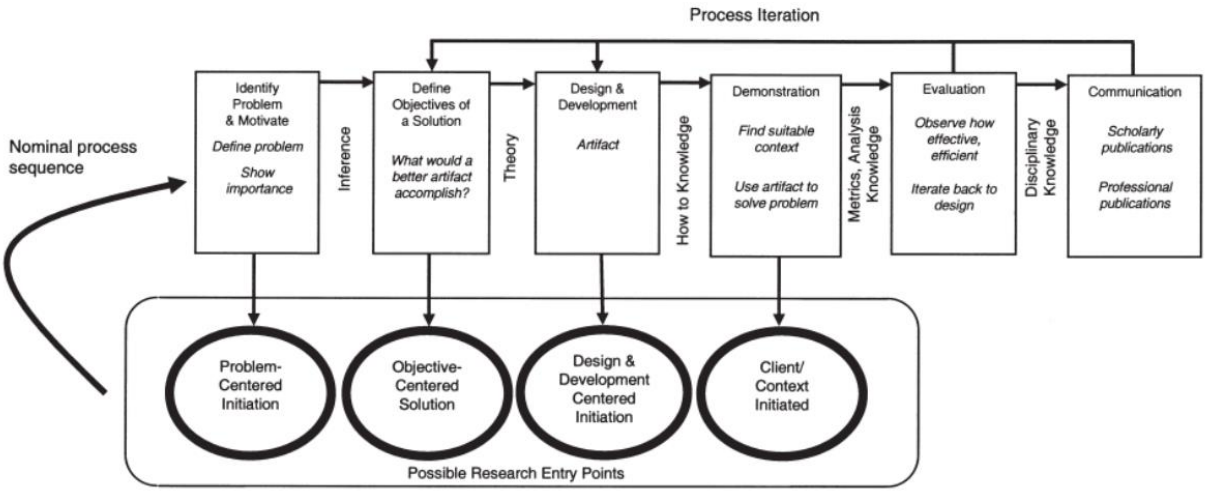

Process Iteration
Nominal process sequence
Identify Problem & Motivate
Define problem
Show importance
Inference
Define Objectives of a Solution
What would a better artifact accomplish?
Theory
Design & Development
Artifact
How to Knowledge
Demonstration
Find suitable context
Use artifact to solve problem
Metrics, Analysis Knowledge
Evaluation
Observe how effective, efficient
Iterate back to design
Disciplinary Knowledge
Communication
Scholarly publications
Professional publications
Problem-Centered Initiation
Objective-Centered Solution
Design & Development Centered Initiation
Client/ Context Initiated
Possible Research Entry Points

Figure 1-2: DSRM Process Model By Peffers

Design Science Research Methodology leads in this research framework to the following steps:

1. ***Problem identification and motivation***

The objective of the first phase within the DSRM is to develop a sound method that can effectively provide a solution and emphases the potential added value for the acceptance of the research results. By identifying the problem and determining its relevance, we focus on the problem identification and motivate why there is a need for the proposed solution. We describe this phase in *Chapter 1.*

2. ***Define Objectives of a Solution***

The research aims to develop a machine learning model which accurately predicts the disability probability based on long-term absenteeism. We give insight into how the aim of the research will be achieved using a developed research method. We elaborate the context to clarify the background and the application domain of the research in *Chapter 2*. In *Chapter 3* we perform literature research relating to applicable machine learning algorithms and performance measures. This provides support for the design process and helps to legitimize the research. In *Chapter 4* we perform an analysis of raw data and we prepare the data.

3. ***Design and Development***

The third phase in the DSRM encompasses the design and development of the artifact which we develop in *Chapter 5.* In this case, we design and develop a predictive model for the disability probability based on long-term absenteeism. We compare different algorithms and select based on this comparison the most accurate algorithms.

4. ***Demonstration***

The fourth phase in the DSRM methodology encompasses the demonstration step where the created solution will be demonstrated. In this phase, we demonstrate how the artifact can be used to solve the problem. The developed method is demonstrated in an experiment on the relevant data of the disability insurance within Achmea. The experiment shows the ability of the method to provide guidance and provides a real-life example giving insight to other practitioners. We comprehend the demonstration phase in *Chapter 6.*

5. ***Evaluation***

Evaluation is the fifth phase of the design science methodology for information systems. The assessment is essential to observe and to measure if and to what level, the designed method supports the problem solving that is defined in the first step of the design science methodology. We discuss the results of the artifact concerning the problem. The results of the experiment provide insight if the developed method is a successful machine learning project. Therefore, in C*hapter 7*, we discuss the problem, conclude the research, and provide recommendations.

6. **Communication**

In the final phase of the DSRM, we focus on communication. It is essential to communicate the design artifact to understand how much potential value the solution has and for the creation of transparency. The problem and its importance are communicated in *Chapter 7*. Communicating the contribution of the design method to existing methods enhances validity and effectiveness. We communicate this research twofold, through sharing this thesis on the University of Twente repository and via Presentations to both relevant stakeholders and during the Public Defense.

1.6 Research questions

The main research question is formulated as follows:

How can we utilize the 42nd- week reports data to increase the accuracy of the disability probabilities prediction with machine learning techniques?

To answer this main question, we split the research into three sub-questions, which are further divided into steps. An overview is given in Table 1-1.

Table 1-1: Overview Research questions

Research questions with corresponding steps
1. How to develop a forecasting model for determining the transition probability from long-term absenteeism (42 weeks illness) to disability (WIA-influx)?
A. Features available in the data
B. Machine learning algorithms applicable for predicting disability probabilities
C. Insights created by the forecasting model
D. Handling class imbalance
E. Machine learning algorithm with accurate disability probabilities
2. How can the outcomes of a forecasting model be used to create added value?
A. Metrics for model performance
B. Evaluation the performance
3. Which visualization of trends and relationships can be noticed based on the 42nd-week notifications?
A. Trends in long-term absenteeism during the COVID-19 period

1.7 Programming languages

In this master's thesis, we program both in Statistical Analysis System (SAS) and Python. We execute the merging of datasets in SAS. For the data pre-processing process and the implementation of machine learning techniques, we use Python. SAS is the name for software that can be used to, among others, analyze and report data. Python is a commonly used language for data analysis. The programming language has tons of modules that are publicly accessible to plot curves, calculate with matrices, perform statistical analyzes, and so on. Of course, there are still available languages, each with its advantages, such as MATLAB [17]. Python has the advantage that it contains the sci-kit learn library, which features all state-of-the-art machine learning algorithms.

1.8 Ethical framework

Collecting data is the basis of insurers' work to estimate risks and determine premiums. Only if the customer, regulator, and legislator have sufficient confidence in the correct use of data, insurers will be able to incorporate new technologies in their business processes in a future-proof manner. It may be that a particular data technique is legally permitted, but is contrary to the premise of the ethical framework. The ethical framework for data-driven decision-making requires insurers to perform several checks when using modern technologies such as artificial intelligence [18]. The framework is based on the recommendations of the High-Level Expert Group on Artificial Intelligence and emphasis the importance of data protection.

During this research, we perform several concrete actions to ensure data protection. First of all, the data file is protected with a password. Furthermore, the merged data is saved in a safe environment. Moreover, to ensure the sensitive data cannot be traced back to private individuals, we do not work with the citizen service number.

2. Context Analysis

In this chapter, we outline the context in which the research is conducted. First, we discuss the Disability Insurance system in the Netherlands to increase the understanding of WIA Insurance. Subsequently, we elaborate on the disability insurance products. Next off, we describe the current forecasting model. Finally, we discuss how analytics and a forecasting model can be used within the Disability Income Insurance.

2.1 The Dutch Disability Insurance

The focus of this research is on implementing advanced tools on Disability Income Insurance. The DII in the Netherlands is designed to protect policyholders in case of incapacity for work. When an employee becomes incapacitated for work, one can suffer from a significant loss of income [19].

Within the European Union, countries use different definitions of incapacity for work and apply different systems. Therefore, incapacity for work in the Netherlands cannot directly be compared with other countries in Europe. In the Netherlands, employees who are ill for a *maximum of two years* or who are (partially) incapacitated for work are assured of continued payments of salaries by the employer, who can cover this risk by the sickness absenteeism insurance.

Employees who have been ill for *at least two years* or who are (partially) incapacitated for work are assured of continued payments of wages by the Dutch legislation “Werk en Inkomen naar Arbeidsvermogen”(WIA), which is a Disability Income Insurance [20]. With this DII, the employee, after a disability period of two years, is insured to a maximum of 70% of the social insurance pay of € 58.307,40 in 2021 [21]. Like many other Dutch insurance companies, Achmea offers WIA insurance which covers the relapse in income for at least 70% after a disability to work occurs for more than two years.

Insurance against disability is one of the many types of insurances that are available offered by insurers. According to the Financial Supervision Act, which regulates the supervision of the financial sector in the Netherlands, disability insurance falls under the insurer's non-life sector. Non-life insurance in addition to disability insurance also includes insurance such as a car, fire, and home insurance. The characteristic of the latter “standard” non-life insurance policies is that they usually have a short-term (usually 1-year contract) and the payment in case of damage often concerns a one-off payment.

However, the disability insurance deviates from these “standard” non-life insurance policies because disability insurance is usually for a longer term and in the event of damage the payment is not a one-off, but the insured person receives a payment during the period that he or she is incapacitated for work. This benefit period ends when the final age at which the insurance is taken out has been reached or at the moment when the employee is recovered. The characteristics of occupational disability insurance are therefore more in line with a life insurance policy.

With life insurance, however, only mortality rates are included in the pricing and the determination of the technical provision. In the case of disability insurance, also opportunities for illness, disability, and rehabilitation are included. Where traditional life insurance only uses the probability that an insured person is alive, DII links the probability of being alive to the active or inactive state of the insured person. Active means that the insured person is fully or partially active in the labor market. The inactivity of the insured indicates that the insured is ill or disabled and is therefore no longer (fully) active on the labor market. In addition to the risk of death, a DII also takes into account the changes of disability and rehabilitation in the determination of the premium and valuation of the provision. Therefore, there are several situations in which an insured person may find himself which are: active, death, incapacity for work with a return, and incapacity for work without return.

On the occasion where the employer receives a benefit for the WGA from UWV, the premium is paid through the Tax Authorities. However, the employer can also choose to become a self-insurer. In that case, the contribution will no longer be paid through the Tax Authorities. Instead, the employer will be responsible for the WGA benefit and the reintegration of its employers for 10 years. In this way, the employer keeps control, but also runs a financial risk. This risk can be insured at disability insurances, such as Achmea.

2.1.2 Disability Insurance Products

In this section, we give an overview of the different disability insurance products within Achmea. With the Income Insurance policies, Achmea limits the financial consequences of incapacity for work for their customers. Disability Income Insurance is insurance for employees (WIA) or self-employed entrepreneurs (AOW) which provides benefits to the insured in the event of illness or disability, Figure 2-1. In this research, we focus on DII for employees after a waiting period of 2 years (WIA).

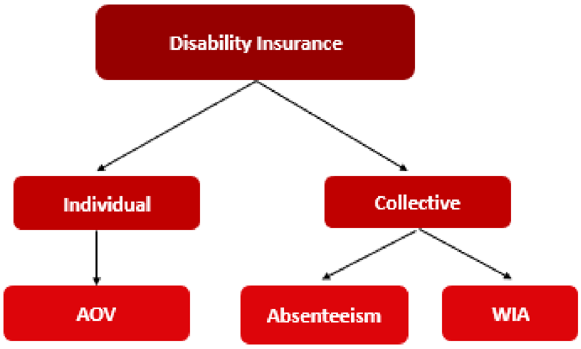


Figure 2-1: Overview Disability Income Insurances

In the first 2 years of illness, the Absenteeism insurance is active. According to the law, the employer has a wage payment obligation of 100% of the wage in the first year an employee is ill. However, in the second year of illness, the employee will get most often 70% of the wage. Notwithstanding, this is dependent on the comprise stated in the Collective Labour Agreement (CAO). During the first 2 years of illness, the employee receives a benefit from their employer which is mandatorily stated in the law “Wet Uitbreiding Loondoorbetalingsverplichting Bij Ziekte” (WULBZ) and is dependent on the last earned salary. After these 2 years, an inspection will take place by the UWV in which the percentage of an employee is incapacitated for work is determined.

Also, the residual earning capacity is determined which is the salary that the employee can still earn himself. After more than 2 years of illness, the disability insurance WIA becomes active and has a maximum of 10 years. The WIA consists of two different schemes: the WGA and the IVA. Depending on the outcome of the inspection which is performed after 2 years of illness, the employee gets classified in one of the two branches of the WIA.

1. WGA

The WGA ("Werkhervatting Gedeeltelijk Arbeidsgeschikten") is part of the WIA. It has two distinctions:

WGA-Partial: Employees who are between 35% and 80% incapacitated for work.

WGA-Entirely: Employees who are completely, but not long-term incapacitated for work. Therefore, employees who enter the WGA can still have a current salary. There is a distinction between employees who can learn more or less than 50% of their salary, also known as the "Restverdiencapaciteit" (RVC).

2. IVA

The IVA ("Inkomensvoorziening Volledig Arbeidsongeschikten") is the legislation for employees who are long-term incapacitated for work (at least 80%) and therefore have no or a very small chance of recovery. Also, employees who have been ill for more than 10 years will end up in the IVA.

Only in the WGA division, Achmea should pay the WIA benefit. In the case of the IVA, the obligation transfers to UVW for the WIA benefit. Employees who are less than 35% incapacitated for work, indicated as disability rate (AO) are not entitled to a statutory benefit. Figure 2-2 gives an overview of the disability process in the Netherlands.

The maximum benefit an employee can receive is coupled to the SV-wage, which is the salary of an employee over which taxes and social premiums are paid [22]. This SV-wage changes each year. In 2021 this was set at € 58.307,40 [23].

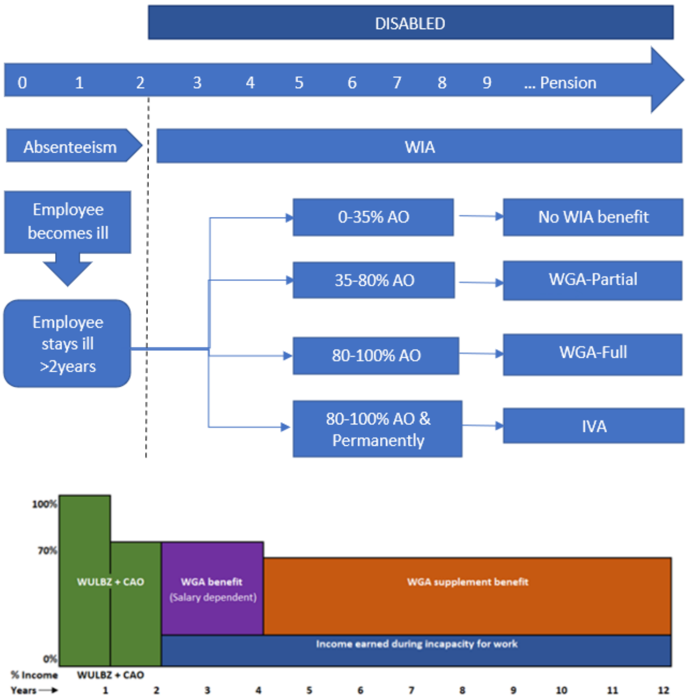


Figure 2-2: Dutch Disability Insurance Process

2.2 Disability Prediction Model

In this section, we elaborate on the current prediction method for the WIA Influx of Achmea. This is necessary to get a deep understanding of the core research question.

2.2.2 Association of insurance companies

An employee who becomes partially incapacitated for work is eligible for a benefit after two years under the Resumption of Work Scheme for Partially Disabled People (WGA). An employer is by law insured with UWV for this risk. An alternative for employers is to become a self-insurer (ERD-‘Eigenrisicodrager’), also known as the WGA-ERD. In this case, the employer chooses to remain responsible for employees and former employees for whom they are no longer obliged to pay wages in the event of illness. The employer will largely take over the role of UWV. The advantage for the employer is that he or she has more control over absenteeism and is responsible for the reintegration. The employer can ensure this risk at a disability income insurance such as Achmea by paying a WIA premium. This premium consists of a fixed basic premium and a differentiated premium, which depends on the risk. Insurers can only insure the WGA-ERD risk if they are enabled to make a proper assessment of the risk.

In this research, we focus on the disability probability (WIA-Influx). Currently, the WIA-influx prediction is based on the “Verbondsmodel”, as defined the “Verbond van Verzekeraars”. The “Verbond van Verzekeraars” is the association of insurance companies in the Netherlands [24]. Based on extensive data this association reports probability numbers of events happing for different common insurance products.

For the insurances that provide disability income insurance for WGA-ERD, the association of insurers provides a covenant model which is based on the data set provided by the UWV. To estimate the chances of disability, the Generalized Linear Model with Poisson distribution and log link function (Poisson model) is used by the association of insurers [25]. Thus, the WGA-ERD covenant model estimates the costs of the coverage. However, there are some uncertainties. Therefore, only considering the WGA-ERD model is not sufficient for the WGA-ERD coverage. Actuaries at Achmea perform calculations and evaluate financial-economic risks to increase the probability of having enough coverage.

The disability probability provides insights into the expected damage burden which is calculated with the following formula:

*Expected damage burden of an employee = chance of disability (inflow) * Expected benefit * Expected duration of benefit.*

The expected damage burden is essential for both pricing the premium and the required provisions.

1. *Pricing:* Determination of the customer premium for a specific WIA product.

   Customer premium = Expected claim burden + Costs + Commercial margins and discounts

2. *Provisions*: Determination of money that must be reserved now to be able to pay future benefits (claims)

2.2.3 Current forecasting method

Currently, the actuarial world is using the Generalized Linear Model (GLM) with Poisson distributed claim numbers. The GLM requires a deep understanding of the patterns in the data to make informed decisions, especially by actuaries. In the last twenty years, statistical techniques have been developed, including algorithmic models. In contrast to GLM, these models do not assume a linear relationship but consider the data structure as unknown.

2.3 42nd weekly reports

In this section, we elaborate on the 42nd weekly reports since these notifications are the key part of this research. No later than the first working day after an employee has been sick for 42 weeks, the employer must notify the UWV that the employee is still being ill. This is also included in the policy condition of Achmea. For the insurer, this notification is important for the timely deployment of targeted actions for resumption for work.

In addition, the 42nd-week reports give the insurer a good indication of the future WIA influx. Ideally, as an insurer, Achmea receives all 42nd-week reports who have taken out a WIA product. Taking into account the 42nd weekly reports within the predictions of the WIA influx can increase the accuracy of the expected WIA influx.

Within the DII of Achmea, the 42nd-week report's data is not yet used for actuarial models since the data is only recently available. Once a year, the Actuarial Department researches the principles of pricing and provisions with the data that is available at that time. For the 2017 rate, for example, damage data is used from 2010 to 2013. This means that the developments in the last three years cannot be included. With the use of 42nd-week reports, models can be made of developments in more recent years.

The registration of the 42nd-week reports can be used to estimate the number of long-term sick employees in the Netherlands and their recovery pattern. In the event of long-term sickness, we can locate where the increase or decline mainly occurs such as the age groups, which sectors, gender, and large or small companies. This information helps to explain the changes in the WIA influx. Conversely, the registration also helps to better predict how many employees will become a WIA-influx in the next calendar year. A side note remains that the registration of the 42nd-week reports is still incomplete.

2.4 Factors determining the disability probability

The current probability of disability is based on the following factors:

- Age
- Gender
- Type of employment relationship
- Salary

3. Theoretical Framework/ Literature Review

In this chapter, we discuss the literature studied for this research.

3.1 Machine Learning versus AI

The terms Artificial Intelligence (AI), Machine Learning (ML), and deep learning are often used interchangeably. As can be seen from Figure 3-1, the terms are closely related. The umbrella term is AI, where a computer is trained to mimic human intelligence. The idea is to train the computer such that it can perform actions the way a human would do, like creative thinking. The aim is to stimulate human intelligence as closely as possible.

Machine learning is a subset of AI, where machines/computers are trained to perform certain tasks without being explicitly programmed. ML is a type of Artificial Intelligence that aims to build systems that can learn from the processed data or use data to perform better. An important difference between AI and ML is that while ML is always under AI, AI is not always under machine learning. Finally, there is also deep learning, which uses artificial neural networks inspired by the human brain. Within this master's thesis, we focus on machine learning.

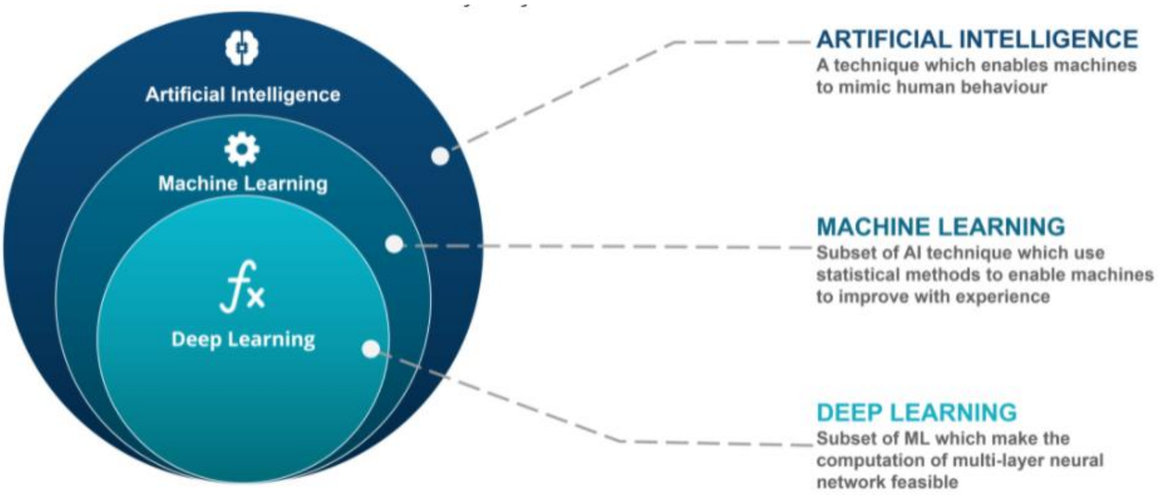

Artificial Intelligence
Machine Learning
fx
Deep Learning
ARTIFICIAL INTELLIGENCE
A technique which enables machines to mimic human behaviour
MACHINE LEARNING
Subset of AI technique which use statistical methods to enable machines to improve with experience
DEEP LEARNING
Subset of ML which make the computation of multi-layer neural network feasible

Figure 3-1: Machine Learning vs Artificial Intelligence [26]

3.2 Machine Learning in the actuarial world

As mentioned in Section 3.1, machine learning is the computer science industry studying algorithms that perform a particular task without being explicitly programmed. Instead, machine learning algorithms make use of statistical models and a dataset. As a result of the improved availability of powerful hardware and large datasets, we are experiencing explosive growth in applications of these techniques [7] [8].

Currently, the actuarial world is using most often traditional statistical methods (linear regression, GLM) for pricing which have an overlap with machine learning models. The current GLM model which is used is occasionally incorporated with time series forecasting [27].
The requirement for more advanced machine learning models is increasing. The main reasons are the large amounts of varying data and the exponential increase in computing power. Within the insurance industry, more advanced machine models have been applied [28]. However, the previously conducted researches within the actuarial world have the focus mainly on car insurance [29], customer retention [30] and claim fraud detection [31].

In the case of car insurance, the distribution is completely different in comparison to disability insurance. A claim in car insurance is much more likely to occur than invalidity as a percentage of the total policyholders, which leads to an extremely imbalanced dataset. Moreover, the machine learning techniques used in car insurance had the focus on the accuracy of the prediction rather than on the predicted probability of claim occurrence.

Within the actuarial world, the GLM is extensively used with Poisson distribution and log link function (Poisson model). The first use of GLM in DII was by Renshaw [32]. The Poisson distribution is used for the number of claims an individual policyholder reports during the insurance term. For this, the assumptions are that policies are independent of each other, that the time intervals of the policies are independent of each other, and that policies are homogeneous in a certain rate scale.

It follows that all individual claims are independent of each other and of time and for that reason, the number of claims for an individual policy follows the Poisson distribution. This applies not only to the number of individual claims but also to the number of claims for all policies in a certain risk class.

Since the insurance business is highly regulated, it poses some challenges to implement machine learning algorithms. Moreover, insurance companies are responsible to provide solidarity among policyholders. Therefore, extreme discrimination should be avoided [33]. There needs to be a proper trade-off between customer segmentation and risk pooling.

3.3 General procedure machine learning application

Before an algorithm is implemented, the data must be prepared. It essential to make the input data readable for the chosen algorithms. Therefore, the data is often represented numerically in data frames and arrays. Furthermore, preparing the data also means filling in empty values, detecting incorrect data, and detecting extreme outliers. The process of putting the data into a readable form for the algorithm is called data pre-processing.

According to the International Business Machines Corporation (IBM), a professional data analyst spends 80% of his time preparing the data. The remaining 20% of the time effectively is spent to train models and perform analyses [34].

Depending on the chosen algorithm, the data will undergo several further operations. The following steps displayed in Figure 3-2 are general and can change from situation to situation.

After the pre-processing, the data is split into a training and test set. A model is trained with the training data. The algorithm then tries, based on the training data to transform the inputs and characteristics to the output as well as possible.

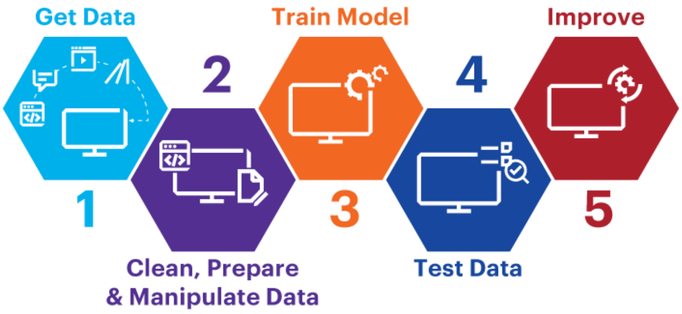


Figure 3-2: Machine Learning Operation [35]

3.4 Machine Learning Types

The applications of machine learning are divided into supervised learning, unsupervised learning, and reinforcement learning [36], Figure 3-3.

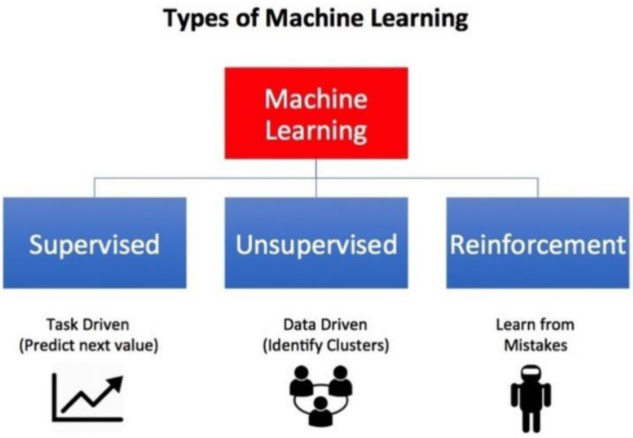

Types of Machine Learning
Machine Learning
Supervised
Unsupervised
Reinforcement
Task Driven (Predict next value)
Data Driven (Identify Clusters)
Learn from Mistakes

Figure 3-3: Machine Learning types [37]

3.4.1 Reinforcement learning

Reinforcement Learning addresses the question of how an autonomous agent can learn to choose the optimal actions to achieve its goal [38]. In this technique, the learning algorithm uses a system of “reward” and “punish”. When the algorithm performs a task well, it will receive a reward. Conversely, if the algorithm has not obtained the desired result, it will be penalized.

3.4.2 Unsupervised learning

Unsupervised learning is applied when unlabeled data is available. This means that learning is done without there being an associated baseline or an answer. Enormous amounts of unlabeled data can be processed to gain insights. For example, grouping customers into distinct categories

(Clustering) based on purchasing behavior. Applications include “pattern recognition”, “market basket analysis”, “web mining”, and many more [39]. The purpose of the algorithm is to notice regularities in the data and from there uncover hidden patterns from unlabeled data. When certain patterns return, the algorithm will look at the general correlation between different clusters of data. This is called “density estimation” in statistics. Unsupervised learning is a variant that does not apply to this master’s thesis [17].

3.4.3 Supervised learning

Supervised learning is a machine learning method that works with labeled data as the input to make predictions [40]. Labeled data here means that the dataset contains both the properties and the outcome of what has to be predicted. Every sample of the available data within this master’s thesis has a corresponding outcome, accordingly, we are dealing with labeled data. The aim is to predict the target variable, given the predictor variables. Therefore, supervised learning will be used in the master’s thesis. The two biggest applications of supervised learning are classification and regression.

3.4.3.1 Classification

Classification is applied when the data is divided into finite categories. The possible results are a discrete number of classes. Thus, it is a predetermined list of possible classes from which the algorithm will select. Models that describe essential data classes are called classifiers. Classification models are used in practice, for example in fraud detection.

In this research, we want to predict whether a policyholder will get disabled or not based on long-term absenteeism reports. There is a certain amount of information known about the policyholder, called features, such as the salary, age, and gender. Since we want to divide policyholders into “disabled” and “not disabled”, this is a classification problem.

3.4.3.2 Regression

Regression differs from classification, it can be used to predict certain values. Regression aims to predict the form of a continuous quantity. Therefore, regression is a method used to predict a numeric value, such as house prices and the temperature at a specific time.

3.4.4 Default probability in Machine Learning

Within this research, the focus is on the probability of becoming disabled, given certain characteristics of a policyholder, also known as conditional probability. With the classification application of supervised learning, we can determine a conditional probability [41]. The conditional probability can be determined by default probability prediction which is in most cases a classification problem [42]. Within this research, the default probability categorizes the status of an individual as being disabled or not, defined as not disabled (=0) and disabled (=1), and calculates the probability that an individual is in one of the two states [42].

3.5 Machine learning preparation techniques

Before explaining the different algorithms, we discuss some important concepts in the world of machine learning.

3.5.1 Dimensionality reduction

Features affect the performance of machine learning algorithms, which is why dimensionality reduction is an essential step in building a machine learning algorithm. Dimensionality reduction aims to remove irrelevant features/noise to resolve the issue of overfitting.

When the aim is to investigate the underlying structure of a group of variables, Factor Analysis (FA) is an appropriate technique to implement [43]. FA is based on variables that cannot be measured directly, also known as *latent* variables. FA is used to see if there are underlying factors in variables. In factor analysis, we try to find the presence of latent variables, which explains the pattern of observed variables. FA is primarily an exploratory technique, which is effective at finding groups of related variables in the data. Factor Analysis looks at underlying patterns and correlations between the different variables and puts the variables that have similar patterns together to form a factor. FA is based on the correlation matrix between indicators.

We perform FA to identify a smaller number of underlying variables for a large number of observed variables. Therefore, in factor analysis, we are trying to find the presence of latent variables, which explains the pattern of observed variables.

3.5.2 Underfitting and Overfitting

The terms “underfitting” and “overfitting” are used to determine whether or not a model is suitable for performing a particular task and will make appropriate predictions on unseen data. This can be observed in the learning curves of the model (Figure 3-4).

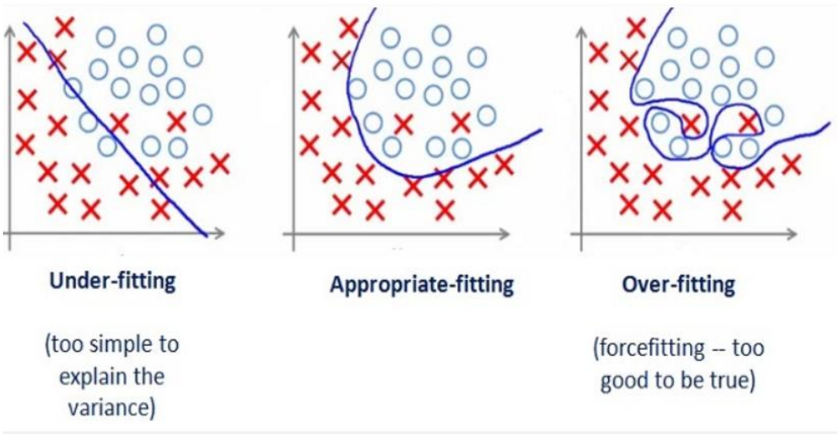

Under-fitting
(too simple to explain the variance)
Appropriate-fitting
Over-fitting
(forcefitting -- too good to be true)

Figure 3-4: Illustration Underfitting and Overfitting [44]

A useful model is generalizable. The better that generalization, the higher the score and the lower the error of the model on the test data will be. When the model becomes too complex, the generalization is lost, with the result that the model no longer makes accurate predictions on the test data. This phenomenon is called “overfitting”, the model has memorized the training data and scores very well on this, but not on unseen data [45]. Underfitting occurs when the model is not complex enough and the accuracy is too low [46].

A method to overcome overfitting and underfitting is to find the so-called “sweet spot” where the accuracy on the training data is high, without losing accuracy on the test data (Figure 3-5).

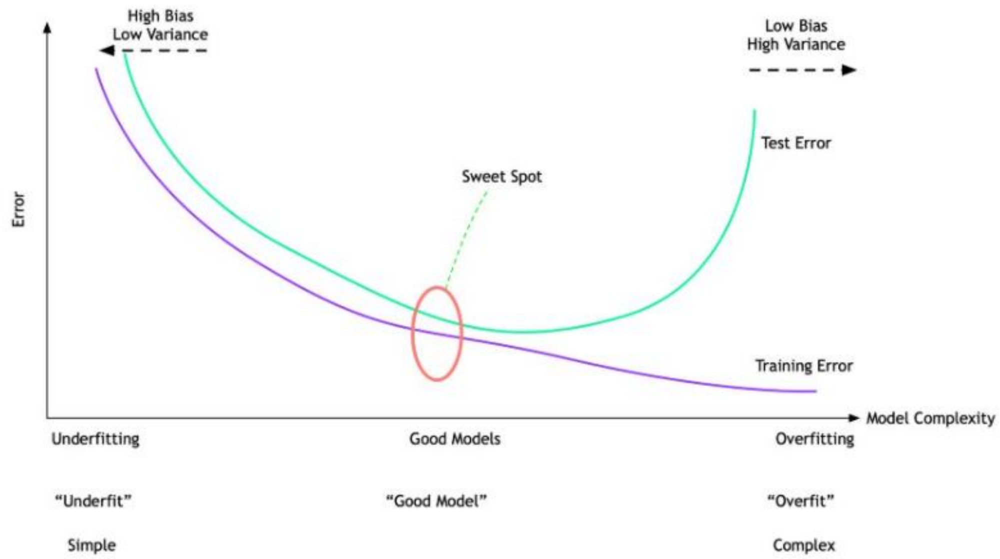


Figure 3-5: Visualization "Sweet Spot" [47]

3.5.3 K-Fold Cross Validation

If we build the whole model based on the entire data, the model cannot be validated. A solution is to create a training and testing set. Consequently, we can create a model on the training set and test the model on the testing set. However, this can cause dependency on the testing set. It is extremely important to make sure that the model does not accidentally see the test set during training, this is a form of “data leakage”. For example, a prediction on the test set, which serves as an indication of how well the model can predict new data, would give an optimistic picture. Especially with small data sets, it is therefore important to mix the data randomly first, so that the model is not trained on data from one particular scenario.

In machine learning, a hyperparameter such as the model type and model architecture is not affected by training. A hyperparameter is a configuration that is external to the model and whose value cannot be estimated from data [48]. The best setting values for parameters are found with grid search techniques [49], Section 3.5.4. When hyperparameter settings need to be tested, which is the case at most models, testing those parameter settings should not be done on the test set. That way the parameters would be changed until the model performs well on the test set and information from the test set would “leak” to the model.

A solution to this is a separate validation set, which serves to validate the settings of the hyperparameters after training the model. The data must therefore be divided into three (for example 70% - 15% - 15%), with the result that the number of samples to test or train the model will be reduced significantly. For the test set, this means a decrease of 10%.

Another solution for validating the hyperparameters is cross-validation (CV). A test set for final evaluation remains necessary, but a separate validation set is no longer necessary. The basic method is a k-fold CV, the training set is divided into k smaller subsets. For each “k” subset, a model is trained with “k-1” subsets and validated with the remaining subset. Ultimately, “k” models are trained (with the same settings) that are each time trained and validated with different data [49]. Cross-validation is used to predict the model’s ability to predict unseen data and catch any possible occurrence of overfitting. Moreover, the advantage of CV is that it gives a better estimate of the accuracy of the model than the evaluation on a single validation set. The disadvantage of CV is the long computation time required to train the models.

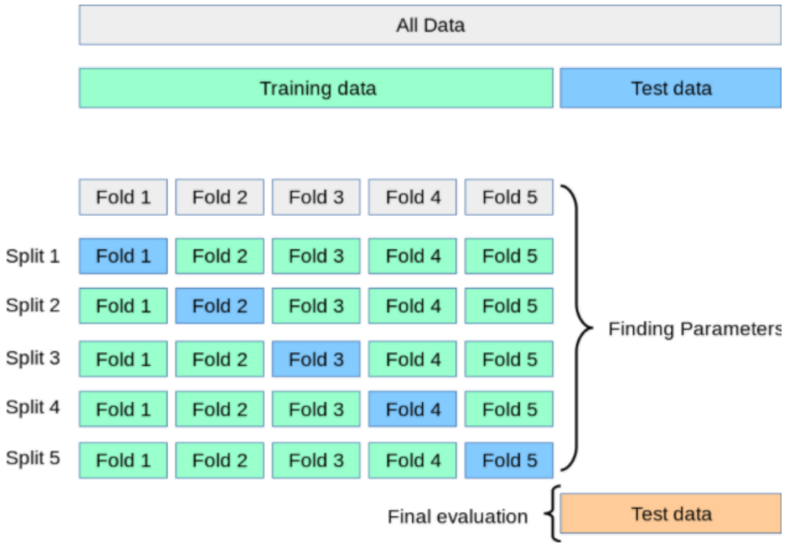


Figure 3-6: Splitting Data 5-fold Cross Validation

Therefore, the way cross-validation works are that it will set aside a portion of the data and it will perform multiple iterations (Figure 3-6). In this research, we use K-Fold cross-validation instead of splitting the data into train, test, and validation sets because it takes a large portion of the data. We want to get the maximum data points in both the testing and training set to get the best learning result.

First, it is going to split the training data into multiple folds. One of the folds will be used as the validation set and the rest will be used as the training set. When we are satisfied with the result, we will compare it with the test data. The higher the k-fold, the smaller the test set will get. The higher the k value, the more data is available for training the dataset, which will return often to higher accuracy.

Hence, we divide our dataset as visualized in Figure 3-6. First, the data is divided into two parts based on the class, one of 70%, and one of 30%. The 70% portion of the training examples determine the best parameters per model with 5-fold cross-validation [50]. With the parameters of the best model, we develop a model on the full 70%. The model is then compared with the models prepared with other techniques on the test set that contains 30% of the data. We use the same random seed in the division of the sets. This means that the division of the data is done in the same way for each model. Using this setup, the results can be compared with each other using paired tests. We calculate the mean of the performance of each iteration to evaluate the model (Figure 3-7). In this way, we go across all the samples.

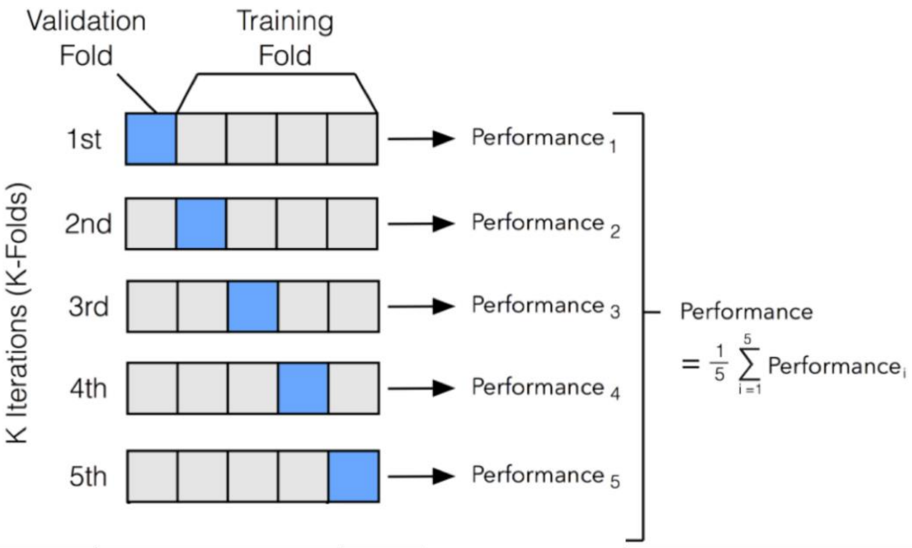


Figure 3-7: Performance K Iterations

The target variable within this research consists of two classes; disabled and not disabled. Since we have a class imbalance in our dataset, randomly splitting the data can influence the outcomes. Stratified sampling is a method that tries to keep the same proportion of each class in each subset. We perform Stratified K-fold to make sure the under-represented class label is equally split among all the splits. Therefore, with stratified sampling, we ensure that the disabled class will appear in the same amount in both the training and testing set.

3.5.4 Hyper-Parameter tuning

Machine learning algorithms depend on certain parameters. The optimal value of a parameter can be found using a validation set. Then the observations in the original training set are divided into a validation set and a training set [51]. For different values of parameters, the model is trained on the training set and tested on the validation set. The optimal value for the parameters is the values for which the deviance on the validation set is the lowest. Instead of a validation set, K-fold cross-validation can also be used (Section 3.5.3).

Almost all Machine Learning models have single or more parameters with which the model can be refined. These parameters are called hyperparameters. More precisely, a hyperparameter is a parameter whose value is set before the Machine Learning process starts. For example, with neural networks, the number of hidden layers and the number of nodes in each layer can be controlled.

Model performance is highly dependent on hyperparameters. Hyperparameter tuning, also known as hyperparameter optimization, is the process of finding the configuration of hyperparameters that provide the best performance. In most cases, hyperparameters have to be set manually, with which the results can be optimized (based on a chosen evaluation criterion).

An alternative is Grid search cross-validation which is an exhaustive search through a manually specified subset of the hyperparameter space of a learning algorithm. A grid search must be guided by a performance metric, usually measured by cross-validation on the training set or evaluation on a deferred validation set. This can lead to a significant improvement in model performance.

3.6 Machine Learning algorithms

After discussing the necessary preparations we describe the relevant Machine Learning algorithms. Despite the vast number of machine learning techniques that exist, we attempt to provide a clear overview of the most common and potentially useful machine learning techniques for this research. Logically, not every algorithm is equally suitable for the prediction of the disability probability. These models should be based on structured data and require supervised classification algorithms. We have taken into account several preferences within the selection of machine learning models. The algorithms are preferably applied more often, have an available implementation in Python, and can handle the data and type of variables available. Moreover, the algorithms need to be transparent and easy to communicate to all stakeholders.

The prediction of disability probability is in essence a classification problem. The aim of a classification model is in general to accurately predict the target class, for each observation in the dataset. Given certain characteristics of a WIA policyholder, put a label on it. Machine learning algorithms are used to discover patterns in data. In the current context, we discuss algorithms that can be applied in supervised classification. Supervised means that the algorithm learns from examples of which the outcome is known. For *m* training examples, the input variables or features $x = x_1, \ldots, x_n$ are known of a policyholder (age, gender, etc.) and an output or target *y* that is predicted. A class is predicted during classification (here the binary variable is disabled, *y* = (0;1).

Standard machine learning algorithms often deal with balanced data. However, in the insurance industry, the data is most often highly imbalanced with excess zeros. In this research, We first discuss the current insurance pricing model. Thereafter, we examine simple base classifiers and describe more advanced ML models.

3.6.1 The state-of-the-art insurance pricing model

To appraise the potential value of ML models, there needs to be a fair benchmark. A commonly used pricing model in the actuarial world to estimate the disability probability is the Generalized Linear Model (GLM). A GLM is a generalization of the linear model that is often used in statistics. Compared to linear models, for GLMs the independent variables do not have to follow the normal distribution. Furthermore, the dependent variable does not have to vary linearly with the independent variables: these are related to each other via a link function. The parameters for the GLM are chosen so that the “mean squared error” between the prediction and the actual target is kept as small as possible. The analysis of claim frequencies examines how the expected number of claims (λ_i) of policyholder *i* behaves as explanatory variables x_{ij} varies.

3.7 Base Classifiers

3.7.1 Naïve Bayes

Naïve Bayes (NB) is a method of calculating the probability of something happening (or classifying positive or negative), based on knowledge of previous times the same thing happened. Naïve

Bayes is a simple type of machine learning, but the algorithm usually performs quite well. One drawback is that it assumes that the variables used are independent of each other hence the name “Naïve”. However, the advantage is that it is faster to train and test than other machine learning algorithms. In practice, it is used, for example, to classify spam.

3.7.2 Logistic Regression

When the aim is to investigate the influence of one or more independent variables (X) on a dependent variable (Y), linear regression analysis might be suitable. Such a regression model assumes that the dependent variable is continuous, measured at interval or ratio level. However, it regularly happens that the dependent variable is of different measurement levels, for example, a nominal variable with only a few categories. Linear regression analysis is then impossible to perform. Various analysis techniques have been developed to be able to investigate the influence of all kinds of independent variables on a nominal variable [52]. The technique that most closely resembles linear regression analysis is logistic regression analysis. Logistic regression analysis is suitable for a dichotomous dependent variable: there are only two categories.

Logistic regression (LR) is one of the most popular machine learning algorithms for binary classification [53]. The logistic model is based on probabilities, or rather probability rations: odds. The odds in this study are the probability of becoming disabled (p_1) divided by the probability of not becoming disabled (p_0). An odds has a range of 0 (the probability of becoming disabled is zero) to infinity (the chance of becoming disabled is one). Since we prefer to calculate with a variable that runs from minus infinity to plus infinity, the natural logarithm of the odds is taken. So probability, odds, and logit are actually three ways of mentioning the same thing. If we call the independent variables X_1, X_2, and so on, the logistic model is:

$$\text{In}(\frac{P1}{P0}) = a + b_1X_1 + b_2X_2 + \ldots.$$

With a being the intercept, b_1 the parameter that indicating the effect of X_1, b_2 indicating the effect of X_2, and so on. We can also convert the logistic model into a probability model. The chance that someone will become disabled is:

$$\text{P}_1 = \frac{e^{(a+b1X1+b2X2+\cdots)}}{e^{(a+b1X1+b2X2+\cdots)}+1}$$

And the chance that someone will not become disabled is:

$$\text{P}_0 = \frac{1}{e^{(a+b1X1+b2X2+\cdots)}+1}$$

From these formulas, it becomes clear that the probabilities P_1 and P_0 add up to one. Furthermore, the probabilities P_1 and P_0 are dependent on X_1, X_2, and so on, but this dependence is not linear. Logistic regression is not like a straight line, but like an S-shaped curve (Figure 3-8).

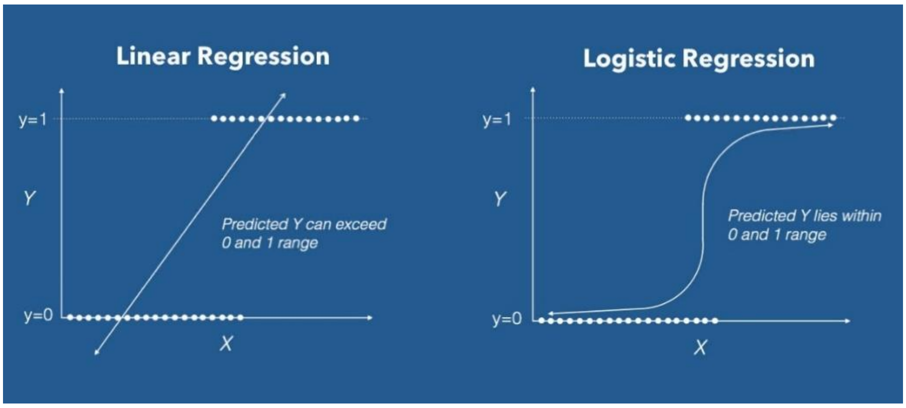
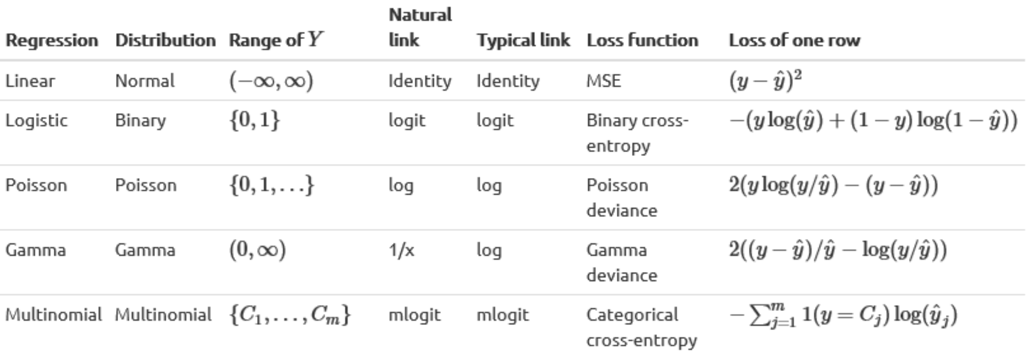


Figure 3-8: Linear Regression vs Logistic Regression [53]

Therefore, logistic regression is a method where the outcome variable is categorical, and the predictor variables are continuous or categorical. Logistic regression is a technique that can be used for traditional statistics as well as for machine learning. Simply put, logistic regression predicts which category of people falls into based on other information. If the outcome variable has two categories, (Disabled: Yes/No), it is called binary logistic regression. If it has multiple categories, it is called multinomial logistic regression.

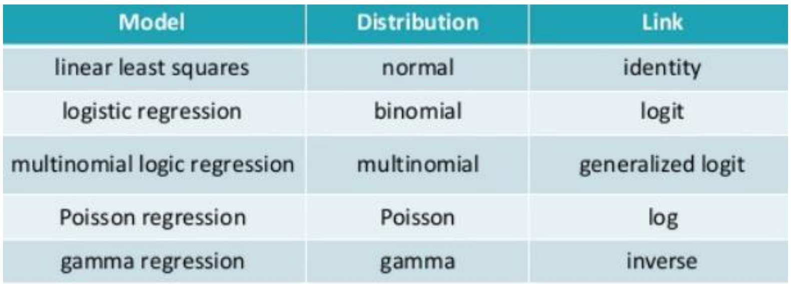
As discussed in Section 3.6.1, the GLM has different link functions depending on the response type. Table 3-1 displays some of the most used GLMs. When the output can be classified as 0 or 1, the distribution is binary. Binary logistic is a specific instance of a GLM (with a logit link function and binomial distribution). Therefore, using the GLM, the link function 'logit' needs to be used. When using the link function 'logit' we will get equivalent results as the logistic regression, Table 3-2. therefore, logistic regression is a special case of Generalized Linear Models. when we mention logistic regression, we are specifying the logit link function. When using the logit link function in the GLM, we will get the Logistic Regression.

Table 3-1: Most used GLMs [54]



Regression	Distribution	Range of Y	Natural link	Typical link	Loss function	Loss of one row
Linear	Normal	$(-\infty, \infty)$	Identity	Identity	MSE	$(y-\hat{y})^2$
Logistic	Binary	$\{0,1\}$	logit	logit	Binary cross-entropy	$-(y\log(\hat{y})+(1-y)\log(1-\hat{y}))$
Poisson	Poisson	$\{0,1,\ldots\}$	log	log	Poisson deviance	$2(y\log(y/\hat{y})-(y-\hat{y}))$
Gamma	Gamma	$(0,\infty)$	1/x	log	Gamma deviance	$2((y-\hat{y})/\hat{y}-\log(y/\hat{y}))$
Multinomial	Multinomial	$\{C_1,\ldots,C_m\}$	mlogit	mlogit	Categorical cross-entropy	$-\sum_{j=1}^{m} 1(y=C_j)\log(\hat{y}_j)$

Table 3-2: Overview GLM Models [55]



Model	Distribution	Link
linear least squares	normal	identity
logistic regression	binomial	logit
multinomial logic regression	multinomial	generalized logit
Poisson regression	Poisson	log
gamma regression	gamma	inverse

3.7.3 Decision Tree

The next algorithm which we investigate is decision trees (DT), which is the foundation of all tree-based models. These models are easy to understand and are often used for classification and regression problems. Decision trees are based on a non-parametric method as opposed to linear systems and the GLM. In a parametric method, the algorithm assumes to obtain data from a known distribution, for example, a normal distribution in the linear regression. With a non-parametric method, no assumptions are made about the distribution of the data [56].

In this method, it is assumed that comparable inputs provide comparable outputs. The algorithm learns a hierarchy of “if/else” questions. In this way, the data is divided into regions (see the top of Figure 3-9). The tree consists of nodes, leaves, and edges (see the bottom of Figure 3-9). The nodes represent the questions, the leaves are the result. The edge is the connection between the answer to the question and the next question in the hierarchical line [57].

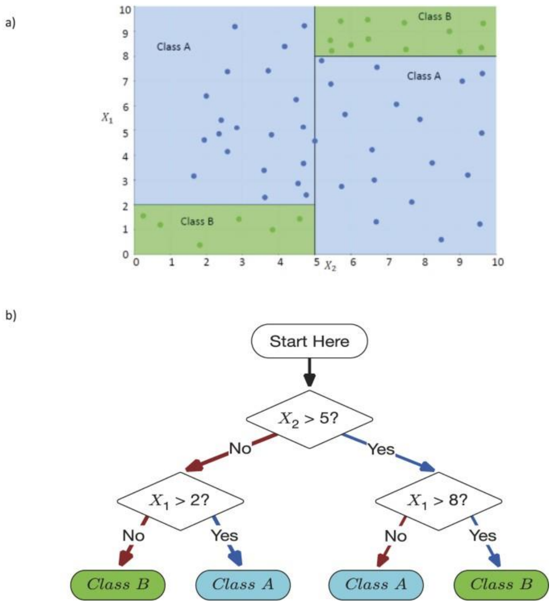


Figure 3-9: Example Decision Tree

When building a decision tree, it is essential to find the most important information question. This is a recursive process in which every possible combination of questions is passed through. Each question becomes a node and will in turn be responsible for further splitting the data. As can be seen in Figure 3-9, the root of the tree is the one that takes care of the initial division of the data. Namely, if X_2 is greater than 5, then the edge is tracked “Yes” and the algorithm moves to the next hierarchical question on the right. When X_2 is less than or equal to 5, the edge is tracked “No” and the algorithm moves to the next question on the left.

When the algorithm is not assigned a limit, the complexity will quickly increase and the chance of overfitting becomes high. This can be adjusted based on a few parameters, the most important parameter is the maximum depth of the decision tree [58]. One can also set the maximum number of leaves and the minimum number of data points needed to create a node. This strategy to prevent overfitting and thus abort the algorithm prematurely is called pre-pruning. Also, there is the possibility to simplify the decision tree after training, this is called post-pruning [57].

An advantage of decision trees is the ability to determine feature importance after the model has been trained. This means how heavily a feature weighs on the algorithm when making a prediction. Note that this is not the same as with linear regression where the coefficients can also be retrieved, namely the weights and bias of the parametric model. In the case of decision trees, this is a score from 0 to 1 to indicate the extent to which a feature is important. The total score of all features is equal to 1.

In the insurance industry, a decision tree mimics the human decision-making process in the form of yes-no questions. Groups of homogeneous risk profiles are created based on certain characteristics of policyholders. The benefits of decision trees are very clear, it is easy to understand the model and when complexity is limited, it is also relatively easy to visualize. Also, scaling the data is not required in advance. The disadvantage is that the model is always overfitted, despite the parameters such as the depth of the tree being adjusted.

3.8 Ensemble Methods

Ensemble Methods are part of Machine Learning where one combines multiple individual models to jointly make better predictions. Ensemble methods are mainly applied where other, simpler models cannot achieve the desired prediction quality. There are three different Ensemble Methods.

With the Stacking method, we train several different algorithms with the same data. The results of the different algorithms are the input of an end model, which gives the final prediction. The emphasis here is mainly on the use of different algorithms on the same data. In practice, Stacking is used less often concerning Bagging and Boosting, because it is often less accurate and it has a lack of interpretability.

With the Bagging method, one divides the dataset into separate subsets segments. We then train separate versions of a specific algorithm on each of the subsets. To make a prediction, each of the models that have now been made gives a prediction, of which we can take the average. Bagging stands for bootstrap aggregating. Bootstrapping is a statistical method of sampling, aggregating means collecting. Bagging is often used where real-time predictions are required. An example of a bagging technique is random forest.

With the Boosting method, we train successive algorithms with the data of incorrect predictions from the previous algorithm. Therefore, there is always a new algorithm that is trained on the wrong predictions made earlier. By focussing on an ever-decreasing number of incorrect predictions with Boosting, accurate prediction results can be achieved. Well-known Boosting models are, for example, Gradient Boosting. Boosting is an algorithm that helps reduce variance and bias in a machine learning ensemble. Moreover, boosting is a resilient method that easily curbs too many passes. Since boosting methods are known to build strong predictive models, we focus on several boosting ML models [59].

3.8.1 Gradient Boosting

Gradient Boosting (GB) is based on several decision trees, which are calibrated sequentially; after the first calibrated decision tree, the second decision tree is fitted, and so on. Boosting does not take samples from the training set. Instead, decision trees are generated sequentially. Therefore, a GBA is an additive model.

The Gradient Boosting Algorithm is a popular boosting algorithm. GB, like any other machine learning procedure, sequentially adds predictors to the ensemble and follows sequence when correcting previous predictors to arrive at an accurate predictor at the end of the procedure. GB uses gradient descent to indicate the challenges in the previously used predictions. The previous error is highlighted and by combining one weak predictor with the next predictor, the error is significantly reduced over time.

The most common form of this algorithm optimizes the Mean Squared Error (MSE). This is the average difference between the predicted targets and the actual values. This model combines the weaker forecasting models, or different weak predictions, to create a strong model. With a strong forecasting model, the average prediction corresponds to the actual value. The more weak predictions are added to this algorithm, the lower the MSE will be.

3.8.2 Extreme Gradient Boosting (XGBoost)

XGBoost, an abbreviation of Extreme Gradient Boosting is a specific implementation of Gradient Boosting, with some improvements which make the model faster and more accurate than traditional Gradient Boosting. XGBoost is an ensemble decision tree-based machine learning algorithm and uses a gradient enhancement framework.

3.8.3 Adaptive Boosting

Another ensemble method is Adaptive Boosting (AB), an algorithm that combines the output of “weak learners”. The output is a linear combination of the output of simple models. AdaBoost stands for Adaptive Boosting and is adaptive such that the simple models are weighted based on previously misclassified training examples.

3.8.4 Light Gradient Boosting Machine

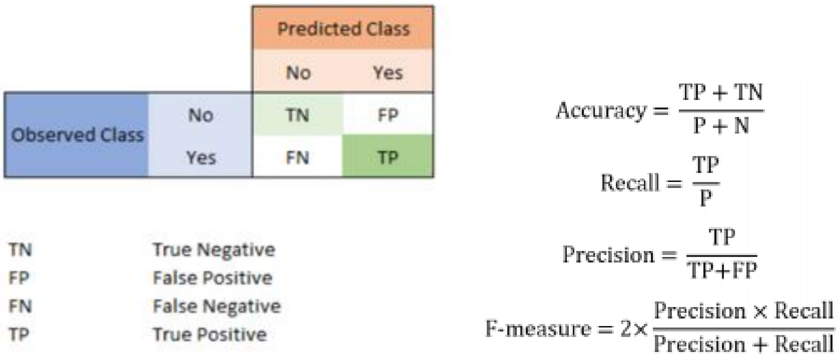
The Light Gradient Boosting Machine (LightGBM) is another variation of gradient boosting. “Light” stands for the lighter version since this model is faster. Its advantages are the rapid training speed and the higher efficiency.

3.9 Evaluation metrics

The aim is to find the model with the most accurate predictions of the disability probability. Therefore, the tested models need to be evaluated. In this section, we describe specific methods for quantifying the quality and reliability of machine learning algorithms. For classification, performance measures can be divided into two categories: deterministic and probabilistic [60]. Deterministic performance measures only consider the predicted label and are based on the confusion matrix (Figure 3-10) or can be derived from it. Probabilistic performance measures also include the certainty of prediction.

3.9.1 Confusion matrix

Classification models can be evaluated using a confusion matrix which is a table with the relationships between positive and negative predictions and reality.



		Predicted Class	
		No	Yes
Observed Class	No	TN	FP
	Yes	FN	TP

TN	True Negative
FP	False Positive
FN	False Negative
TP	True Positive

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{P}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

$$\text{F-measure} = 2\times\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Figure 3-10: Confusion matrix & Performance measures [61]

Accuracy is defined as the percentage that is properly classified. The recall is the percentage of positive examples classified as positive. Precision is the percentage of the number of positive examples classified as positive. The F-measure combines precision and recall by taking the harmonic mean. Each performance measure emphasizes a different aspect; a better performance according to one measure may correspond to a lower performance according to another measure.

For classification, there is a class of algorithms that produce a probability in addition to only the most likely class. These probabilistic classifiers thus provide a degree of certainty about the prediction that can be included in the evaluation. Many probabilistic performance measures are graphical and show how the model performs over a range. Examples are the Precision and Recall curve, the ROC curve, and the lift graph.

One visualization parameter for evaluating the performance of a classification model is the Receiver Operating Characteristic (ROC) curve. ROC graphs provide a simple way to summarize the information within the confusion matrix. A ROC curve is a way to visualize the performance of a machine learning algorithm. In the ROC curve (Figure 3-11), the sensitivity of the test (true positives) on the y-axis is plotted against the false positives (1 – specificity) on the x-axis on different cut-off values. The most optimal cut-off value is in the top left corner of the curve (high proportion of true positives and low proportion of false positives). The 'Area Under the Curve' (AUC) indicates how accurate a test is: 1 is a perfect test, which can identify disabled policyholders without false positives, and 0.5 is a worthless test, which detects as many true positives as false positives.

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

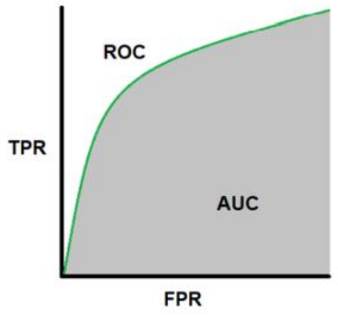


Figure 3-11: ROC-AUC Curve [62]

3.10 Handling data imbalance techniques

Machine learning techniques perform best when the number of observations of each class of the target factor is equal. However, the analysis of the available data shows that there are unbalanced classes. The policyholders who become disabled is considerably smaller than the policyholder who does not become disabled. In other words, becoming disabled in comparison with not becoming disabled for the DII case.

Therefore, in this section, we focus on techniques to handle imbalanced data since we are dealing with class imbalance as discussed in Section 3.5.3. The techniques to handle imbalance will be executed and evaluated on the base classifiers.

Class imbalance exists when the data has an unequal number of examples in each of its classes. Unbalanced data quickly leads to a bias (preference) or overfitting of the majority class. The performance of a predictive model is therefore greatly reduced. Unbalanced data is a well-known problem in the machine learning literature and has been studied for several decades. It is a common challenge that one group is significantly more strongly represented than the other. What can arise is called the Accuracy Paradox, a model that can overfit the largest group, so that only the large group is predicted well. Several techniques can be applied to the data before that training data is presented to the learning algorithm [63].

3.10.1 Undersampling and Oversampling

One way to overcome the bias of the excess class is to adjust the training data and manually restore the balance. This method can be performed in two ways, by using undersampling or oversampling (Figure 3-12). When using oversampling, the minority class is copied until the two classes are back in balance. In addition to the fact that this results in longer training times because there is more data, it must be taken into account that this does not lead to overfitting. If it does lead to overfitting, the machine learning model is trained too much on the training data, so that it can only classify known data properly and unknown data (the validation set and test data set) is classified worse.

In oversampling we increase the number of samples in the minority class to match up to the number of samples in the majority class. The basic fundamental idea of oversampling is taking the minority class and create new samples that match up to the length of the majority class samples.

Another way to deal with class imbalance is undersampling. Undersampling is a technique wherein the number of samples in the majority class is used to match up to the total length of the minority class samples to create balance within the dataset. In undersampling, the minority class data set remains as it is and the majority class data set will be reduced to fit the size of the minority class.

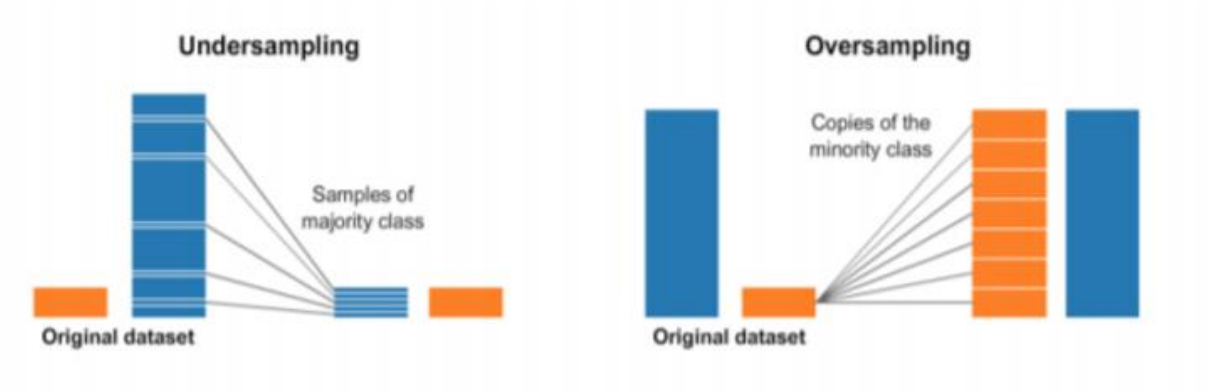


Figure 3-12: Undersampling and Oversampling [64]

3.10.2 Synthetic Minority Over-Sampling Technique

Another method of oversampling is Synthetic Minority Over-Sampling Technique (SMOTE) [65]. SMOTE is a method of synthesizing data to amplify datasets containing rare events or scenarios whose detection is critical, such as cancer detection. It is an oversampling approach in which the minority class is increased by creating “synthetic” points in the dataset. [66] In SMOTE, artificial samples are created of the minority class. Synthetic data can be especially useful in cases where there are too few minority class examples for a model to know the decision boundary effectively.

Rather than duplicating existing information, data scientists can synthesize examples around the minority class using SMOTE. This technique constructs new minority class observations by creating ‘synthetic samples’ rather than simply replicating the minority class [67]. By using SMOTE to create more samples in and around the minority class, the network can better define the minority class and the critical boundaries around it.

With SMOTE, the minority class is sampled by sampling under each of the minority classes and introducing synthetic data points along the line segments corresponding to one or all of the k-NN of the minority classes. The nearest neighbors are chosen randomly depending on the amount of oversampling required. The SMOTE algorithm finds a record that is similar to the record being upsampled and creates a synthetic record that is a randomly weighted average of the original record and the neighboring record, where the weight is generated separately for each predictor [68].

3.11 Hyper Parameter Tuning

Machine learning algorithms train on data to find the best set of weights for each independent variable affecting the predicted value or class. For a machine learning algorithm, there are some default parameters defined. However, we want to achieve the best possible performance of a model, and therefore get the best parameters for the model. To achieve this, hyper-parameter tuning needs to be performed to find the optimal parameters for a specific model.

Hence, algorithms themselves have variables called hyperparameters. They are called hyperparameters as opposed to parameters because they control the operation of the algorithm rather than determining the weights. Hyperparameter tuning, also known as hyperparameter optimization, is the process of finding the configuration of hyperparameters that provide the best performance. In this section, we discuss the process of hyperparameter tuning and the impact on overfitting.

With the function *GridSearchCV* within Python, the predefined hyperparameters can be looped through and be fitted to the model on the training set. This package offers automatic hyperparameter tuning. Essentially, we tell the system which hyperparameters we want to vary, and possibly which metrics we want to optimize. On all the combinations, the package will fit the model. It is called GridSearch CV since the combinations are set in a grid. For all the intersections in the grid, the model will run once.

4. Data preprocessing

Data preprocessing is an essential step in building a machine learning application. The algorithm needs to receive the data in a “correct form”. In this chapter, we first explore the available internal data. Next off, we will discuss some techniques for data preprocessing. It is also important that research is carried out into the mutual correlation of the characteristics and against the output variables.

4.1 Data Exploration

In the first stage, we explore the data.

4.1.1 Reading the data

In this section, we describe the internal data which has been used within this study. Within the Disability Income Insurance, several data sources are available. The most important data sources for this research are policy-, cover- and employee data, 42nd-week reports (Absenteeism data), disability from the UWV (SUAG), and WIA claims.

4.1.1.1 Policy

For the WIA policyholders, Achmea has information about the type of coverage and the insured employees in the policy administration. An exception to this is employers who have taken out insurance with Achmea utilizing proxies. Policy information is available in a self-regenerated data mart. The scope of the research is to predict WIA claims and only contracts of this product within the policy data are considered. The policy data contains contract, company, and employee level information, which we distinguish in Table 4-1, Table 4-2, and Table 4-3. In total, the policy data contains 382,772 policyholders in 2017 and 357,194 policyholders in 2018. Basic data which do not apply to the scope, e.g. payment method, are removed from the dataset. In Appendix A we give an overview of all the individual variables in the policy data.

Table 4-1: Policy data description – Contract level (limited overview)

Feature	Description
Policy number	Contract number
Producttype	Type of WIA insurance product
Start date	Start date WIA insurance
End date	End date WIA insurance
Insurance year	Year of WIA insurance contract
Number of insurance years	Length of insurance in one year
Contract System	Type of contract system.

Table 4-2: Policy data description – Company level (limited overview)

Feature	Description
Company name	Name of the company where the employee is employed
Sector	The sector of the company where the employee is employed
Sector code	Code of the company sector
Company name	Name of the company where the employee is employed

Table 4-3: Policy data description – Employee level

Feature	Description
PSN	Unique ID per policyholder/employee
Birthdate	Birthdate of the employee
Gender	Sex of the employee
Employment relationship	Type of labor contract (fixed-term, indefinite-term contract, etc.)
Salary	The income of the employee per year

4.1.1.2 Absence

42nd weeks reports (long-term absenteeism) are reported by employers and registered by an external party (Verzuimdata- VDN). Also, VDN registers the recovery reports (between the 42nd-week reports and WIA influx). The variables which are used from the VDN data are displayed in Table 4-4. VDN creates dashboards based on this data (supplemented with no policy data that is also managed by VDN) that Achmea uses to control various processes. One of the dashboards is the WIA Care 42nd-week dashboard that provides insight into the reporting behavior of employers to submit their 42nd-week reports in time.

Table 4-4: Absence data information

Feature	Information
PSN	Unique ID per policyholder
Date sick	First day of illness of a policyholder
Date 42nd-week report	Reporting date of the 42nd-week notification
Claimyear	Year of illness

The 42nd-week report's data are only available since 2017. Moreover, within the 42nd-week reports, there are some unusual cases such as the date of the 42nd-week report being later than the date of illness, these records are removed from the dataset. Since the data is recently available and to ensure we have enough data points, we have set the boundaries of the 42nd-week report date from 3 months too early till 3 months too late. In total, there are 7269 unique policyholders with long-term absenteeism in 2017 and 2018.

Figure 4-1 displays the percentage of long-term absenteeism in proportion to the total WIA policyholders. From this, it is obvious that the proportion of long-term absenteeism is minuscule in comparison with the total policyholders.

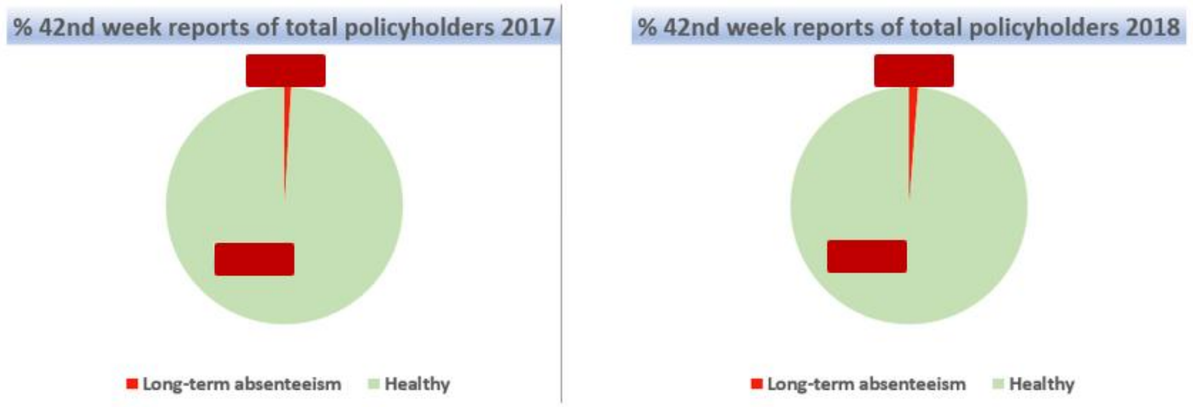


Figure 4-1: Proportion long-term absenteeism of total WIA Policyholders 2017 & 2018

4.1.1.3 Claims

This data source contains all disability data, policy data, and employee data for each claim to calculate a benefit. From this dataset, we determine whether long-term absenteeism has become disabled. In other words, whether a policyholder has been ill for at least two years. Table 4-5 displays the variables from the claims dataset which are included. The 'AOKlasseCode' variable indicates the percentage of disability in case of disability. The disability percentage is distinguished in three different values (Table 4-6).

Table 4-5: Claims data information

Feature	Information
PSN	Unique ID per policyholder
AOKlasseCode	Disability percentage in codes

The target of this research is the WIA-Influx. Within the research, an additional column for the target variable is added. A policyholder is labeled as disabled when the disability code is 02, 03, or 04. When there is no code or the code is 01, the policyholder is label as not disabled since in that case, the policyholder will not become in the WIA. The domain of the target value is [0-1], where 0 indicates the policyholder is not disabled and 1 indicates the policyholder is disabled.

Table 4-6: Disability Percentage

Code	Information
01	< 35%
02	WGA 35% - 80%
03	WGA 80% - 100%
04	IVA

4.1.2 Exploratory Data Analysis

After having a first impression of the available data, we analyze the data in more detail to familiarize ourselves with the data. The purpose is to determine the relationship with the target attribute and how the input characteristics are distributed. Visualizing the data is an appropriate method to quickly gain more insight. We perform the visualizations both in Excel and Power BI.

First of all, Figure 4-2 displays the distribution of the total active WIA policyholders in 2017 and 2018. It appears that in 2017 there were more policyholders in comparison with 2018. Figure 4-2 displays the total number of long-term absenteeism cases (42nd week reports) both in 2017 and 2018. Figure 4-3 displays the percentage of the target prediction, WIA influx, of the total long-term absenteeism cases. As can be seen in Figure 4-3, the proportion of disabled persons in comparison with not disabled persons is small, which means the data is imbalanced.

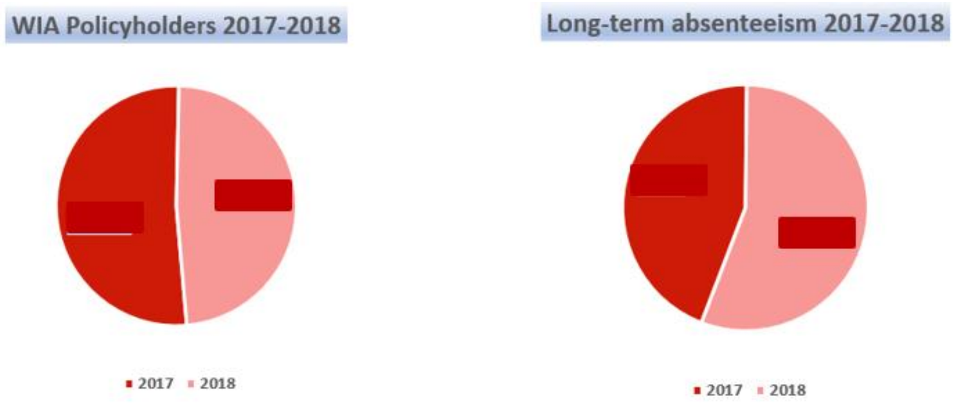


Figure 4-2: Active WIA policyholders in 2017 & 2018 (left) and Total long-term absenteeism 2017 & 2018 (right)

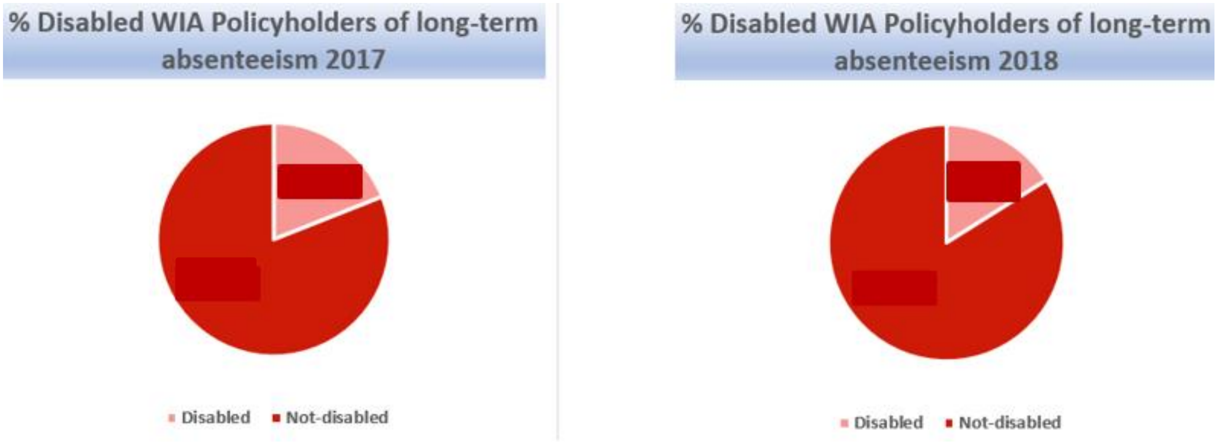


Figure 4-3: % WIA Influx of total 42nd week reports

4.1.3 Univariate Analysis

In a univariate analysis, only one variable is treated at a time. The purpose is to explore the distributions of the available features. We analyze the features which are currently used in disability prediction. Due to privacy, the axes are hidden.

4.1.3.1 Gender

Figure 4-5 and Figure 4-4 display the division of long-term absenteeism and disability cases by gender. It appears that more women have a 42nd-week report in comparison to men since the percentage of WIA policyholders being a woman is approximately 42% and 41% consecutively for the insurance years 2017 and 2018. This difference appeared in both 2017 and 2018 with only a marginal change, therefore, indicating that it is likely a consistent difference. However, the difference is minimal.

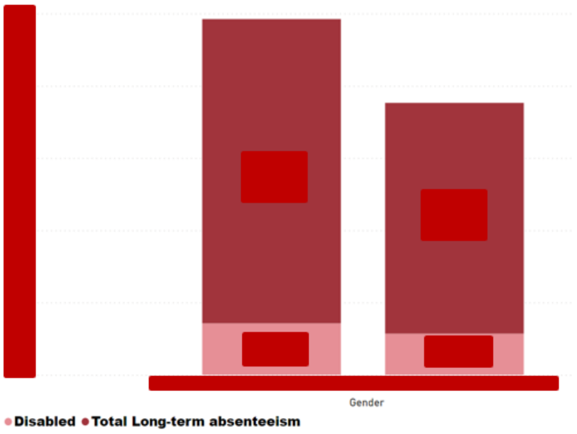


Figure 4-5: Long-term absenteeism versus Disability by Gender

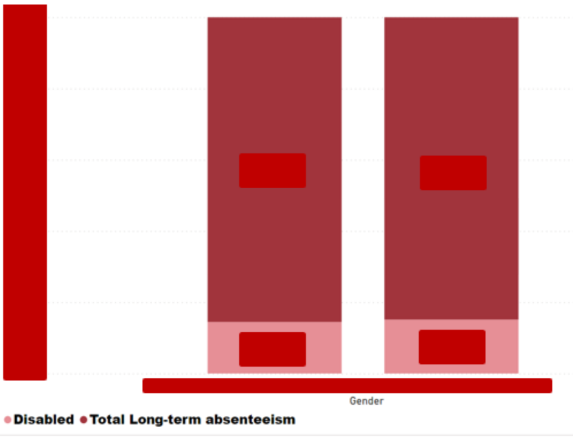


Figure 4-4: % Disability of total Long-term absenteeism by Gender

4.1.3.2 Age

Figure 4-6 displays the distribution by the variable age. It appears, for both long-term absenteeism and disabled individuals, that there is a linear relationship for the age variable.

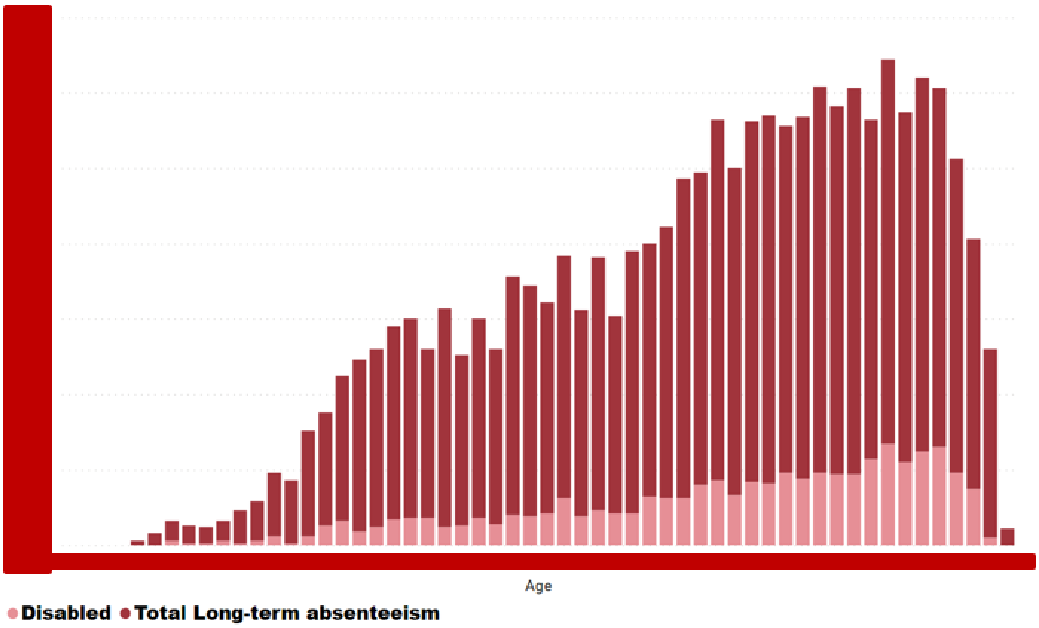


Figure 4-6: Long-term absenteeism versus Disability by Age

4.1.3.3 Employment

From **Error! Reference source not found.** it appears that the proportion of permanent contracts is m uch larger than fixed-term contracts and on-call workers. However, this is in line with the distribution of the total WIA policyholders.

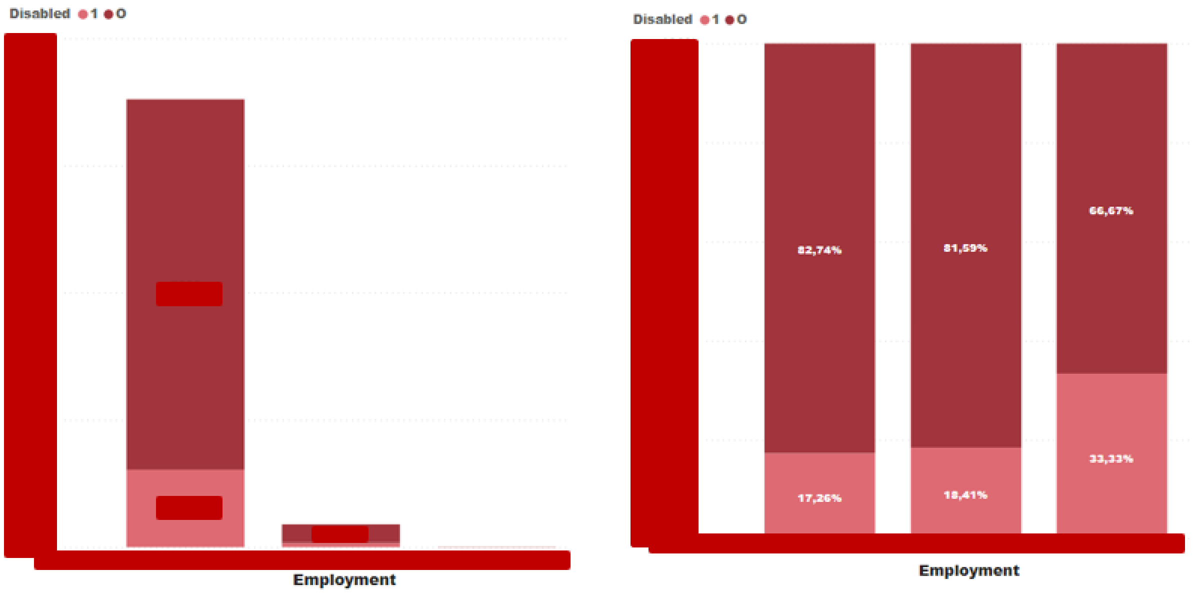


Figure 4-7: Long-term absenteeism versus Disability by Employment

4.1.3.4 Salary

From Figure 4-9 it becomes clear that the gross annual salary has a skewed distribution, more specifically right-skewed.

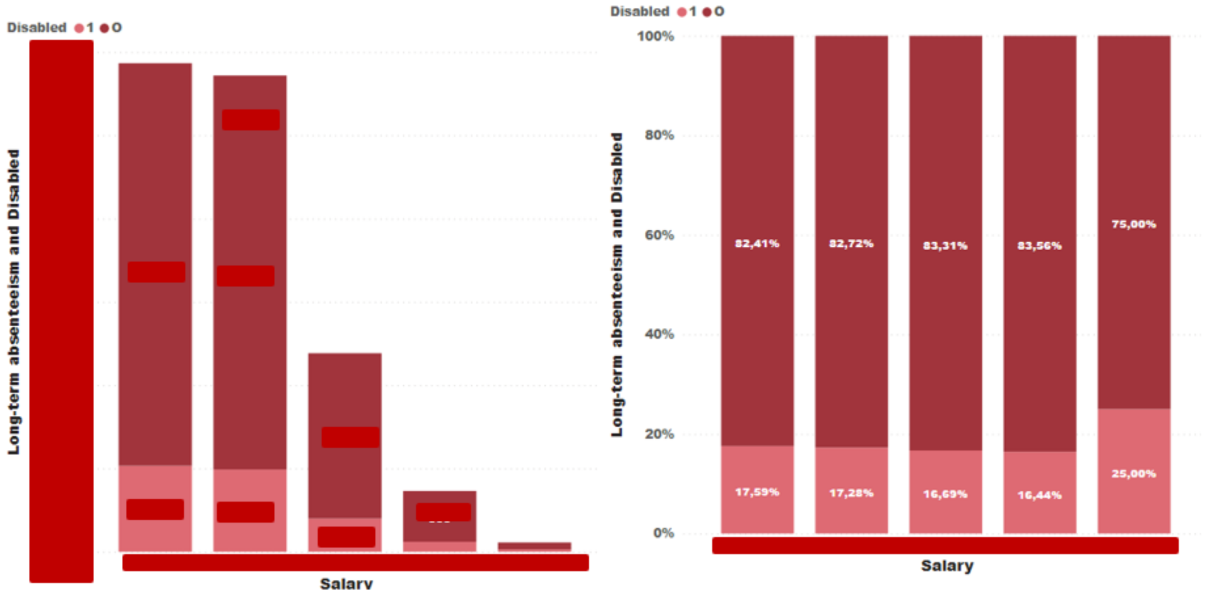


Figure 4-8: Salary distribtution

4.2 Merging Process

In this section, we elaborate on the merging process to gain the final dataset we use in this research. The data is merged within SAS, statistical software for mainly data management, business intelligence, and predictive analytics. The data of 42nd week reports can be merged with the policyholders based on the primary key: PSN. The policy data contains multiple rows for one unique PSN. To get unique records, we perform multiple steps within SAS. We made several assumptions and choices in merging the datasets.

First of all, only the WIA policyholders that appear in both the policy data and VDN data are included. Moreover, only the policyholders which were active in 2017 and 2018 are included in the merging process. To ensure a policyholder had active WIA insurance at the date of illness, we include a condition that the date of the illness is higher than the starting date of the insurance contract.

For each policyholder, the necessary features are obtained and aggregated based on the unique PSN. Policyholders with different WIA insurance products have an indicator for each product to prevent double records. In this research, we do not distinguish between different WIA insurance products.

Unfortunately, there are cases where there are several values for one variable, such as different genders, birth dates, and salaries. We have taken the maximum of each of these variables.

After achieving unique records per PSN within the policy data, this data is first merged with the long-term absenteeism data with an inner join statement. Thereafter, the composite dataset is merged with the claims data with a left outer joint statement. The tables are merged per year on the PSN since one policyholder might be in insurance for several years. The result gives a merged dataset of both WIA policyholders who had a claim and who did not have a claim together with the needed variables (Figure 4-9).

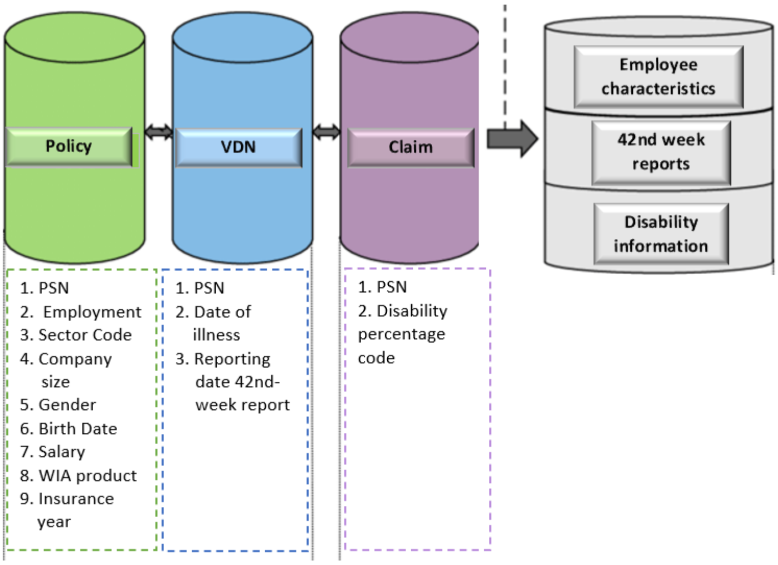


Figure 4-9: Overview of the merging process with 42nd week reports

4.3 Data Preparation

The next step of the machine learning project is to prepare the data. We perform this process in Python.

4.3.1 Missing values and zero values

Treating missing values is a fundamental step in preparing data for analysis. Missing values are data points of a variable that are missing. It can have a significant effect on the conclusions which are drawn based on the data and hence the method for dealing with missing values needs to be selected carefully [69]. It often happens that the dataset contains missing values and/or zero values, especially when data gathering did not have the intention of using machine learning techniques for prediction. The treatment of these cases depends on the situation. If the number of missing or zero values of an attribute is rather small, they can be replaced by, for example, the mean or median. When there are many missing values for a feature, it is better to remove that features. There are several more advanced methods such as k Nearest Neighbor and mice. Several functions in Python are available for this.

At Nearest Neighbor, the k closest examples are selected by the Nearest Neighbor algorithm and then the missing values are replaced with either the mean (for numerical variables) or the most common value (for categorical variables). Determining the closest examples can be done with different techniques and depends on the data and type of variables.

Mice, multivariate imputation by chained equations, is a method that can be applied when more than one variable contains a missing value [70]. “Chained equations” are used; missing values are replaced several times and the results are combined. With mice, it is possible to maintain relationships that come from transformations, combinations, or coding of variables. Mice also have a lot of freedom in choice: when specifying an “imputation model” there are seven choices to make, such as the shape of the model and the sequence of imputation.

4.3.2 Anomaly detection

Outliers are unusual data points that are different from the rest of the observations, it is a data point that is statically different from other data points. To train a model correctly, it is necessary to remove those outliers from the training set because those data points are not representative. It can disturb the statistical power of the data analysis process. There are several ways to detect and remove outliers, two of them are briefly discussed here.

The first method is the z-score method, where the z-score is calculated for each value, per column. All values that have a z-score above a predetermined limit value (often a value between 2 and 3.5) are then taken from the dataset.

A second method is an unsupervised Machine Learning technique, called “isolation forests”. It is based on random forest and can also be used for anomaly detection. To “isolate” a sample, a characteristic is chosen at random. Afterward, a “split value” is randomly chosen between the maximum and minimum values of the selected characteristic. The recursive splitting is represented in a tree. The average number of times to split for a particular sample is a measure of the normality of that sample. The advantage of this technique is that it scales better with high-dimensional data. In this research, we make use of outlier detection and removal for the variable ‘Age’.

For the feature ‘Salary’, there are marginal outliers. With the function: RobustScaler within python, the median is subtracted rather than the mean. Instead of taking into account the minimum and maximum values, the Inter Quantile Range (IQR) is used. Therefore, it is “robust” to outliers, but it will not completely remove outliers.

4.3.3 Feature Scaling

Each feature has two important components: magnitude and units. Magnitude is the value of the feature in a specific record and the unit is the way how the magnitude is measured. For example, for the feature 'Age = 40', 40 is the magnitude and years is the unit. Most of the machine learning algorithms work with Euclidean distance which is the shortest path, a straight line, and the distance can be calculated as the square root of the sum of the squares of the differences between the coordinates, according to the Pythagorean theorem. Whenever an algorithm is working on the Euclidean distance, the magnitude might play a very important role since the distance between 200 and 300 is much higher in comparison to the distance between 20 and 30 for example. Therefore, for some algorithms, it is important to scale the features before feeding them to the model.

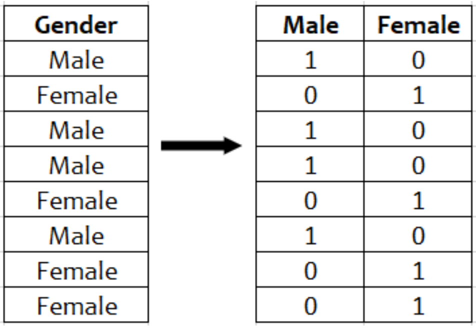
For the feature 'Age', the boundaries are predetermined since only individuals from 16 till 65 can become a WIA Policyholder as stated in the policy conditions. Therefore, for the feature 'Age' we manually scale between 0 and 1.

4.3.4 Encoding

This method is often used when the data is divided into categories and is not continuous/numerical. As an example, a database of employees is viewed in which the gender of the employees is stored. The gender may be stored in the form "V" or "M" in the same column. The algorithm cannot calculate this. Therefore, the attribute must be converted to a numeric value. One option would be to use the number "0" for men and "1" for women. Again, the "gender" attribute will take up one column. The disadvantage of this way of working is that the model will interpret a value (which is coding for a certain category) as a probability of a certain event. This is because most algorithms look for a function between the input parameters and the intended target.

A more efficient way is Dummy encoding. The feature is divided into several binary features, equal to the number of possible categories of the original characteristic. In the case of gender, there are two characteristics: male and female. Where the original attribute took up one column, it now uses two. Only one of these two is equal to "1" per sample and the rest is "0". For example, if the person is a woman, the column "woman" will contain a "1" and the other column "man" will contain a "0". This process is displayed in Figure 4-10.

To prevent multicollinearity, one of the columns needs to be deleted. For example, we delete the column woman. It is obvious that if M=1, then V=0 since both are not possible. In this research, we perform the Dummy encoding process for the non-ordinal categorical variables: Gender, Employment, and SectorCode.



Gender
Male
Female
Male
Male
Female
Male
Female
Female

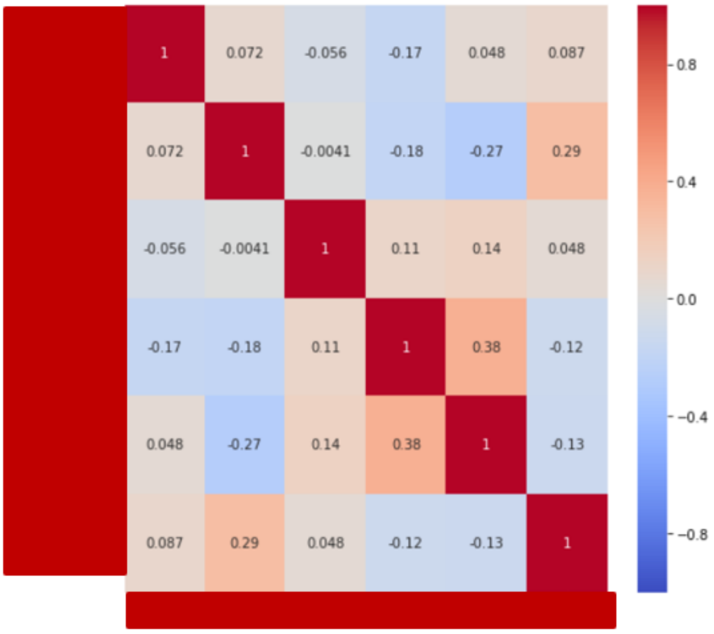
Male	Female
1	0
0	1
1	0
1	0
0	1
1	0
0	1
0	1

Figure 4-10: One-Hot Encoding example Gender

The variable salary is an ordinal variable and therefore has a hierarchy. For this variable, we transform the values to numerical in one column.

4.3.5 Dimensionality reduction

We perform Factor Analysis (FA) to reduce dimensionality. The starting point of FA is a correlation matrix. In a correlation matrix, a large amount of data is summarized to see patterns. Figure 4-11 displays a concise correlation matrix.



[redacted]

[redacted]

1	0.072	-0.056	-0.17	0.048	0.087
0.072	1	-0.0041	-0.18	-0.27	0.29
-0.056	-0.0041	1	0.11	0.14	0.048
-0.17	-0.18	0.11	1	0.38	-0.12
0.048	-0.27	0.14	0.38	1	-0.13
0.087	0.29	0.048	-0.12	-0.13	1

Figure 4-11: Concise Correlation Matrix (hidden for privacy)

In FA, only factors with high Eigenvalues are preserved. The Eigenvalue indicates how much additional variance is explained by the extra factor. The process of determining how many factors to keep is called extraction.

We create a graph by plotting the eigenvalue on the y-axis against the factor on the x-axis. This graph is called a scree plot (Figure 4-12). There are as many possible factors as there are variables, but there are only a few factors with high eigenvalues and many with a low eigenvalue.

To determine if an eigenvalue is high enough to main the factor we use the Kaiser's criterion [71]. The Kaiser's criterion indicates that factors with eigenvalues greater than one must be preserved. In Appendix C the eigenvalues of the factors retrieved from the factor analysis are displayed. In total there are 54 factors with an eigenvalue above one.

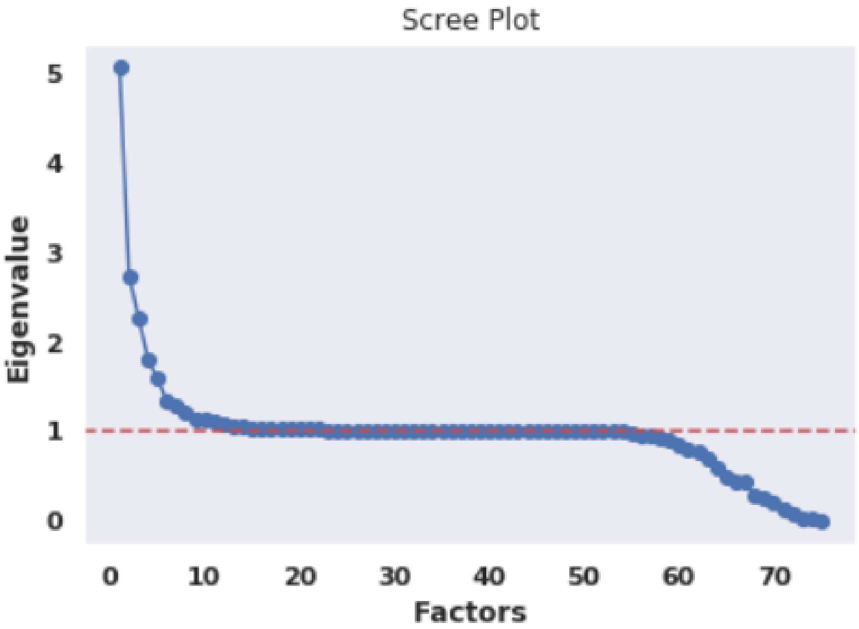


Figure 4-12: Scree Plot

5. Design and Development

The research aims to classify disability, given long-term absenteeism based on features of a policyholder. In this chapter, we develop the predictive model for the disability probability based on long-term absenteeism described in Chapters 3 and 4. The developed model can predict the target variable of a new observation if the observation has the same variables as those in the training set. We compare different algorithms and select the most suitable model based on performance evaluation methods described in Section. Thereafter, we apply class imbalance methods to review the effects on the model performance. We test and validate the models with the use of confusion matrixes, k-fold cross-validation, and the ROC curve.

The machine learning models are implemented by used a training set to build the build. Thereafter, we use the test set to validate the developed model. We start with highly interpretable and simple algorithms. From there on, we evaluate several boosting models.

5.1 Probability accuracy

Besides the classification of disability and non-disability, we are also interested in the predicted probability. Probability metrics quantify the skill of a model with the use of predicted probabilities rather than the class labels. To measure the accuracy of the predicted probabilities, we use the Brier Score (BS) [72]. The Brier Score measures squared distances of probabilities from the true class labels, therefore the mean squared error between the predicted probabilities and the observed values. Hence, the Brier Score indicates how accurate a prediction has been. The value of the BS is always between 0.0 and 1.0, where a model with perfect performance has a score of 0.0 and the worst model has a score of 1.0. The BS can be calculated with the following formula where f_t is the predicted value, o_t is the observed value, 1 if claim, 0 if non-claim, and N is the number of observations.

$$\text{BS} = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$

5.2 Comparing predictive models

Within predictive analytics, the focus is on predicting future events or outcomes. It is solely based on data-driven approaches and techniques to reach conclusions or solutions. We use analytical techniques and predictive modeling to find relevant patterns in the data set.

In this section, we compare the different machine learning algorithms based on the performance metrics: Accuracy and the area under the curve (AUC) of the ROC curve. First, we determine the cross-validation scores achieved after cross-validating to determine which models will be tested based on the achieved accuracy. Next off, we compare the base classifiers after re-balancing the data. Thereafter, we build and fit the models and compare the results with hyper-parameter tuning. The models are built-in Python.

5.2.1 K-Fold Cross-validation

Cross-validation is a way to select different “train” and “test” samples from a dataset. It is a way to validate whether a trained machine learning model performs well on new data. In K-Fold Cross-validation, we split the training data into k folds (folds are subsets of the total dataset). We use k-1 folds to train the model and validate with the remaining fold. Five to then folds is widely used in practice, we perform 5 Fold Cross Validation.

The mean accuracy and the standard deviation of the 5 Fold Cross Validation are displayed in Figure 3-8. Since the mean accuracy of the Naïve Bayes is low, we do not use this model further in our study. In Section 3.8 we have discussed different types of ensemble methods: Bagging, Stacking and Boosting. Since the bagging method, Random Forest is scoring low in comparison to the boosting methods, we do not consider bagging models further in this research. Moreover, we do not consider the ensemble method stacking due to the lack of interpretability.

The standard deviation is a measure of the spread of the distribution, it is the square root of the variance. The variance is the mean of the squared differences from the mean. The extent to which a new prediction is reliable is higher with a lower standard deviation. The lower the standard deviation, the more close the numbers are together. The three algorithms with the lowest standard deviation and hence with the lowest spread are the LR, GB, and XGB. We further examine these three models since these models have the highest accuracy and lowest standard deviation.

Table 5-1: Results 5-Fold Cross Validation

ML Models	Mean Accuracy – 5 fold	Standard Deviation Accuracy
Naïve Bayes (NB)	18.49%	0.0126
Logistic Regression (LR)	81.56%	0.0093
Decision Tree (DT)	72.12%	0.0102
Random Forest (RF)	80.31%	0.0114
AdaBoost (AB)	81.73%	0.0121
Gradient Boosting (GB)	81.90%	0.0093
XGBoost (XGB)	82.07%	0.0095
LightGBM (LGBM)	81.32%	0.0114

5.2.2 Evaluation

In this section, we train the models with the training set and evaluate the models in the testing set. We select the three models with the highest mean accuracy and the lowest standard deviation of 5 fold cross-validation (Table 5-1). Hence we examine the results of the LR, GB, and XGB.

5.2.2.1 Classifiers

In this research, the ROC curve is a graph that shows the ability of a test to distinguish between disabled and non-disabled individuals. In a ROC curve, at different cut-off values, the sensitivity of the test (true positives) on the y-axis is plotted against the false positives on the x-axis. The most optimal cut-off value is in the upper left corner of the curve (high proportion of true positives and low proportion of false positives). The 'Area under the Curve (AUC)' indicates the accuracy of the test: 1 is a perfect test, which can identify all policyholders without false positives, and 0.5 is a worthless test, which detects as many true positives as false positives. The AUC can be used to indicate the added value of a test.

The ROC curves of the LR, GB, and XBG models are plotted in Figure 5-1. The random prediction line does not discriminate between the classes and will predict a random class in all cases. All models are performing better than the random prediction. The XGBoost model is performing worst and the logistic regression (LR) has the best score in discriminating between classes when taking into account the ROC-AUC value. However, the differences are minimal when considering the AUC of the ROC curve. All models do perform better than the random model.

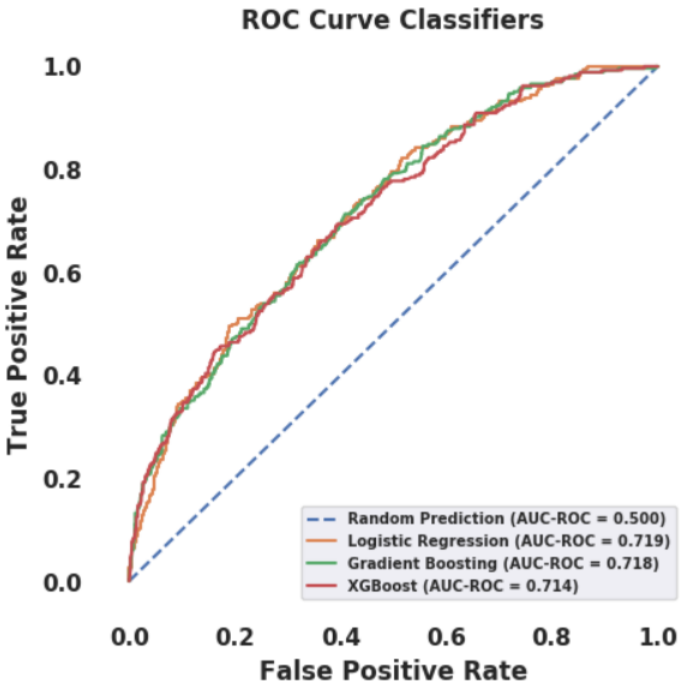


Figure 5-1: ROC-Curve with AUC of the LR, GB, and XGB

As mentioned in Section 5.2, we are also interested in the accuracy of the predicted probabilities. Therefore, we also analyze the Brier Score (BS). A Brier score of 0 means perfect probability accuracy, and a Brier score of 1 means perfect probability inaccuracy. So a lower Bier score indicates greater accuracy. Therefore, based on the predicted probabilities accuracy, the boosting methods are performing better. However, the differences with the Brier Score of the Logistic Regression are minimal. The Brier Scores are displayed in Table 5-2.

Table 5-2: Brier Scores

Classifiers	BS
LR	0.175
GB	0.173
XGB	0.173

5.2.3 Re-balancing Data

The analysis performed in Chapter 4 shows that there are unbalanced classes within the dataset. The number of disabled individuals (the minority) is much lower than not disabled individuals, Figure 5-2. When a prediction model is developed based on such data, it will be dominated by the contribution from data in the not disabled (the majority) class. The accuracy of the model will be better when predicting the “0” class versus the “1” class. However, we are more interested in predicting the “1” class. Therefore, there is a class imbalance problem.

Table 5-3 displays the proportions of the classes. The class of ‘disability’ is considerably smaller than the class of ‘not disability’ from the total long-term absenteeism events. The ratio is so skewed that we attempt to compensate for this to improve the training process of the ML models.

Table 5-3: Number of Disabled vs Not Disabled from long-term absenteeism

	Percentage
Long-term absenteeism disabled (minority)	17.86%
Long-term absenteeism not disabled (majority)	82.14%

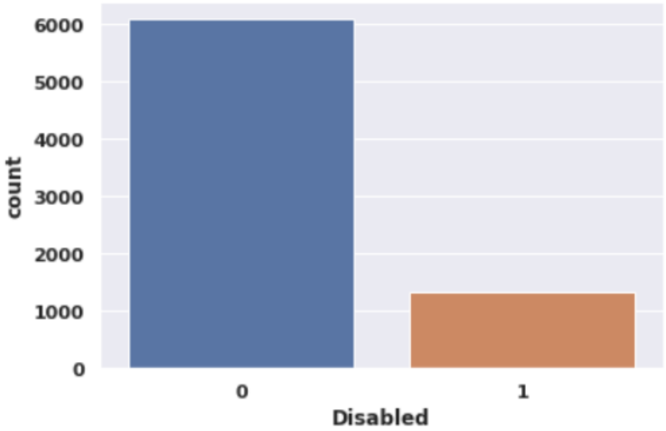


Figure 5-2: Visualization class imbalance

Unbalanced data is a well-known problem in the machine learning literature and has been studied for several decades. There are methods to reduce the bias of a certain class in the data, such as oversampling and undersampling. We re-balance the data by adding minority samples or randomly sampling from the majority samples. The aim is to examine whether re-balancing the data will improve the performance of the model. Hence, we evaluate the data re-balancing techniques by testing the Logistic Regression.

5.2.3.1 *Random undersampling*

Random Undersampling (RUS) is a technique wherein we randomly sample data records in the majority class so that they will match up to the total length of the minority class samples.

5.2.3.2 *Synthetic Minority Oversampling Technique*

With randomly oversampling, the minority points are copied, which might lead to over-fitting. As an over-sampling method, we apply SMOTE instead of randomly oversampling. As mentioned in Section 3.10.2, SMOTE creates new samples, it selects the minority class that is close and then draws lines between them, new sample points are located on these lines. To be more precise, a random sample is chosen and the K-Nearest Neighbor algorithm is used to select neighbors to which lines are drawn. We use SMOTE as the oversampling technique in this study. Appendix B shows the pseudo-code for the SMOTE algorithm [73].

To illustrate the possible impact of the re-balancing data techniques, we implement the techniques on the Logistic Regression model. Table 5-4. displays the results of both undersampling and SMOTE. Both the accuracy and the area under the ROC curve have decreased. Also, the BS has increased, which means a lower probability accuracy. Nonetheless, both the precision and recall increase after undersampling and oversampling the dataset. Therefore, based on the recall and the precision, it can be concluded that the models perform better after re-balancing the data. Within this research, the accuracy of the disability probability which is indicated by the Brier Score plays an important role. Hence, considering the Brier Score and the accuracy, the effect of data re-balancing is not preferred.

Table 5-4: Results class imbalance

	Accuracy	AUCROC	Precision	Recall	Brier
LR (original)	0.827	0.719	0.609	0.495	0.172
LR (undersampled)	0.646	0.713	0.679	0.556	0.354
LR (oversampled)	0.648	0.716	0.681	0.630	0.401

6. Implementation and Demonstration

In this chapter, we discuss the results of the implemented algorithms. Thereafter, we discuss the usefulness of the model in extraordinary cases with a practical example: COVID-19.

6.1 Evaluation of the algorithms

In this section, we summarize the results of the different ML models. To evaluate the performance of the models, we indicate the accuracy, area under the ROC curve, and the Brier Score. The higher the scores of the accuracy and the AUC and the lower the Brier score, the better the algorithm is performing. From Table 6-1**Error! Reference source not found.** it becomes clear that the Gradient Boosting and XGBoost perform best in terms of accuracy and the Brier score. The logistic regression has the highest AUC of the ROC curve. However, the differences are minimal. This shows that the three selected models could be a good match to use with the prediction of disability.

Table 6-1: Results of ML algorithms

Classifiers	Accuracy	AUC	BS
LR	81.56%	0.719	0.175
GB	81.90%	0.718	0.173
XGB	82.07%	0.714	0.173

In Section 5.2.1 we have seen that the Naïve Bayes and the Decision tree have a low accuracy score. One reason for the low score of the Naïve Bayes might be that we have a lot of categorical predictors. The Naïve Bayes models these as following a normal distribution. However, we have converted these categorical variables to ones and zeros and hence the model might throw away some useful information. Moreover, the Naïve Bayes assumes the features are independent, which is not always the case within our dataset (Section 4.3.5).

The other way around, the Decision Tree assumes that every feature interacts with each other with every variable further up the tree. The fact that we have features that do not interact with each other could explain the low performance of the Decision Tree.

If we consider the Random Forest, it is also an ensemble training method and it predicts by combining the outputs from individual trees. However, they differ in the way the trees are built (the order and the way the results are combined). The RF trains each tree independently with the use of a random sample from the data. In this way, the model is more robust and is less likely to overfit the training data in comparison with a single decision tree. On the other hand, the Gradient Boosting model builds trees one at a time. Where each new tree helps to correct errors made by the previously trained tree. Therefore, training the GBM takes longer since trees are built sequentially. However, results have shown that the GBM is a better method than the RF [74].

6.2 Advantages and disadvantages of the algorithms

There is ‘no free lunch’ in machine learning since no single machine learning algorithm is universally the best-performing algorithm for all problems [75]. Every algorithm has its advantages and disadvantages. In this section, we discuss the advantages and disadvantages of the three best-performing machine learning algorithms.

6.2.1 Logistic Regression

The Logistic Regression is the most useful for understanding the influence of several independent variables on a single dichotomous outcome variable. In general, logistic regression is a powerful algorithm. In contrast to linear regression, only several statistical conditions apply to logistic regression. For example, no assumptions need to be made about the distribution of the outcome variable. The predictors (or explanatory variables) can be either discrete or continuous. However, when large datasets are considered, the problem of overfitting might occur.

6.2.2 Gradient Boosting

The Gradient Boosting method has high flexibility since it has several hyperparameter tuning options. Moreover, the GB can optimize different loss functions. Gradient boosting also contributes to solving several multicollinearity issues where there are high correlations between prediction variables.

However, the GB is computationally expensive and due to the many trees it requires, it needs exhaustive memory. Also, the GB is prone to overfitting which we overcome with hyperparameter tuning.

6.2.3 Extreme Gradient Boosting

XGBoost, short for Extreme Gradient Boosting, is an advanced version of the Gradient Boosting method that is designed to focus on computational speed and model efficiency. The XGBoost is a specific implementation of Gradient Boosting, with several improvements that make the model faster than the traditional Gradient Boosting. However, just like other boosting methods, it is sensitive to outliers.

6.3 Feature Importance

The feature importance indicates how much the model considers a specific feature when classifying. The higher the importance, the more important the feature. It is calculated by comparing how the machine learning model performs when the data it represents is arranged and when this data is shuffled. The feature importance is determined for the logistic regression model, Figure 6-1.

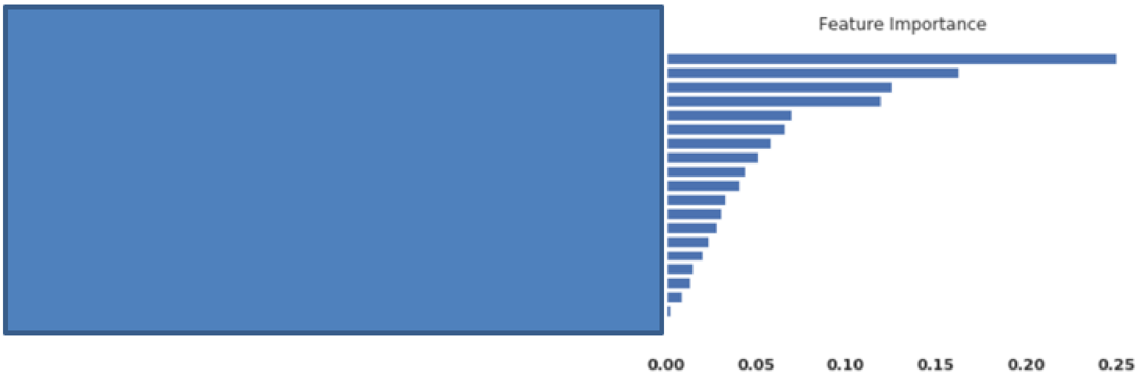


Figure 6-1: Feature Importance of the LR Model (hidden for privacy)

7. Conclusions and Recommendations

7.1 Conclusions

For the WIA insurance products, there are almost two years between the moment of becoming ill and whether or not to become disabled. For pricing purposes, this means the actuaries have to wait at least two years before comparing the forecasted disability probability with the actual realized WIA-inflow. Also for provisioning purposes, the waiting period is two years before basing the provisions for a specific claim year on actual WIA cases. With this research, we examined both the added value of machine learning algorithms and the use of 42nd-week report data.

Currently, the Disability Income Insurance of Achmea is not using the long-term absenteeism data to predict disability more accurately and make decisions based on more recent data. Moreover, the DII is not making use of any machine learning algorithm.

Consequently, we sought to decrease the two years waiting period and to increase the accuracy of the disability probability prediction by using the long-term absenteeism data (42nd-week reports) together with the implementation of machine learning.

To this end, we aimed to answer the main research question stating:

How can we increase the accuracy of the disability probabilities prediction based on 42nd- week reports data and machine learning techniques?

We tested eight distinct ML models to assess the prediction performance on the merged dataset. The intent is to use the probabilities in a pricing model. Hence, the classification of the disabled and non-disabled classes needs to be accurate.

In the current literature, there are only a few articles available concerning the tradeoff between prediction probabilities and the classification of the classes. To evaluate the prediction performance, we used the evaluation metrics. We used the frequently used ROC curve and the area under this curve. Moreover, we used the accuracy which can be extracted from the confusion matrix. We evaluated the predicted probabilities with the Brier Score metric.

We first assessed the accuracy of the ML models by K-fold cross-validation. Especially the Decision Tree model performed poorly regarding the accuracy of class separation. Eventually, we selected the three ML models (Logistic Regression, Gradient Boosting, and Extreme Gradient Boosting) with the highest accuracy performance by taking the standard deviation of the cross-validation into account.

Considering both the accuracy and the area under the ROC curve, the XGBoost has the highest performance. However, there is not a big difference between the Gradient Boosting and the Logistic Regression models. Both the XGBoost and the Gradient Boosting have equal Brier Scores. The Logistic Regression is performing slightly less regarding the probability accuracy.

Since the dataset is highly imbalanced (high non-disabled cases, low disabled cases), we performed literature research on data re-sampling methods. We balanced the data by both undersampling and oversampling the data and evaluated the result through the implementation of the Logistic Regression. To evaluate the usefulness of re-balancing the data, we performed Random Undersampling and as an Oversampling method, we used SMOTE for the Logistic Regression model. Both the re-balancing methods caused a decrease in the accuracy and the area under the ROC curve. This means that the total correct predictions have been reduced. Moreover, the Brier Score increase meaning the re-balancing methods have a negative impact on the predicted probabilities. However, both the precision and recall increased, which indicates that due to re-balancing the data, the accuracy of classification of positive incidents increased.

We also focused on the predictive power of the features. The three features with the highest predictive power, based on the Logistic Regression, are the disability percentage and the expectation of becoming disabled at the moment of 42-week illness, and the age of a long-term absenteeism employee. For the development of the models, these features must be stored as complete as possible in the database. Improving the data, especially the important features, can have a big effect on the performance of the ML models.

For dimensionality reduction, we performed factor analysis to reduce a large number of features into fewer numbers of factors which explain almost as much of the variation as the observed features. With the factor analysis, we reduced the 73 features to 54 factors.

We showed that in extraordinary circumstances, such as the current COVID-19 pandemic, the model cannot be used directly. Instead, the data of the 42nd-week reports, which were not used for the prediction of disability, could be used to make a hands-on analysis.

This thesis aimed to investigate the impact of machine learning on the insurance industry in combination with the use of long-term absenteeism data. In the literature study, we discussed how the insurance industry is expected to change in the future regarding the data-driven approach. This study can be seen as practical guidance for insurers on how to implement and execute data-driven opportunities.

7.2 Recommendations

We recommend using the Logistic Regression, the Gradient Boosting or the Extreme Gradient Boosting algorithm for the prediction of disability based on the 42nd-week reports. The accuracy and the AUC of these three models show that there is potential in using machine learning to predicting disability. Based on certain requirements, one of the three selected models should be chosen.

In addition to the accuracy, AUC, and the brier score of a model, other factors play a role in choosing a suitable model for implementation in an industrial environment. One of those factors is the extent to which extensive preprocessing of data is required. The intention is that the business can set up, maintain and use these models in a relatively simple way. Other factors that play a role are the calculation time when training the model.

First of all, since currently no machine learning models are used for the prediction of disability and the performance of the three selected ML models are comparable, we suggest the Logistic Regression when taking into account the interpretability. The LR is easy to implement, interpret, and efficient to train. Logistic Regression is simply the GLM when describing it in terms of the logit link function. Also, the computing time of the LR in comparison with boosting methods is low. Since the actuarial business is already quite familiar with the use of the GLM, implementing the LR will be more accessible in comparison with the boosting methods. Moreover, according to Occam's razor, we should apply the simplest solution with the fewest assumptions.

However, the Gradient Boosting and the Extreme Gradient Boosting have a lot of flexibility due to the various hyperparameters which can be tuned. Moreover, the data pre-processing is less exhaustive in comparison with the Logistic Regression.

According to, Laet [76], a Gradient Boosting method provides better estimates for claim frequencies than the Generalized Linear Model [76]. Unlike the GLM, the GB does not have the restrictions that explanatory variables have a linear relationship with the log of the claim frequency, resulting in higher flexibility. Moreover, in a GB, interactions between variables are made by the algorithm itself, while in a GLM it has to be added in the formula.

However, Gradient Boosting is computationally expensive since it requires many trees and it requires a large grid search during tuning due to the high number of parameters. Moreover, the boosting methods are less interpretative. The advantage of XGBoost over GB is that it controls over-fitting and has a more efficient approach regarding speed and memory utilization. Therefore, in comparison with gradient boosting, the XGBoost focuses more on computational power.

Hence, this research shows that the ML models can make reasonable predictions based on the data of long-term absenteeism. However, it is questionable whether the ML models work well enough to make these predictions fully automatically. A hybrid system where the actuary is assisted in the determination of disability probabilities by an ML model is probably the best solution. The benefits of applying ML in classifying disability would still apply if a hybrid system were used. Because all available relevant cases are taken into account by the ML model and the actuary then selects the correct final classification, it will probably also yield higher accuracy of the prediction.

The feature importance analysis shows that the disability percentage and the predetermined expectation of the WIA are relatively important for the classification. As a result, more attention would be given by the actuaries of 42nd-week reports with these indicators. The danger of a vicious cycle arises in this case since more disability cases will be predicted by the ML model which increases the machine bias. This can also be prevented by developing a hybrid system instead of a fully automatic system. In this way, there is always the supervision of an actuary who can make adjustments.

Certainly, there is no black and white choice among the traditional actuarial models and new machine learning algorithms. Applying machine learning in combination with domain knowledge provides a better result. This means that as an actuary one has an optimal position to make use of new data and techniques to arrive at better models and insights. Therefore, the algorithm will help to further shape the transition of the industry, with the highest relevance in the pricing of an insurance product.

The availability of the data is key in constructing an ML model since machine learning (almost) always benefits from more data. Hence, data is at the heart of any Machine Learning research. Currently, approximately 65% of the 42nd-week reports are reported, despite the fact it is stated as mandatory as stated in the policy conditions. Also, if reported, the reports are often too late or too early. Improving the process of reporting 42nd-week reports could realize an improvement in the prediction power of the model. Achmea does not yet actively request these reports from employers. To bring the percentage of the submitted 42nd reports to a decent level, Achmea could send messages to their customers who have not submitted 42nd-week reports (for a while) whether it is true that no 42nd-week reports have been made. Within this message, the importance of the reports could be explained to increase the 42nd-week reports and improve the usefulness of these notifications.

7.2 Discussion and further research

In this section we provide a discussion of the study performed and make recommendations for further research.

We developed ML models in this research based on the available feature within the dataset. An opposite approach could be realized by performing literature research on features that may have high predictive power in predicting disability from a scientific point of view. These features could be used in the ML model to increase performance. Possible features could be the geographical location such as the postal code, the type of illness, the number of hours working per week, and the state pension age. However, it must be taken into account that the insurance business is highly regulated which poses some challenges for collecting privacy-sensitive data.

With the use of 42nd-week reports, we can predict which companies have bad reporting behavior. This information can be taken into account in the portfolio development in the forecast of both pricing and provisions. Since the 42nd-week reports are not complete yet and the information is just currently known, it is not representative to make conclusions regarding the reporting behavior. However, in the future, the impact of the reporting behavior can be analyzed based on the 42nd-week reports.

The 42nd-week reports could also be used to initiate targeted actions for returning to work in time. In other words, to ensure that the sick employee recovers earlier (and thus ends up in the WIA for a shorter time or not at all). Quantifying this impact is an interesting issue. However, unfortunately, there is insufficient data available for this (yet).

A lot of research takes place in the ML domain. Every year there are new and improved ML models that can do better classifications. It is important to keep testing different ML models and to keep looking for improvements. In the literature study, we have described the machine learning algorithms which have been compared in this study. Since there is a high expectation there will be developed more algorithms that are a good match with the used dataset, future research could include these algorithms and compare them with the selection algorithms.

The machine learning algorithms in the comparison were determined through a literature study. It is, however, imaginable that a more suitable algorithm will be developed. For future research, it might be interesting to add these new algorithms to the comparison process. Hence, more algorithms could become available that could be a good match for the dataset.

Hence, machine learning has progressed a remarkable amount over the past few years. It seems that in each passing week new research and discoveries are published. Many factors are driving this rapid growth and expansion.

One particular area of research is the proliferation of Automated Machine Learning (AutoML) tools. AutoML covers all aspects of the machine learning workflow that can be potentially automated [77].

First of all, with AutoML, the type of data could be identified, therefore dividing the data types into Boolean, discrete, continuous, or text for example. Moreover, with AutoML, task detection can be automated. In other words, based on the dataset, it can be determined whether it is a binary classification, regression, or clustering for example. The next step is feature engineering which is a time-consuming process. With AutoML, for instance, feature selection can be automated. Other steps within the machine learning project e.g. model selection, hyperparameter optimization, model interpretation, and prediction analysis can also be automated with AutoML.

Many companies do have domain experts. However, concerning machine learning algorithms and the parameters, they may not have expertise. Hence, Automated Machine Learning can be used to implement the machine learning project despite the lack of thorough data science expertise.

References

[1] „Data Science vs. Machine Learning,” Master in Data Science, 2021. [Online].

[2] M. M. X. P. S. W. J.-F. Denuit, „ Actuarial modelling for claim counts: Risk classification, credibility and bonus-malus systems,” *Chichester: John Wiley & Sons, Ltd. ,* 2007.

[3] J. Brownie, „Ensemble Learning Algorithm Complexity and Occam's Razor,” Machine Learning Mastery, 21 12 2020. [Online].

[4] R. C. M. A. K. a. V. R. Henckaerts, „Boosting insights in insurance tarieff plans with tree-based machine learning methods”.*Faculty of Economics and Business.*

[5] W. Q. a. H. Z. Y. Yang, „Insurance Premium Prediction via Gradient Tree-Boosted Tweedie Compound Poisson Models [Online],” 2016. [Online]. Available: https://arxiv.org/pdf/1508.06378.pdf.

[6] X. &. H. S. Yang, „Classification of risk to support decision-making in hazardous processes,” *Safety Science,* nr. doi:https://doi.org/10.1016/j.ssci.2015.07.011, pp. 80, 115-126, 2015.

[7] W. L. Y. J. P. S. S. R. D. A. D. E. V. V. a. A. R. Christian Szegedy, „Going Deeper with Convolutions,” p. arXiv:1409.4842, 2014.

[8] Y. C. N. W. a. Z. Z. Yanghao Li, „Scale-Aware Trident Networks for Object Detection,” *arXiv preprint arXiv:1901.01892,* 2019.

[9] Adfiz, „Arbeidsongeschiktheid. Wat zijn de financiële gevolgen voor u?,” Geijsel Kroon, 2021. [Online]. Available: https://www.geijselkroon.nl/december-2014/arbeidsongeschiktheid.

[10] „Hoe groot is de kans dat ik arbeidsongeschikt word?,” AOV, [Online]. Available: https://doehetzelfaov.nl/hoe-groot-is-het-arbeidsongeschiktheidsrisico/.

[11] RIZIV, „Statistieken over invaliditeit van werknemers en werklozen in 2014,” 2016. [Online]. Available: https://www.inami.fgov.be/nl/statistieken/uitkeringen/2014/Paginas/statistieken-invaliditeit.aspx .

[12] K. A. a. S. C. L. S. Lee, „Why High Dimensional Modeling in Actuarial Science?,” 2015. [Online]. Available: https://pdfs.semanticscholar.org/ad42/ c5a42642e75d1a02b48c6eb84bab87874a1b.pdf.

[13] K. T. T. R. M. A. &. C. S. Peffers, „A Design Science Research Methodology for Information Systems Research,” *Journal of Management Information Systems,* pp. 45-77, 2007.

[14] K. Peffers, T. Tuunanen, M. Rothenberger en S. Chatterjee, „A Design Science Research Methodology for Information System Research,” *Journal of Management Information Systems,* pp. 45-78, 2008.

[15] G. L. Geerts, „ A design science research methodology and its application to accunting information systems research,” *Elsevier,* pp. pp. 142-151, 2011.

[16] S. T. M. a. V. C. Storey, „Design Science in the Information Systems Discipline: An Introduction to the Special Issue on Design Science Research,” *Jstor,* pp. 725-730, 2008.

[17] J. P. M. a. L. Massaron, Machine Learning for Dummies, vol. 53, no. 9. , 2013.

[18] „Ethich Kader datatoepassingen,” Verbond Van Verzekeraars, 1 1 2021. [Online]. Available: https://www.verzekeraars.nl/branche/zelfreguleringsoverzicht-digiwijzer/ethisch-kader-datatoepassingen. [Geopend 10 5 2021].

[19] M. D. a. J. G. D. Autor, „Moral Hazard and Claims Deterrence in Private Disability Insurance,” *National Bureau of Economic Research, Cambridge, MA, Tech.,* nr. [Online]. Available: http://www.nber.org/papers/w18172.pdf, pp. Rep., 6, 2012.

[20] I. R. Z. a. L. E. Beedon, „Long-Term Disability Programs in Selected Countries,” *Social Security Bulletin,* pp. Vol. 50, No.9, 1987.

[21] „Zilveren Kruis,” WIA Loongrens per 1 januari, 2021. [Online]. Available: [Online]. https://www.zilverenkruis.nl/zakelijk/producten-en-diensten/wettelijke-bedragen .

[22] UWV, „Wat is sv-loon? - UWV - Voor Particulieren,” 2018. [Online]. Available: https://www.uwv.nl/particulieren/veelgestelde-vragen/actueel/detail/wat-is-sv-loon.

[23] UWV, „WIA Loongrens per 1 januari 2021,” [Online]. Available: https://www.zilverenkruis.nl/zakelijk/producten-en-diensten/wettelijke-bedragen. [Geopend 8 2 2021].

[24] „Over het Verbond,” Verbond van Verzekeraars, 2021. [Online]. Available: https: //www.verzekeraars.nl/het-verbond/over-het-verbond.

[25] V. v. Verzekeraars, "Nieuw rapport Kansenstelsel WGA-ERD 2019," 2019. [Online]. Available: https://www.verzekeraars.nl/academy/activiteitenoverzicht/httpssamenwerkenverzekeraarsnlactueelevenementenpaginasmarktbijeenkomst-kansenstelsel-wga-erd-2019-9-april-2019aspx. [Accessed 2 10 2021].

[26] Atul, „AI vd Machine Learning vs Deep Learning,” Edureka!, 28 7 2020. [Online]. Available: https://www.edureka.co/blog/ai-vs-machine-learning-vs-deep-learning/. [Geopend 25 3 2021].

[27] B. D. a. B. Löfdahl, „A hidden Markov Approach to Disability Insurance,” *Cornell University,* 9 2 2020.

[28] L. K. ´. s. a. M. N. V. Kasˇcelan, „A nonparametric data mining approach for risk prediction in car insurance: a case study from the Montenegrin market,” *Economic Research-Ekonomska Istrazivanja,* pp. vol. 29, no. 1, pp. 545–558.

[29] Guelman, „Gradient boosting trees for auto insurance loss cost modeling and prediction,” *Expert Systems with Applications,* pp. 39(3):3659-3667, 2012.

[30] A. D. C. P. L. Spedicato, „Machine learning methods to perform pricing optimization. A comparison with standard GLMs,” *Variance,* pp. 12(1):69-89, 2018.

[31] Y. a. X. W. Wang, „Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud,” *Decision Support Systems,* pp. 105:87-95, 2018.

[32] A. E. a. H. S. Renshaw, „On the graduation associated with a multiple state model for permanent health insurance,” *Insurance, Mathematics and Economics,,* pp. pp. 1-17, 1995.

[33] O'Neil, Weapons of math destruction: How big data increases inequality and threatens democracy., Broadway Books, 2017.

[34] „Breaking the 80/20 rule: How data catalogs transform data scientists' productivity," IBM, [Online]. Available: https://www.ibm.com/cloud/blog/ibm-data-catalog-datascientists-productivity. [Geopend 17 3 2021].

[35] R. Kress, „A board primer on artifical intelligence," Cyber Resilient Business, 22 10 2019. [Online]. Available: https://www.accenture.com/hu-en/insights/security/board-primer-artificial-intelligence. [Geopend 15 3 2021].

[36] J. P. M. a. L. Massaron, in *Machine Learning for Dummies*, 2013, pp. vol. 53, no. 9..

[37] H. Heidenreich, „What are the types of machine learning?," Towards Data Science, 4 12 2018. [Online]. Available: https://towardsdatascience.com/what-are-the-types-of-machine-learning-e2b9e5d1756f.

[38] T. M. Mitchell, Machine Learning, Boston: McGraw-Hill, 1997.

[39] M. E. C. a. K. Aydin, „Unsupervised learning algorithms," 2016.

[40] W. L. Y. J. P. S. S. R. D. A. D. E. V. V. a. A. R. Christian Szegedy, „Going Deeper with Convolutions," vol. arXiv:1409.4842 , 2014.

[41] I. B. a. C. Mues, „An experimental comparison of classification algorithms for imbalanced credit scoring data sets," *Exert Systems with Applications,* pp. vol. 39, no. 3, pp. 3446-3453, 2012.

[42] H. C. a. D. R. Hyeongjun Kim, „Corporate Default Predictions Using Machine Learning: Literature Review," *Sustainability,* p. p. 5, 6 8 2020.

[43] A. W. A. a. A. A. Alkarkhi, Factor Analysis, Easy Statistics for Food Science with R, 2019.

[44] A. Bhande, „What is underfitting and overfitting in machine learning and how to deal with it," Greyatm, 2018. [Online]. Available: https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76. [Geopend 26 2 2021].

[45] „ML – Python (VII) – Overfitting & Underfitting – Binary Coders," [Online]. Available: https://binarycoders.dev/2019/10/17/ml-python-vii-overfitting-underfitting/. [Geopend 25 3 2021].

[46] A. C. M. a. S. Guido, Introduction to Machine Learning with Python, 2016.

[47] P. Menon, „Data Science Simplified Part 12: Resampling Methods," Towards data science, 2019. [Online]. Available: https://towardsdatascience.com/data-science-simplified-part-12-resampling-methods-e029db77fa9c. [Geopend 3 15 2021].

[48] J. Brownlee, „What is the Difference Between a Parameter and a Hyperparameter?," Machine Learning Mastery, 2017. [Online]. Available: https://machinelearningmastery.com/difference-between-a-parameter-and-a-hyperparameter/. [Geopend 15 2 2021].

[49] "3.1. Cross-validation: evaluating estimator performance — scikit-learn 0.22.2 documentation," [Online]. Available: https://scikitlearn.org/stable/modules/cross_validation.html.. [Accessed 28 3 2021].

[50] T. T. R. a. F. J. Hastie, The Elements of Statistical Learning, Springer, 2017.

[51] G. W. D. H. T. T. R. James, An Introduction to Statistical Learning, New York: Springer, 2013.

[52] J. P. B. H. J. a. E. R. Lammers, Categorische data analyse met SPSS: inleiding in loglineaire analysetechnieken, Assen: Van Gorcum, 2007.

[53] H. Rajput, „MachineX: Simplifying Logistic Regression," Knóldus, 28 3 2018. [Online]. Available: https://blog.knoldus.com/machinex-simplifying-logistic-regression/. [Geopend 10 5 2021].

[54] M. Mayer, „Basics and Linear Models," 11 09 2020. [Online].

[55] X. Meng, „Generalized Linear Models in Spark MLLib and SparkR," Databricks, 17 02 2016. [Online].

[56] „Differentiate between parametric and nonparametric statistical analysis," [Online]. Available: http://agrimetsoft.com/faq/DifferentiatieBetweenParametrixAndNonParametric . [Geopend 19 3 2021].

[57] A. Ethem, Introduction to Machine Learning, Third., The MIT Press, 2014.

[58] D. C. e. al., „Documentation scikit-learn: machine learning in Python — scikitlearn 0.16.1 documentation," 2014. [Online]. Available: https://scikitlearn.org/0.16/documentation.html. [Geopend 25 3 2021].

[59] C. Mc, „Why Boosting Works?," TowardsDataScience, 4 4 2021. [Online]. Available: https://towardsdatascience.com/gradient-boosting-is-one-of-the-most-effective-ml-techniques-out-there-af6bfd0df342.

[60] N. J. a. M. Shah, „Evaluating learning algorithms: a classification perspective," *Cambridge University Press,* 2011.

[61] „Confusion Matrix," Everything About Data Science, 2016. [Online]. Available: https://scaryscientist.blogspot.com/2016/03/confusion-matrix.html?view=classic . [Geopend 18 3 2021].

[62] S. Narkhede, „Understanding AUC - ROC Curve," Towards Data Science, 28 6 2018. [Online]. Available: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5. [Geopend 11 4 2021].

[63] J. M. a. K. T. M. Johnson, „Survey on deep learning with class imbalance," *Journal of Big Data,* p. 6(1):27, 2019.

[64] W. Badr, „Having an Imbalanced Dataset? Here Is How You Can Fix it.," Towards Data Science, 22 2 2019. [Online]. Available: https://towardsdatascience.com/having-an-imbalanced-dataset-here-is-how-you-can-solve-it-1640568947eb . [Geopend 13 4 2021].

[65] N. V. C. e. al., „SMOTE: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research,* pp. 321-357, 2002.

[66] B. H. a. K. Chawla, „SMOTE: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research,* pp. vol. 16, pp. 321-357, 2002.

[67] K. B. L. H. W. K. N. Chawla, „SMOTE: Synthetic Minority Oversampling Technique," *Journal of Artificial Intelligence Research,* 2002.

[68] A. B. P. G. P. Bruce, Practical Statistics for Data Scientists, United States of America: O'Reilly , 2020.

[69] Y. X. H. K. E. C. D. E. E. A. M. T. a. R. G. P. Gromski, „Influence of Missing Values Substitutes on Multivariate Analysis of Metabolomics Data,” *Metabolites,* pp. vol. 4, no. 2, pp. 433-452, 6 2014.

[70] S. B. a. K. Groothuis-Oudshoorn., „mice: Multivariate imputation by chained equations in r,” *Journal of statistical software,* p. 45(3), 2011.

[71] J. a. D. W. Kaufman, „Determinng the number of factors to retain: A Window Based Fortnan-IMSL program for parallel analysis,” *Behaviour Research Methods, Instruments and Computers,* pp. 389-395, 2000.

[72] T. F. G. a. C. N. V. Raeder, „Learning from Imbalanced Data: Evaluation Matters,” *Springer, Berlin, Heidelberg,* pp. pp. 315-331, 2012.

[73] K. H. L. a. K. W. Chawla, „SMOTE: Synthetic Minority Oversampling Technique,” *Journal of Artificial Intelligence Research,* 2002.

[74] J. a. A. M. Bradley, „Random Forests and Boosting in MLlib,” Databricks, 21 01 2015. [Online].

[75] A. Mavuduru, „What "No free lunch" really means in machine learning,” Towards Data Science, 12 11 2020. [Online].

[76] B. d. Laet, „Regression trees and ensembles of trees in P&C pricing,” KU Leiven master-thesis, 2013-2014.

[77] X. Z. K. a. C. X. He, „AutoML: A survey of the state-of-the-art,” *Elsevier,* 5 1 2021.

[78] K. A. R. M. Son Dao, 18 January 2017. [Online]. Available: https://link.springer.com/article/10.1007%2Fs40092-017-0183-0.

[79] F. S. L. P. F. R. C. C. Rosa, September 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S235197891730745X?via%3Dihub.

[80] A. Azizi, July 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2351978915000281?via%3Dihub.

[81] J. M. Rohani, 2 July 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2351978915000037?via%3Dihub.

[82] [Online]. Available: https://www.theleansixsigmacompany.nl/dmaic-model/.

[83] [Online]. Available: https://bureautromp.nl/dmaic/.

[84] 2015. [Online]. Available: https://nl.wikipedia.org/wiki/DMAIC.

[85] J. R. Veldof, „Data driven decisions: using data to inform process changes in libraries,” *Library & Information Science Research,* Vols. %1 van %2doi:https://doi.org/10.1016/S0740-8188(99)80004-8, pp. 21(1), 31-46, 1999.

[86] N. D. M. L. S. &. K. T. Klein, „Nonlife ratemaking and risk management with Bayesian generalized additive models for location, scale, and shape,” *Insurance: Mathematics and Economics,* pp. 55, 225–249, 2014.

[87] R. C. T. R:, „A Language and Environment for Statistical Computing,” VR Foundation for Statistical Computing, 2014. [Online]. Available: http://www.R-project.org/.

[88] T. S. Review, „Pros and Cons of a BI Solution for Data Visualization," 2017. [Online]. Available: https://www.sam-solutions.com/blog/tableau-software-review-pros-and-cons-of-a-bi-solution-for-data-visualization/.

[89] Q. S. A. a. Limitations, „ Explore the Pros & Cons," 2019. [Online]. Available: https://data-flair.training/blogs/qlik-sense-advantages-and-limitations/.

[90] M. P. B. P. a. Con, „AbsentData,," 2019.

[91] Interpolis, „Crystal ClearBusiness Model & Brand Strategy Business," MBA-FSI , Tilburg.

[92] R. R. A. &. M. B. Hoffman, „What Is Design in the Context of Human-Centered Computing?," *Intelligent Systems,,* pp. 89-95, 2004.

[93] Y. Reich, „The study of design research methodology," *Mechanical Design,* 1995.

[94] S. T. a. S. G. March, „Design and Natural Science Research on Information Technology," *Decision Support Systems,* pp. pp. 251-266, 1995.

[95] A. v. L. B. v. d. W. J. d. Heerwaarden, „Schadeverzekering 1: Syllabus voor module AN 17 van de opleiding tot Actuarieel Analist," Actuarieel Instituut, 2012.

[96] E. H. S. Pitacco, „Actuarial Models for Disability Insurance," *Chapman & Hall/ CRC,* 1999.

[97] J. Grosfeld, „De voorspelbaarheid van de individuele verzuimduur," *Swets & Zeitlinger BV,* 1998.

[98] P. Smulders, „Balans van 30 jaar ziekteverzuimonderzoek," NIPG-TNO, Leiden, 1984.

[99] R. &. S. D. Rigby, „A Flexible Regression Approach Using GAMLSS in R," 2010. [Online]. Available: http://www.gamlss.org/wp-content/uploads/ 2013/01/book-2010-Athens1.pdf . [Geopend 24 March 2021].

[100] J. &. W. R. Nelder, „Generalized Linear Models," *Journal of the Royal Statistical Society,* pp. 135 (3), 370–384, 1972.

[101] E. &. J. B. Ohlsson, „Non-life Insurance Pricing with Generalized Linear Models," *Springer, Heidelberg,* 2010.

[102] P. &. N. J. McCullagh, „Generalized Linear Models," *Londen: Chapman & Hall,* 1989.

[103] R. G. M. D. J. &. D. M. Kaas, Modern Actuarial Risk Theory – Using R., Berlin: Springer, 2008.

[104] J. Starkweather, „Bayesian generalized linear models in r," *Benchmarks RSS Matters,* p. 29, 2011.

[105] J. &. W. R. Nelder, „Generalized Linear Models. Journal of the Royal Statistical Society," pp. Vol. 135, 370-384, 1972.

[106] M. &. L. S. Denuit, „Non-life rate-making with Bayesian GAMs," *Insurance: Mathematics and Economics,* pp. 35, 627–647, 2004.

[107] J. H. Friedman., „Greedy function approximation: a gradient boosting machine. Annals of statistics," p. pages 1189–1232, 2001.

[108] R. (. .. Rojas, Neural networks: a systematic introduction, Springer Science & Business Media, 2013.

[109] H. Heerkens, Geen Probleem: Een aanpak voor alle bedrijfskundige vragen en mysteries, Nieuwegein: Van winden Communicatie, 2012.

[110] Achmea, „Organisatiestructuur, medewerkers en informatie afdeling BFTI," 2021. [Online]. Available: https://intranet.achmeanet.nl/ . [Geopend 2021 2 1].

[111] O. M. a. L. Rokach, „Introduction to Knowledge Discovery and Data Mining," Data Mining and Knowledge Discovery Handbook. Boston, MA: Springer US, 2009. [Online]. Available: http://link.springer.com/10.1007/978-0-387-09823-4_1. [Geopend 5 3 2021].

[112] I. B. a. C. Mues, „An experimental comparison of classification algorithms for imbalanced credit scoring data sets," *Expert Systems with Applications,* pp. vol. 39, no. 3 pp. 3446-3453, 2012.

[113] J. Suykens, Artificial neural networks, 2011.

[114] A. Géron, „Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems," 2017.

[115] I. G. a. A. Elisseeff, „An introduction to variable and feature selection," *The Journal of Machine Learning Research,* p. 1157–1182, 2003.

[116] R. K. K. P. e. a. George H John, „Irrelevant features and the subset selection problem," *In Machine learning: proceedings of the eleventh international conference,* p. 121–129, 1994.

[117] S. L. Salzberg, „On comparing classifiers: Pitfalls to avoid and a recommended approach. Data mining and knowledge discovery," p. 1(3):317–328, 1997.

[118] J. M. B. a. D. G. Altman, „Multiple significance tests: the bonferroni method," *Bmj,* p. 310(6973):170, 1995.

[119] „Biological Neuron versus Artigicial Neural Network," ResearchGate, [Online]. Available: https://www.researchgate.net/figure/Biological-Neuron-versus-Artificial-Neural-Network_fig4_325870973. [Geopend 26 3 2021].

[120] „Interpret the key results for Correlation - Minitab Express," [Online]. Available: https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modelingstatistics/regression/how-to/correlation/interpret-the-results. [Geopend 24 3 2021].

[121] A. &. J. H. V. Labrinidis, „Proceedings of the VLDB Endowment," *Proceedings of the VLDB Endowment,* pp. 5(12), 2032–2033, 2012.

[122] A. &. H. M. Gandomi, „Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management,* pp. 35, (2), p.137-144, 2015.

[123] S. K. Cronin, „Deriving Data," Correlation, 2021. [Online]. Available: https://siobhankcronin.com/notes/deriving_data/correlation/ . [Geopend 31 3 2021].

[124] E. Huizingh, „SPSS 12.0 voor Windows en Data Entry," *Academic Service,* 2004.

[125] O. M. a. L. Rokach, Introduction to Knowledge Discovery and Data Mining in Data Mining and Knowledge Discovery Handbook, Boston: MA: Springer US, 2009.

[126] S. e. al., „A survey on generative adversarial networks for imbalance problems in computer vision tasks," *Journal of Big Data,* pp. https://doi.org/10.1186/s40537-021-00414-0 , 2021.

[127] L. e. a. Hyland, „Real-Value (Medical) Time Series Generation With Recurrement Conditional GANs".

[128] G. A. G. K. Fekri M., „Generating Energy Data for Machine Learning with Recurrent Generative Adversial Networks,” *MDPI Energies,* 26 December 2019.

[129] R. Shaikh, „Feature Selection Techniques in Machine Learnign with Pyton,” Towards Data Science, 28 10 2018. [Online]. Available: https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e. [Geopend 13 4 2021].

[130] A. Harsha, „AI vs Machine Learning vs Deep Learning,” Edureka, 28 7 2020. [Online]. Available: https://www.edureka.co/blog/ai-vs-machine-learning-vs-deep-learning/. [Geopend 23 4 2021].

[131] L. Breiman, „Machine Learning Random Forests,” *Statistics Department,* pp. 5-32, 2001.

[132] D. D. Lewis, Naive (Bayes) at forty: The independence assumption in information retrieval, Berlin Heidelberg: Springer , 1998, pp. 4-15.

[133] R. C. a. A. Niculescu-Mizil, „An empirical compariosn of supervised learning algorithms,” *ACM: In Proceedings of the 23rd international conference on Machine Learning,* pp. 161-168, 2006.

[134] J. C. M. J. C. a. G. I. W. Nayyar A. Zaidi, „Alleviating Naive Bayes attribute independence assumption by attribute weighting,” *The Journal of Machine Learning Research,* p. 14(1), 2013.

[135] L. Y. J. S. G. M. H. Y. a. G. B. G. Haixiang, „Learning from class-imbalanced data: Review of methods and applications,” *Expert Systems with Applications,* pp. vol. 73, pp. 220-239, 5 2017.

[136] R. K. K. P. George H John, „Irrelevant features and the subset selection problem.,” *In Machine Learning: proceedings of the eleventh international conference,* pp. 121-129, 1994.

[137] I. G. a. A. Elisseeff, „An Introduction to Variable and Feature Selection,” *Journal of Machine Learning Research,* 2003.

[138] X. a. W. M. Chawla, „Editorial: special issue on learning from imbalanced data sets,” *ACM Sigkdd Explorations Newsletter ,* pp. no. 1 (2004): 1-6.

[139] M. Kuhn, „Recursive Feature Elimination,” [Online]. Available: https://topepo.github.io/caret/recursive-feature-elimination.html. [Geopend 26 2 2021].

[140] S. Kotsiantis, „Supervised Machine Learning: A review of classification techniques,” *Imformatica,* pp. pp. 249-268, 2007.

[141] J. Brownlee, „Tour of Evaluation Metrics for Imbalanced Classification,” Machine Learning Mastery, 8 2 2020. [Online]. Available: https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/ . [Geopend 25 5 2021].

[142] R. C. a. P.-R. C. w. I. Classification. [Online].

[143] F. J. E.-A. M. Boughorbel S., „Optimal classifier for imbalanced data using Mattews Correlation Coefficient Metric,” *PLOS ONE,* 3 1 2017.

[144] N. S. M. Japkowicz, „Evaluating learning algorithms: a classification perspective.,” *Cambridge University Press,* 2011.

[145] S. Glen, „Brier Score: Definition, Examples," Statistics How To: Elementary Statistics , 2020. [Online]. Available: https://www.statisticshowto.com/brier-score/. [Geopend 2 5 2021].

[146] J. Brownlee, „A Gentle Introduction to Probability Metrics for Imbalanced Classification," Machine Learning Mastery, 10 1 2020. [Online]. Available: https://machinelearningmastery.com/probability-metrics-for-imbalanced-classification/. [Geopend 25 5 2021].

[147] Z. Lateef, „A Comprehensive Guide To Boosting Machine Learning Algorithms," Edureka!, 28 6 2019. [Online]. Available: https://www.edureka.co/blog/boosting-machine-learning/. [Geopend 24 4 2021].

[148] Wikipedia, „COVID-19 Pandamic in the Neterlands," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/COVID-19_pandemic_in_the_Netherlands.

[149] B. Boehmke, „Gradient Boosting Machines.," US Business Analytics R Programming Guide, [Online]. Available: http://uc-r.github.io/gbm_regression#proscons.

Appendices

Appendix A: Data Description

In this Appendix, we list the available variables in the individual datasets with a brief description and the data format.

A.1 Policy Data

Table A-1: Variables Policy Data

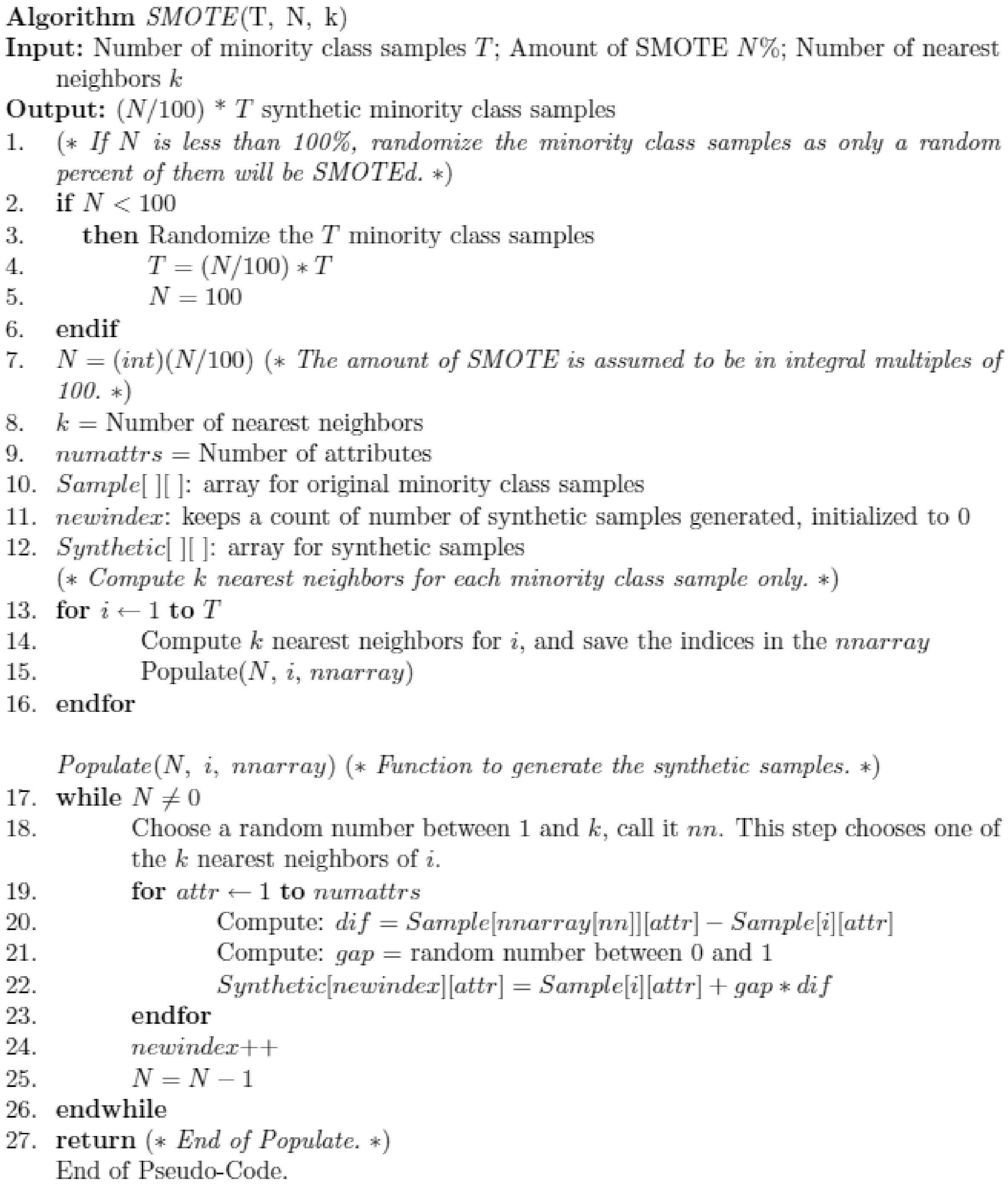
	Variable	Description	Data format/Values
1	Polisnummer	Contract number	Integers
2	Basisproduct	Type of WIA insurance product, select "Uitstap"	String
3	Werknemernummer	Employee number	Integers
4	Ingangsdatum	Start date WIA insurance	Date
5	Einddatum	End date WIA insurance	Date
6	Verzekeringsjaar	Year of WIA insurance contract	Integer, representing year
7	AantalVerzekeringsjaren	Length of insurance in one year	Float, representing proportion of one year
8	Polisingangsdatum	Starting date of the contract	Date
9	Poliseinddatum	Enddate of the contract	Date
10	Bedrijfsnaam	Name of the company where the employee is employed	String
11	Postcode	Postal code of the company	Integers
18	Sector	The sector of the company where the employee is employed	String
19	SectorCode	Code of the company sector	Integers
41	Klantpremie	Premium	Float
47	EindleeftijdUitkering	Stage pension age	Integers

49	PSN	Unique ID per policyholder	Integers
50	Geslacht	Gender of employee	String
51	Geboortedatum	Birth date of employee	Date
52	Indienstdatum	Employment date	Date
53	Uitdienstdatum	Unemployment date	Date
54	Dienstverband	Type of employment	String
55	Brutojaarloon	Gross annual salary	Float
56	Verzekerdloon	Insured wage	String

Table A-2: Variables Absenteeism data

	Variable	Description	Data format
1	BedrijfsID	Company ID	Integer
2	BedrijfsNaam	Company name	String
3	Polisnummer	Policy ID number	Integer
4	Ingangsdatum	Startdate policy contract	Date
5	Einddatum	Enddate policy contract	Date
6	PSN	Personal Number	Integer
7	Datumziek	Date of illness	Integer
8	Schadejaar	Claimyear	Integer
9	Melddatum 42eweeksmelding	Reporting date 42nd week report	Date
10	AOPerc	Disability percentage	Float
11	Status Claim	The state of the claim	String
12	Datum volledig hersteld	Date fully recovered	Date
13	Herstelreden	Recovery reason	String

Appendix B: Pseudo Code for the SMOTE algorithm

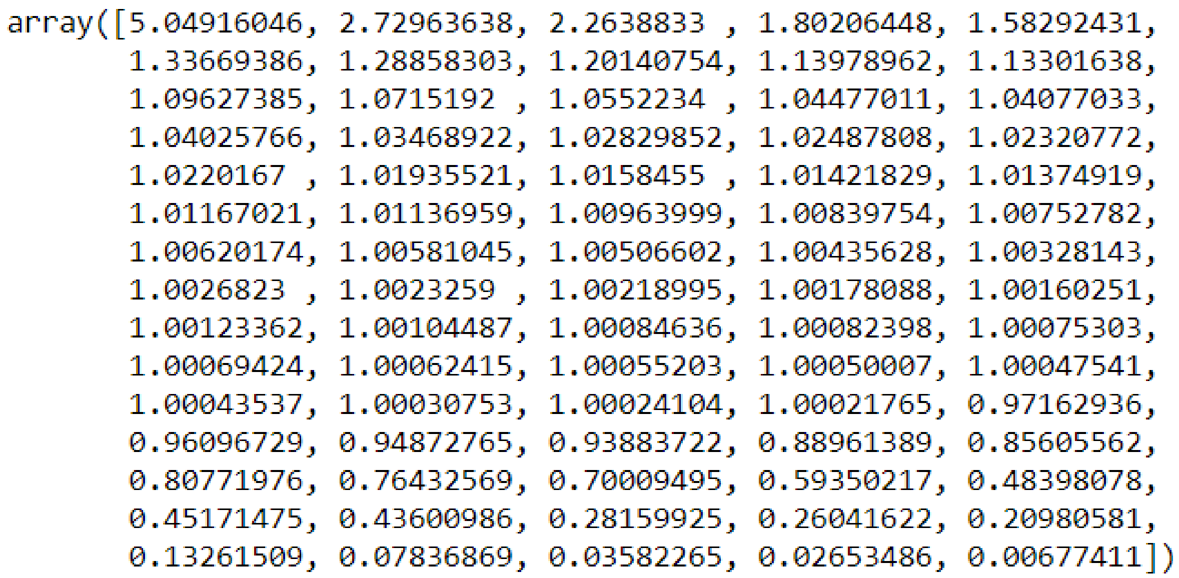


```
**Algorithm** $SMOTE$(T, N, k)
**Input:** Number of minority class samples $T$; Amount of SMOTE $N\%$; Number of nearest
    neighbors $k$
**Output:** $(N/100)$ * $T$ synthetic minority class samples
(* If $N$ is less than 100%, randomize the minority class samples as only a random
percent of them will be SMOTEd. *)
if $N < 100$
  then Randomize the $T$ minority class samples
       $T = (N/100) * T$
       $N = 100$
endif
$N = (int)(N/100)$ (* The amount of SMOTE is assumed to be in integral multiples of
100. *)
$k$ = Number of nearest neighbors
$numattrs$ = Number of attributes
$Sample[\,][\,]$: array for original minority class samples
$newindex$: keeps a count of number of synthetic samples generated, initialized to 0
$Synthetic[\,][\,]$: array for synthetic samples
(* Compute $k$ nearest neighbors for each minority class sample only. *)
for $i \leftarrow 1$ to $T$
        Compute $k$ nearest neighbors for $i$, and save the indices in the $nnarray$
        Populate($N$, $i$, $nnarray$)
endfor

Populate($N$, $i$, $nnarray$) (* Function to generate the synthetic samples. *)
while $N \neq 0$
        Choose a random number between 1 and $k$, call it $nn$. This step chooses one of
        the $k$ nearest neighbors of $i$.
        for $attr \leftarrow 1$ to $numattrs$
                Compute: $dif = Sample[nnarray[nn]][attr] - Sample[i][attr]$
                Compute: $gap$ = random number between 0 and 1
                $Synthetic[newindex][attr] = Sample[i][attr] + gap * dif$
        endfor
        $newindex$++
        $N = N - 1$
endwhile
return (* End of Populate. *)
End of Pseudo-Code.
```

Appendix C: Eigenvalues

As visualized in this table, there are 54 factors with an eigenvalue above one.



```
array([5.04916046, 2.72963638, 2.2638833 , 1.80206448, 1.58292431,
       1.33669386, 1.28858303, 1.20140754, 1.13978962, 1.13301638,
       1.09627385, 1.0715192 , 1.0552234 , 1.04477011, 1.04077033,
       1.04025766, 1.03468922, 1.02829852, 1.02487808, 1.02320772,
       1.0220167 , 1.01935521, 1.0158455 , 1.01421829, 1.01374919,
       1.01167021, 1.01136959, 1.00963999, 1.00839754, 1.00752782,
       1.00620174, 1.00581045, 1.00506602, 1.00435628, 1.00328143,
       1.0026823 , 1.0023259 , 1.00218995, 1.00178088, 1.00160251,
       1.00123362, 1.00104487, 1.00084636, 1.00082398, 1.00075303,
       1.00069424, 1.00062415, 1.00055203, 1.00050007, 1.00047541,
       1.00043537, 1.00030753, 1.00024104, 1.00021765, 0.97162936,
       0.96096729, 0.94872765, 0.93883722, 0.88961389, 0.85605562,
       0.80771976, 0.76432569, 0.70009495, 0.59350217, 0.48398078,
       0.45171475, 0.43600986, 0.28159925, 0.26041622, 0.20980581,
       0.13261509, 0.07836869, 0.03582265, 0.02653486, 0.00677411])
```