Predicting financial distress at Dutch general hospitals: a machine learning approach

MSc Business Informatics thesis

Author
Michael Wessels
m.wessels@students.uu.nl
Utrecht University

First supervisor
Dr. M.J.S. Brinkhuis
m.j.s.brinkhuis@uu.nl
Utrecht University

Second supervisor
Dr. M.R. Spruit
m.r.spruit@uu.nl
Utrecht University

Company supervisor
Ir. A. Kot
a.kot@berenschot.nl
Berenschot

Utrecht University

Berenschot

Abstract

A quarter of the Dutch hospital sector is financially unhealthy according to recent benchmark outcomes and the Dutch government hopes to save 1.9 billion euro structurally each year in the curative care. Therefore, a financial outlook for the coming years is desirable and can be of importance to the hospital sector in order to act before they are in possible financial distress. This thesis focuses on predicting the financial situation of Dutch general hospitals by using a combination of machine learning and text mining techniques. The use of machine learning to predict financial distress has been applied in other sectors before, however the (Dutch) hospital sector has not yet been investigated. This thesis also examines the findings from literature that patient ratings and textual data from annual reports improve the prediction of financial distress, a topic that has not been looked into in the context of this research.

This research shows that a combination of machine learning and text mining techniques can be used to predict financial distress in theory. However, data analysis has shown that the performance of the prediction models using financial statement data was poor to bad on average initially (an average AUC score under 0.6), probably due to a dimensionality problem of the dataset. Further analysis has shown that using the six financial ratios of the stress test to predict financial distress improved the performance of the prediction models in general to a fair AUC score (0.7 – 0.8). Adding patient ratings and/or textual data from annual reports lowered the AUC score in the initial data analysis by at least 0.08, where further analysis showed that adding patient ratings improved the score in some cases.

Acknowledgements

What has been an intensive process from beginning to end has found its climax in finishing the thesis by writing this final part. Long I have been in the dark about which direction the thesis was going: from predicting the productivity to eventually predicting financial distress of Dutch general hospitals. Not only have I gained a lot of knowledge about predictive modelling and machine learning, I have learned and made Python my own as well. Although it was a long and intensive process, I would have never wanted to miss it.

I would like to thank my first supervisor, Matthieu Brinkhuis, for helping me throughout the complete process of this thesis. Started with only knowing in which area I wanted to do research to a finished thesis about predicting financial distress at Dutch general hospitals and the countless steps in between. Furthermore, I want to thank my second supervisor, Marco Spruit, for the feedback and making me aware of the poor initial results. I would also like to thank my co-students that triggered me with their questions during the colloquia presentations which helped me get further on multiple occasions.

Next to the help from academia, I would like to thank the company Berenschot for providing me with a place to write the thesis and especially my company supervisor, Ard Kot, who helped me get through the process from beginning to end. Also, I would like to thank Expert 1 (CWZ, Nijmegen) and Expert 2 (Tergooi, Hilversum) for letting me take a look inside the world of hospital finance.

At last I want to thank my family, girlfriend, and friends who supported me through the entire thesis, kept me motivated at the right moments and always had time to help me relax in stressful moments.

One last time, thank you all for helping me through this process.

Michael Wessels

Utrecht, July 8th 2018

Table of contents

Table of figures

Table of tables

1. Introduction

According to benchmark outcomes concerning the Dutch hospitals, a part of the hospital sector in the Netherlands has financial problems (van den Haak, 2017). The outcomes show that almost a quarter (approximately 23 percent) of the hospitals is financially unhealthy. However, the yearly branch report of the NVZ, the Dutch Hospital Association, shows that the sector has some indicators of decline as well as some positive indicators (NVZ, 2017).

The 2017 branch report of the NVZ show that the total equity capital, which is a company's own capital, of the Dutch hospitals has increased € 3.57 billion to € 3.79 billion (NVZ, 2017). So has the solvency, which is the possession of assets in excess of liabilities, from 21.6 percent to 23.6 percent. Furthermore, the average financial resilience of the sector has increased to 21.6 percent, which is well in between the target figure range of 20 and 25 percent that is indicated by the financial sector and the WFZ (NVZ, 2017). This concerns the capacity to recover quickly from financial difficulties. These numbers show that the sector is doing relatively good financially.

On the other hand the profit as a percentage of turnover was 1,3 percent on average in 2016, which is the lowest value in the past seven years. The NVZ indicates in an earlier branch report that this value should be between 2.5 percent and 3.0 percent for a private, partly regulated market of hospitals (NVZ, 2013). The profitability of the sector has also declined, from an average of 3.8 percent in 2014 to 3.4 percent in 2016. Investments in tangible property by the general hospitals have decreased from 1.5 billion in 2015 to 1.4 billion in 2016, thereby staying at a stabilized level that is approximately 30 percent lower than in 2010. In contrast to the positive financial indicators, these numbers show some financial decline of the sector.

With both positive and negative financial indicators in mind, it is not clear how the sector is going to develop financially. Additionally, the newly assigned government of the Netherlands hopes to save 1.9 billion euro structurally for the upcoming years in the curative care (Rutte, van Haersma Buma, Pechtold, & Segers, 2017). Curative care focuses on the cure and treatment of acute and chronic physical conditions. Therefore, a large part of the curative care is covered by the hospital sector. As a result of the announced annual economic savings by the new Dutch government, the hospitals in the Netherlands are getting less money and are expected to deliver at least the same quality of care.

The financial outlook of the hospital sector in the Netherlands is not that positive when we look at the above-mentioned figures. Financial ratios like solvency, liquidity, and profitability are used to evaluate the performance of hospitals, and are often used as to benchmark the financial health (Zeller, Stanko, & Cleverley, 1996). However, benchmarking only focuses on the present and is not using the available data to predict future financial health. To assist the hospital sector in predicting their financial health and prevent distress, machine learning techniques can be used. In multiple sectors, machine learning techniques are already being used for predicting financial distress, bankruptcy and detection of financial fraud (Geng, Bose, & Chen, 2015; Kim & Upneja, 2014; Lin, Chiu, Huang, & Yen, 2015; López Iturriaga & Sanz, 2015). Research also has been done on predicting financial distress using machine learning techniques in the hospital sector (Holmes, Kaufman, & Pink, 2017; Koyuncugil & Ozgulbas, 2012; Morey, Scherzer, & Varshney, 2004; Price, Cameron, & Price, 2005). These researches have in common that they all use financial data to predict the financial health. However, the data in this research consists of both financial and textual data, whereby it involves the use of text mining as well. The use of text mining as part of the machine learning techniques has been applied in multiple researches that focus on financial health (Khadjeh Nassirtoussi, Aghabozorgi, Ying Wah, & Ngo, 2014; Kloptchenko et al., 2004; Shirata, Takeuchi, Ogino, & Watanabe, 2011).

This research adds a third source of data to predicting the financial health of hospitals: the customer satisfaction. Customer satisfaction has not yet been found to be included in the datasets for predicting financial health using machine learning techniques. Research has shown that customer satisfaction has a positive impact on financial

performance in other sectors (Chi & Gursoy, 2009; Kyoon Yoo & Ah Park, 2007; van der Wiele, Boselie, & Hesselink, 2002).

The purpose of the research is to investigate how a combination of text mining and machine learning techniques can be used to predict the financial health in the Dutch hospital sector. It helps the sector and the individual hospital to assess their future financial health and the corresponding bottlenecks in their finance. As a result, precautionary actions can be taken to prevent individual hospitals and the hospital sector for financial distress before it happens.

1.1 Research questions

The main research question is as follows:

> *"How can a combination of text mining and machine learning techniques be used to predict financial distress at Dutch general hospitals?"*

For answering the main research question, a set of sub-research questions have been formulated. The following sub-research questions have to be answered in order to answer the main research question:

- *How is financial distress (of hospitals) measured?*
  By conducting a literature review, this sub question focuses on identifying ways for measuring financial performance. The influence of financial ratios is discussed briefly.

- *To what extent has text mining been used for extracting information from annual reports?*
  Text mining on annual reports has been applied before in scientific research. This sub question focuses on identifying to what extent text mining has been applied before in this area and which techniques can be used for this research.

- *How can textual data, next to financial data, be used as input for machine learning techniques?*
  This research uses both financial and textual data. Therefore, there is a need to get an overview of the possible techniques that can be used for a combination of them.

- *Which machine learning techniques perform best for predicting financial distress?*
  From a literature review, machine learning techniques are identified that performed best for predicting financial health. They are compared to the results of this research, which compares the prediction accuracy of multiple machine learning algorithms by following the CRISP-DM principle..

1.2 Outline

Chapter 2 focuses on the research framework that is followed during this research. The third chapter discusses the literature study which includes the topics of financial distress (the influence of financial ratios and customer satisfaction), the application of text mining, and the application of classification using supervised machine learning. Chapter 4 focuses on predicting financial distress while following the CRISP-DM process, how the data is prepared and how the algorithms are set up. Chapter 5 elaborates the results of the data analysis on predicting financial distress with machine learning.

2. Research approach

This thesis research focuses on predicting financial distress of Dutch general hospitals for which an extensive data analysis is performed on publicly available/open data. For the data analysis part of the research, machine learning techniques are being used instead of standard statistics. The goal of the research is to create a financial distress prediction system for the general hospitals in the Netherlands based on open data and evaluate the outcomes with financial experts. The research approach that has been used is the design science methodology by Wieringa and is depicted in Figure 1 (Wieringa, 2014).

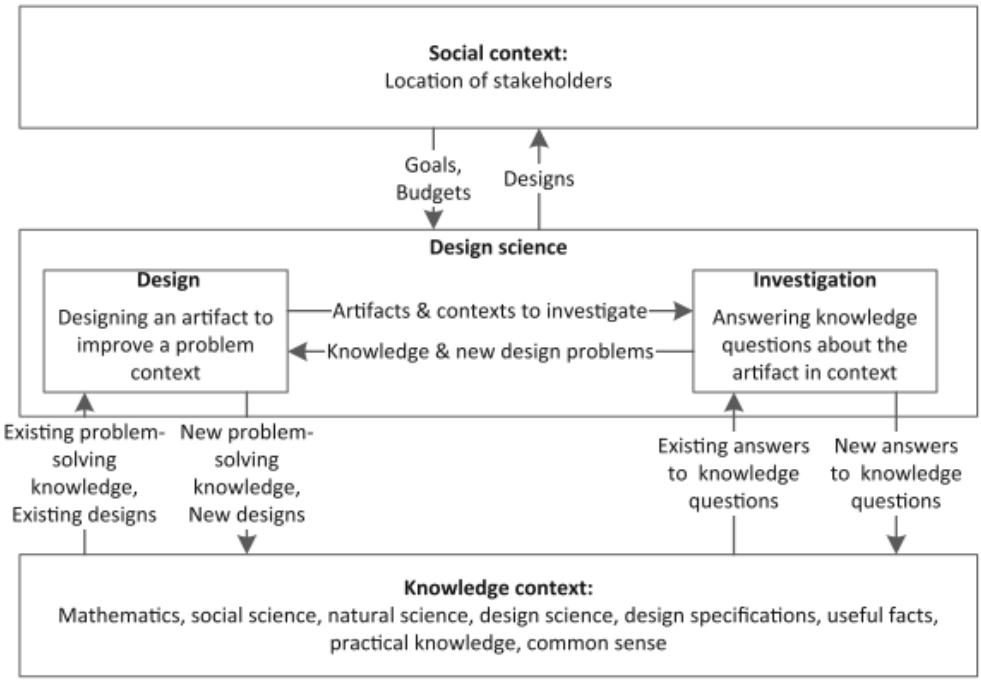

Figure 1. Design science framework (from (Wieringa, 2014))

The research has been divided into several steps in order to separate each individual topic within the thesis:

1. A literature review has been performed to gain knowledge about the main topic of the thesis: financial distress. The use and influence of financial ratios regarding prediction of financial distress is one of the main interests for this phase. Furthermore, the influence of customer satisfaction on financial distress is handled briefly. The goal of this phase is to gather specific knowledge about financial distress that can be used in the data analysis phase.
2. The second phase consists of a literature review as well, this time focussing on the application of text mining. Literature regarding text mining on annual reports is of great importance here, as these documents are used as input for the predictions in the data analysis part. Furthermore, knowledge is gained about text mining techniques that can be implemented in the method of the thesis.
3. Next to knowledge about text mining, knowledge has to be gained about machine learning and in specific for classification. Literature is reviewed for popular classification methods and their use is explained briefly. The discussed machine learning techniques are later used for the prediction of financial distress.
4. With the knowledge from the three literature reviews a start is made for the data analysis. For this the CRISP-DM cycle is used in order to regulate the process throughout the thesis (Wirth, 2000). Each phase from the model except the deployment, from business understanding to evaluation, is went through stepwise. All

necessary datasets are collected and pre-processed so they can be used as input for the machine learning algorithms. For the pre-processing of the data, programming language Python is used.

5. The actual modelling and running of the algorithms is done in this phase with performance/evaluation metrics as output for each classification method. The code for the classification models is written in Python and the scikit-learn package is used for the machine learning part (Pedregosa et al., 2011).
6. The last phase is conducting interviews with financial experts that have knowledge of and experience in the hospital domain. The goal of this phase is to extract knowledge about how they predict financial distress at general hospitals and which features they look for. The outcomes of this phase is compared with the outcomes from the data analysis in phase 5.

For the literature reviews, a snowball approach has been used. Snowballing refers to searching scientific papers by following the references from or to one paper to find other relevant papers (Wohlin et al., 2012).

3. Literature study

The literature study focuses on the first three sub-questions of the main research question. Financial distress prediction of hospitals is discussed first (3.1), after which the application of text mining is elaborated (3.2). The third and last paragraph of the literature study focuses on classification with supervised machine learning techniques (3.3).

3.1 Financial distress prediction of hospitals

The aim of this chapter is to provide an overview and understanding of what financial distress is and how it has been measured in past studies. The use of financial ratios in predicting financial distress is discussed, as is the application of it in the hospital sector. At last, an overview is made of the different machine learning techniques that have been used in literature so far to predict financial distress.

Financial distress is defined as "a condition for a firm where it either cannot meet or has difficulty in meeting its fixed charges" ("financial distress," n.d.). Failure prediction, financial distress prediction, bankruptcy prediction are a few of many terms that are concerned with the research topic of predicting the financial health of a company. The differences between them are marginal, i.e. a company is in financial distress first before it goes bankrupt. Therefore, financial distress can be seen as an early indicator for bankruptcy or business failure. In the remainder of this literature study these terms are used apart from each other, but can be seen as interchangeable.

A lot of research has been done in accounting with regard to predicting bankruptcy from information in companies' financial statements. It goes back as far as the mid-1900s when a ratio was created for the evaluation of credit-worthiness: the current ratio (Beaver, 1966). Nowadays, financial ratios are identified to be a good measurement to predict bankruptcy and are known to have a predictive power up to at least five years prior to bankruptcy (Beaver, McNichols, & Rhie, 2005).

Failure prediction models can be divided into three different categories: classical statistical models, artificially intelligent expert system models, and theoretical models (Adnan Aziz & Dar, 2006). Univariate, multiple discriminant analysis, linear probability model, logit and probit models are different types of classical statistical models, whereas case-based reasoning, decision trees and neural networks are examples of artificial intelligent models. The last category, theoretical models, consists among other things of entropy theory, Gambler's ruin theory and credit risk theories. In the remainder of this research, the focus is on the artificially intelligent expert system models. These models are explained briefly in chapter 5.

In most of the studies that focus on predicting financial distress, financial data from the concerned company is used as input for the prediction. In many cases this is the only input data, but in recent years other variables have been added as input variables as well (Watkins, 2000). It has been shown that market-related variables increase the ability to predict bankruptcy (Beaver et al., 2005).

The remainder of this chapter discusses the use of financial ratios in financial distress prediction, the use of other variables, and the application of financial distress prediction for hospitals.

3.1.1 Financial ratios

Financial ratios are an important tool in predicting business failure and are therefore often used to develop machine learning models and classifiers (Huang, Tsai, Yen, & Cheng, 2008). In the last decade, a lot of research has been conducted on predicting business failure/bankruptcy/financial distress using financial ratios as input for the model (Azayite & Achchab, 2016; Chen, 2011; Holmes et al., 2017; Koyuncugil & Ozgulbas, 2012; W. W. Wu, 2010). To understand what financial ratios are, how they are measured and what they measure, an overview has been made of popular ratios (see table 15, 16, and 17, Appendix A).

A financial ratio is a quotient of two numbers, where both numbers consist of financial statement items (Beaver, 1966). They can be categorized in different groups which each show a different facet of a company's finances and operations (Lin, F., Liang, D., Yeh, C.C., & Huang, 2014). They distinguish the following types of ratios, which are the best known and most used categories:

- **Leverage ratios** measure in which extent debt is used in the capital structure of a company.
- **Liquidity ratios** focus on displaying the short term financial situation of an organization.
- **Operational ratios** are used to show the efficiency of a company's operations and use of assets by applying turnover measures.
- **Profitability ratios** display the return on sales and capital employed by using margin analysis.
- **Solvency ratios** focus on the ability that a company can meet their financial obligations and generate cash flow.

Next to the use of financial ratios, general financial indicators are used for predicting financial distress as well. (Geng et al., 2015) uses a combination of 31 financial indicators (see Table 15, Appendix A), whereas others use raw data from the financial statements as input for the prediction model (Azayite & Achchab, 2016). Furthermore, (W. W. Wu, 2010) suggests that research should not focus on predicting business failure based on failed and non-failed companies at all, but instead focus on key financial ratios only and act when necessary. It can be concluded that financial ratios are used often to predict financial distress and is a good measure to apply in further research.

3.1.2 Hospital sector

As this research focuses on general hospitals in the Netherlands, this section aims at briefly discussing the use of financial distress prediction and financial ratio analysis for hospitals.

Each year, several companies in the Netherlands publish a benchmark concerning the financial situation in the Dutch health care sector. Next to consultancy firms (BDO, 2017; EY, 2017), public organizations like the NVZ and CBS create reports about the financial situation in the Dutch health care sector as well (CBS, 2013; NVZ, 2017). These reports use financial ratios, next to other general financial indicators to assess and benchmark the sector. The CBS uses a list of 24 financial indicators (see Appendix A, table 16), whereas EY uses ten financial ratios and BDO five financial indicators. An overview of which financial ratios are used in each report is shown in Table 1.

Table 1. Overview of financial ratios in Dutch benchmark reports

Ratio	(CBS, 2013)	(BDO, 2017)	(EY, 2017)
Net profit margin	x	x	x
EBITDA ratio	x		x
Staff costs	x		x
Solvency	x	x	x
Loan to value			x
Financial resilience	x		x
Current ratio	x	x	x
DSCR		x	x
ICR			x
Net debt / EBITDA			x
Profitability	x	x	
Quick ratio	x		

Most financial distress prediction research has been conducted on companies rather than hospitals, however academic research in this area has also been conducted on hospitals. (Holmes et al., 2017) uses hospital and community characteristics to predict the risk of financial distress two years in the future by applying a logistic regression forecasting. Furthermore, research from (Koyuncugil & Ozgulbas, 2012) focuses on developing a financial early warning system for hospitals with financial ratios by applying a specific type of decision tree algorithm. These papers are the only academic researches that have been found that apply models to predict financial distress in hospitals. Therefore, there is much to win by doing further research on financial distress prediction in the hospital sector.

Other papers have been found that focus on financial distress prediction for hospitals, but they do not predict the distress in the paper and only discuss the topic. (Price et al., 2005) found that tracking seven specific financial ratios regularly and consistently is the only way to be assured of seeing signs of impending financial distress far enough in advance to respond effectively. (Watkins, 2000) suggests through his findings that non-financial data can be used to predict financial performance and captures aspects that financial data does not capture.

3.1.3 Customer satisfaction

Findings from the literature review on financial distress has shown that non-financial data can be considered to use to predict the distress as it captures other aspects. Therefore, a new non-financial metric is included in this research: customer or patient satisfaction. Research that has been conducted in the hotel sector suggests that customer satisfaction has a significantly positive impact on the financial performance of the company (Chi & Gursoy, 2009). Related research has been conducted in the Greek banking sector, finding that customer satisfaction has no significant impact on the financial performance of a bank (Keisidou, Sarigiannidis, Maditinos, & Thalassinos, 2013). However, the writers put the results of their paper into perspective as the Greek economy was very unstable at the time and the banking sector in general in a crisis.

No academic research has been found that suggests customer satisfaction has an impact on the financial performance of hospitals. As research has shown that non-financial information can be used to measure the financial performance of a company (Holmes et al., 2017; Watkins, 2000), a new topic of interest is included in this research by adding customer satisfaction measures of hospitals as a metric.

3.2 The application of text mining

This chapter focuses on the application of text mining: what it is, classification with textual data and text mining annual reports. The first section explains what text mining is, after which the next two paragraphs focus on the remaining two topics.

Text analysis in the financial domain has grown in popularity since textual reports of companies and other institutions have been made publicly available online on the internet. Before these documents were available online in readable formats, analysis in the financial domain had been focusing on the data it is known for: numerical data. As is shown in the remainder of this chapter, the addition of text analysis to initial numerical analyses has extended the application of financial analysis.

Text mining is defined as ‘the discovery and extraction of interesting, non-trivial knowledge from free or unstructured text’ (Kao & Poteet, 2007). It focuses on textual properties like grammar and structure, and uses techniques from natural language processing (NLP), computational linguistics, corpus linguistics, machine learning (ML) and statistics, next to counting word and term frequencies (Kobayashi, Mol, Berkers, Kismihók, & Den Hartog, 2017). Nowadays text mining is applied in a large number of domain specific applications, i.e. human resource management, customer relation management, company resource planning and security applications (Kumar & Ravi, 2016; Patel & Soni, 2012). This list is only a fraction of all the domains in which text mining has been applied, reaching beyond the initial applications and spreading to other domains.

The difficulty of text mining is that the data often comes in unstructured form in contrast to many other data sources that are structured or numerical. Therefore, extracting useful insights from unstructured textual data is more difficult than with other forms of data and is seen as the most challenging research aspect for fundamental data (Khadjeh Nassirtoussi et al., 2014). As this is seen as a challenging topic, this literature section only focuses on text mining literature that is relevant and close to the research topic of financial distress prediction. The next paragraph focuses on a specific form of text mining technique: classification.

3.2.1 Classification with textual data

Classification is one type of text operation and can be explained as the assignment of objects to predefined classes or categories, where the goal is to construct a model that can predict the category of a given document (Kobayashi et al., 2017). This type of text operation is used in the method of this research, as the hospitals are assigned to either of two classes: financial healthy or unhealthy. More information about the application of classification in this research is given in chapter 6, Predicting financial distress using CRISP-DM.

The process of text classification can be divided in multiple steps (Patel & Soni, 2012) as depicted Figure 2. A collection of text documents is divided into a set of training- and test documents. The training documents are put into a classification algorithm that trains the classifier model. The resulting classifier model can then be tested with the test documents as input and evaluated by how well the classifier model performs in terms of classifying the document into the right classes (Kos, Schraagen, Brinkhuis, & Bex, 2017; Schraagen, Brinkhuis, & Bex, 2017).

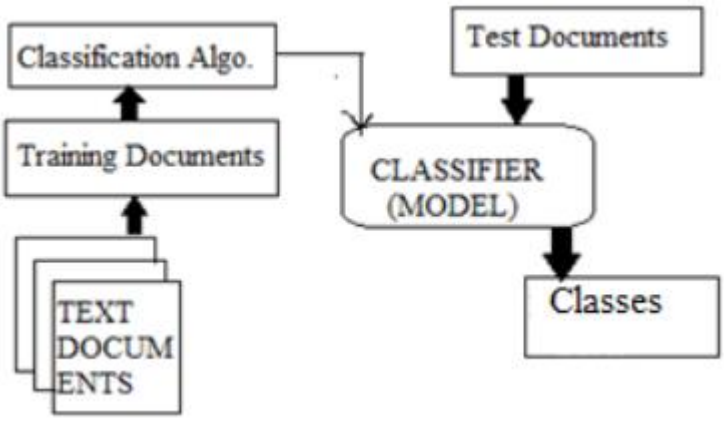

Figure 2. Classification model: construction and prediction (from (Patel & Soni, 2012))

The classification of text is done using machine learning algorithms which are briefly described in chapter 4. An overview has been made for popular machine learning algorithms to predict market change based on collected literature (Khadjeh Nassirtoussi et al., 2014). These categories of algorithms are divided into Support Vector Machine (SVM), regression algorithms, naïve Bayes, decision trees, combinatory algorithms and multi-algorithm experiments.

3.2.2 Text mining on annual report

One of the documents that have become publicly available and easy to access through to rise of the internet are the annual reports of organizations. Listed companies are obligated to publish this document, which contains information about the organization's performance of the previous year, to the public each year. Next to listed companies, the Dutch health care institutions are obligated by the Ministry of Health, Welfare and Sport to publish their annual report each year as well. This obligation is for accounting to the Dutch citizens where they are

spending their tax money on. This section discusses the research that has been conducted in the domain of text mining on annual reports.

Annual reports contain information about the past performance of a company as well as indicators of its future performance. Whereas the financial numbers and ratios contain information about the past, the linguistic structure and text style may indicate how well an organization will do (Kloptchenko et al., 2004). Research has shown that unprofitable companies focus more on the future in the textual part of their annual report than on their past performance (Clatworthy & Jones, 2006). Furthermore, it has been discovered that a change in the text of annual reports indicates a change in the near future most of the time (Magnusson et al., 2005) and the tone of the text changes before the financial change occurs (Kloptchenko et al., 2004).

In the line of this research, the application of text mining on annual reports has been used for predicting the financial performance of companies as well. (Qiu, Srinivasan, & Street, 2006) have confirmed that it is feasible to use text classification on annual reports to predict the financial performance of a company. In other research their prediction with the use financial indicators, annual reports and supervised machine learning models achieved a better prediction performance than the forecast of actual financial analysts (Qiu, Srinivasan, & Hu, 2014). Furthermore, differences have been identified between bankrupt and non-bankrupt companies when particular expressions appear together in the same section of their annual report (Shirata et al., 2011).

The prediction of financial performance is not the only application in the financial domain for text mining annual reports. It has been used for identifying financial statement fraud in multiple occasions (Goel & Gangolly, 2012; Hajek & Henriques, 2017; Sadasivam & Lakshme, 2016), bug report classification (Zhou, Tong, Gu, & Gall, 2016), stock price movements (Doucette & Cohen, 2015) and future accounting and market performance (Balakrishnan, Qiu, & Srinivasan, 2010) among others.

Next up comes a chapter about machine learning techniques for classification. Text mining is a large research field by itself, however for the prediction of financial distress machine learning techniques have to be used for the actual prediction.

3.3 Classification with supervised machine learning techniques

This chapter gives an overview of classification using supervised machine learning techniques. First the overall subject of machine learning is discussed, after which supervised techniques are elaborated in specific. The last section consists of an overview on applications of classification techniques.

In the last decade a lot of research has been conducted in the field of data mining and machine learning. However, the difference between these two is not always clear for people, so it is important to distinguish the two from each other first to get an understanding about the topic of interest. Data mining has been defined as “the application of specific algorithms for extracting patterns from data” and is historically part of the knowledge discovery in databases (KDD) process (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). When data mining is applied, the extracted patterns can help to shape the decision making process. In comparison, the focus of machine learning is on designing algorithms that learn from and predict based on data. Machine learning has been defined as “the systematic study of algorithms and systems that improve their knowledge or performance with experience” (Flach, 2012). In practice, machine learning is used in data mining techniques to identify connections between the resulting relationships.

Machine learning can in turn be divided into three main categories: supervised, unsupervised, and semi-supervised learning. This research focuses on the first, supervised machine learning.

3.3.1 Supervised machine learning techniques

Supervised machine learning is defined as the search for algorithms that reason from externally supplied instances to produce general hypotheses, which then make predictions about future instances (Kotsiantis, 2007). The algorithms used in supervised machine learning are provided with a dataset that is divided into a training and testing set which both have input and output data. The input data has an x amount of features and is the data that is used to predict the output data, which is also known as the label. The training set is used to train a decision model that determines which input data belongs to a certain output. When the model has been trained with the training set, the model is tested on the test set. The input for the testing stage is an unlabelled dataset with the same features as the training set. The decision model is used to predict the labels for the test set and is evaluated by comparing the predicted values with the actual values. The evaluation criteria for a decision model are explained briefly in paragraph 3.3.2 – Classifier evaluation.

To have an idea of what the other types of machine learning techniques are, some differences are briefly discussed here. In comparison to supervised learning, unsupervised learning does not have a separate training and testing data. This technique tries to uncover patterns in the data without being given output data. The best known type of unsupervised learning is clustering, which groups a dataset in subgroups called clusters so that data in the same group is more similar to each other than to the data in other subgroups (Jain, Murty, & Flynn, 1999). With semi-supervised learning, both labelled and unlabelled data are used to perform a task which initially was a supervised or unsupervised task only (Zhu, 2007). An example of semi-supervised learning is to add labelled data to an unsupervised learning task.

A further distinction can be made for supervised machine learning, splitting it into two categories: regression and classification. Regression models is used when the output variable is a so-called real value, such as an integer, double or any other digit type. The other type, classification, is used when the output variable is a class or category. Each entry of a dataset is then e.g. divided into one of two classes: 'good' or 'bad'. As this is classification with two classes, it is called binary or binomial classification (Lin, F. et al., 2014). However, classification can also be used with three or more classes, and is then called multiclass or multinomial classification.

The next few paragraphs explain five well known supervised machine learning algorithms which are used in (Kos et al., 2017) as well: logistic regression, support vector machine, decision tree, naïve Bayes and k-nearest-neighbor (k-NN).

3.3.1.1 Logistic regression

Logistic regression is one of the most used methods in regard to modelling binary response data (Hinde, 2011). The output variable of a logistic regression model is always one of two numbers, zero or one. The model predicts the probability of an object to belong to the default class and is then assigned to the class that is nearest. For example, if the probability is 0.72 the object is assigned to class 1, and if the probability is 0.49 it is assigned to class 0. The fact that the output variable is a number, so a continuous value, is one of the main characteristics of a regression model. This is in contrast to binary classification models which always have class labels as output variable. As the output variable is always either 0 or 1, logistic regression can only be used for binary classification and not for multinomial classification.

For the logistic regression model to perform well, several steps should be taken concerning the preparation of the dataset (Hinde, 2011):

- The output variable should always be binary, either 0 or 1.
- Remove noise by deleting outliers and objects that are classified wrong in the training set.
- The data should be normally distributed.
- Removing correlated input variables

Furthermore, to lower the amount of variables for better accuracy and interpretability purposes, least absolute shrinkage and selection operator (LASSO) is a popular method used for both variable selection and regularization (Meier, van de Geer, & Bühlmann, 2008; Tibshirani, 1996).

3.3.1.2 Support Vector Machine

Support vector machines are considered one of the best predictive models as their accuracy and performance is high (Cristiannini & Shawe-Taylor, 2000). The method was initially designed for binary classification problems, but can also be used for multinomial problems (Bellazzi & Zupan, 2008). A SVM is a linear classifier which uses the principle of margin maximization, where structural risk minimization is used to improve the performance (Adankon & Cheriet, 2009). The SVM model finds a hyperplane with a maximum distance to the closest point of the two classes which is called the optimal hyperplane (see Figure 3). The model also has a support vector which is a set of objects that are closest to the optimal hyperplane. In Figure 3 the hyperplane is shown as the area between the two outer lines, the middle line is the optimal hyperplane, and the outer lines with the red plots on it are the maximum margins of the hyperplane. The two classes are separated by the hyperplane, everything on one side of the hyperplane belongs to class A and everything on the other side belongs to class B.

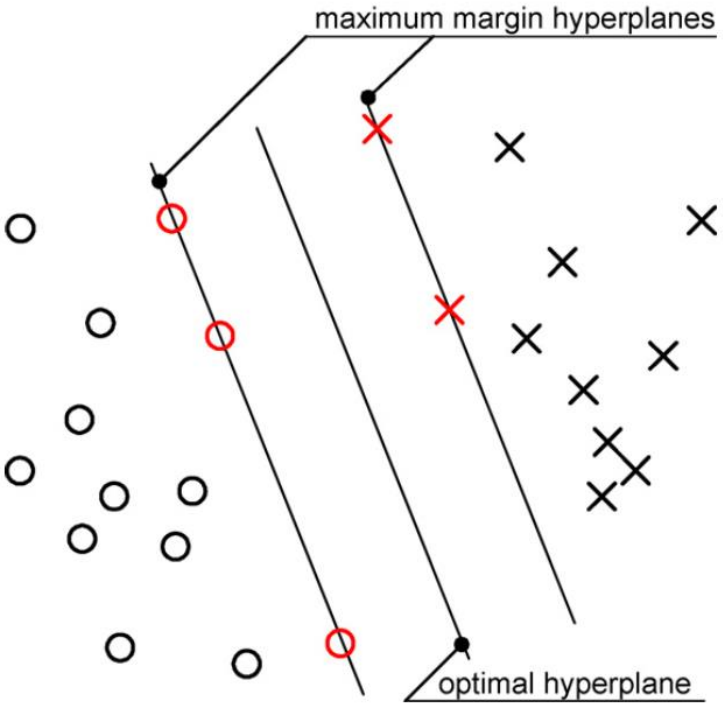

Figure 3. An example of a two-class SVM. The red plots define the margins of the hyperplane and are called support vectors (from (Bellazzi & Zupan, 2008)).

The SVM model works best with continuous variables as input data, other data types can be transformed into continuous variables with one-hot encoding. One-hot encoding converts categorical variables into additional columns that each either has the category (shown by a 1) or does not have the category (shown by a 0). For example, when a variable initially has three different types of entries, 'Blue', 'Red' and 'Green', one-hot encoding converts this to three new columns containing those three entry names. When an observation is blue, the variable of column 'Blue' is assigned the value 1 and the other two columns the value 0. The difference between the tables before and after one-hot encoding is shown in Table 2 and Table 3 respectively.

Table 2. Table before one-hot encoding

Id	Color
0001	Red
0002	Green
0003	Green
0004	Red
0005	Blue

Table 3. Table after one-hot encoding

Id	Red	Green	Blue
0001	1	0	0
0002	0	1	0
0003	0	1	0
0004	1	0	0
0005	0	0	1

3.3.1.3 Decision tree

Decision trees are another popular method for modelling data with machine learning due to the ease to understand them and their ability to find the most important features from a dataset (Flach, 2012). The algorithms of decision trees are based on how well features divide the dataset, ranking them from high to low. The feature that is ranked highest is used to split the dataset on each time, removing the feature on which it has been split from the list for the next split. The point where a dataset is split is called a node, which has three different types: a root node, branch nodes, and leaf nodes (see Figure 4). The root node, also called first node, is the node where the first split happens, and therefore the best split of the entire dataset. Branch nodes are nodes that have at least one child, and therefore the tree continues at this point, where leaf nodes are nodes that have no children and is the end point of one part of the three (Myles, Feudale, Liu, Woody, & Brown, 2004).

Decision tree models are prone to overfitting training data, which means that features which cannot be generalized are included in the model. It describes features that come from noise or variance in the dataset and can lead to a lower accuracy (Blockeel & Leuven, 2017). One way to avoid overfitting is to apply pruning, if two trees have the same accuracy the one with fewer leaves is preferred (Kotsiantis, 2007). When a node is pruned it is changed from a branch node to a leaf node. Although methods like pruning reduces overfitting of decision tree models, they do not have a negative effect on the quality of the rules in the decision tree (Myles et al., 2004).

For the use of decision tree models in the research to be conducted in this thesis, not all models can be used as they are initially designed to handle categorical data and the dataset in this thesis consists of numerical data (Kuhn & Johnson, 2013). However, more recent implementations of decision trees like C4.5 can handle numerical data in addition to categorical data (Quinlan, 2014). In contrast to other learning methods, the dataset of a decision tree does not need to be normalized to get optimal results.

As decision trees are prone to overfitting, random forests are often used instead as they are robust against overfitting (Liaw & Wiener, 2002). Random forests, or random decision forests, construct multiple decision trees during the training phase and aggregate their results into one (better) result (Ho, 1995). As an added outcome, random forests have a measure of variable importance as well as a measure of the internal structure of the data (Liaw & Wiener, 2002).

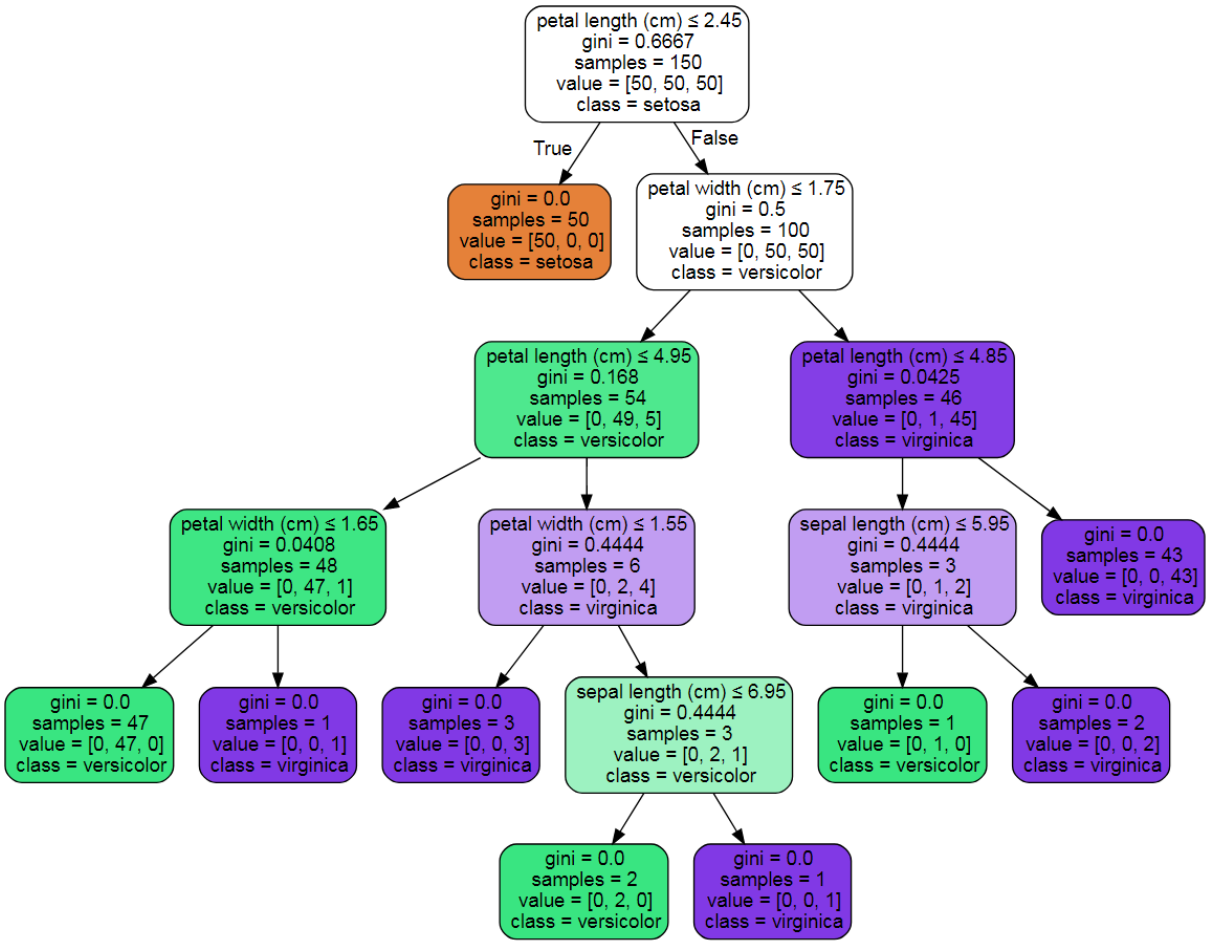

Figure 4. Example of a decision tree. The top node is the root node, the orange and bottom row nodes are examples of leaf nodes, whereas both nodes in the third row are examples of branch nodes.

3.3.1.4 Naïve Bayes

Naïve Bayes is a classification method that is based on the principle of Bayes theorem which assumes that all of the predictors are independent of each other (Kumar & Ravi, 2016). In real life datasets this complete non-correlation is never true, but the naïve Bayes models are nevertheless very accurate in these real life situations and therefore used frequently (Domingos & Pazzani, 1997). The formula of Bayes theorem which is used by Naïve Bayes is:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Bayes models can be used for both numerical and non-numerical datasets, but it has been found that in general it performs best with categorical data (Webb, 2003). For the preparation of the dataset, additional parameter tuning is not required for the basic models. Normalization of the data is also not required as the algorithm of the model does this by default. Though, one aspect of the dataset that must be given attention is the distribution, as different distributions require different sorts models (Pedregosa et al., 2011). The three types of naïve Bayes are Gaussian naïve Bayes, multinomial naïve Bayes and Bernoulli naïve Bayes.

3.3.1.5 k-NN

K-nearest-neighbor (k-NN) is yet another classification method that is used frequently. The model is based on the nearest neighbor rule which identifies the category of unknown data points on the basis of its nearest neighbor whose class is already known (Bhatia & Author, 2010). The method's origin lies in domain experts making

decisions that are based on similar cases from the past and is nowadays used in fields like text categorization, pattern recognition and ranking models (Bellazzi & Zupan, 2008). An example of the model is visualisation is given in Figure 5.

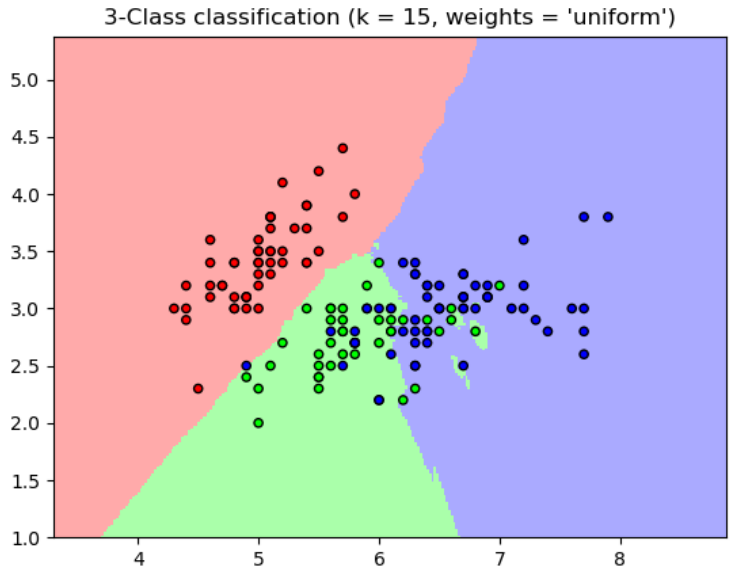

Figure 5. Sample plot of an k-NN classification using scikit-learn

The nearest neighbor method is simple and effective, but it has three characteristics that makes it less suited than other methods (Jiang, Pang, Wu, & Kuang, 2012).

1. The sample similarity computing is complex;
2. A single training sample, like noisy samples, can affect the model's performance easily;
3. As the model is a lazy learning method and therefore does not build the classification model, the model is not well suited for real time applications.

Although its shortcomings, the method is used frequently and therefore used as a method in the research of this thesis.

3.3.2 Classifier evaluation

In the previous paragraph five different classification techniques have been discussed which are used in this thesis research. Each classification technique is going to be evaluated by a range of output metrics, after which the techniques are compared to each other by these metrics. The evaluation of machine learning algorithms is often done by using recall, precision and the f-measure as comparison metrics as accuracy is no reliable measure for machine learning algorithms (Powers, 2011). These metrics are based on a confusion matrix which is shown in Table 4. The axes of the table are 'Actual class' and 'Predicted class' which stand for the actual class from the original dataset and the predicted class by the classification method respectively. When the class in the original dataset is A and the classification model predicts it as A, it is called True Positive (TP). However, when in this case the model predicts B instead of A, it is called a False Negative (FN) as it is falsely marked as a 'negative' outcome. Furthermore, when the actual class was B and the model predicts A, this is called False Positive (FP), and when the actual class is B and the model predicts B as well, it is called True Negative (TN). From here on, the evaluation metrics are calculated (see equations 3.1 up until 3.8).

Table 4. Example of a confusion matrix for the evaluation of classification methods

		Predicted class	
		A	B
Actual class	A	True positive (TP)	False negative (FN)
	B	False positive (FP)	True negative (TN)

$$TP_{rate} = \frac{TP}{TP+FN} \quad \textbf{[3.1]}$$

$$TN_{rate} = \frac{TN}{TN+FP} \quad \textbf{[3.2]}$$

$$FP_{rate} = \frac{FP}{TN+FP} \quad \textbf{[3.3]}$$

$$FN_{rate} = \frac{FN}{TP+FN} \quad \textbf{[3.4]}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \textbf{[3.5]}$$

$$Precision = \frac{TP}{TP+FP} \quad \textbf{[3.6]}$$

$$Recall = \frac{TP}{TP+FN} \quad \textbf{[3.7]}$$

$$F_1 = 2 * \frac{precision*recall}{precision+recall} \quad \textbf{[3.8]}$$

4. Method

The method chapter of this thesis describes how the data analysis part is conducted. The CRISP-DM methodology is used to streamline the process. In the following paragraphs the method is elaborated.

4.1 Predicting financial distress using CRISP-DM

For this research, a knowledge discovery process called CRISP-DM (Wirth, 2000) is used to predict the financial distress of hospitals. The CRISP-DM methodology, short for Cross Industry Standard Process for Data Mining, is the most popular methodology in the field of data mining projects (Mariscal, Marbán, & Fernández, 2010). The model consists of six different phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment (see Figure 6). Each phase of the process is explained shortly in the section below and is afterwards applied in the research.

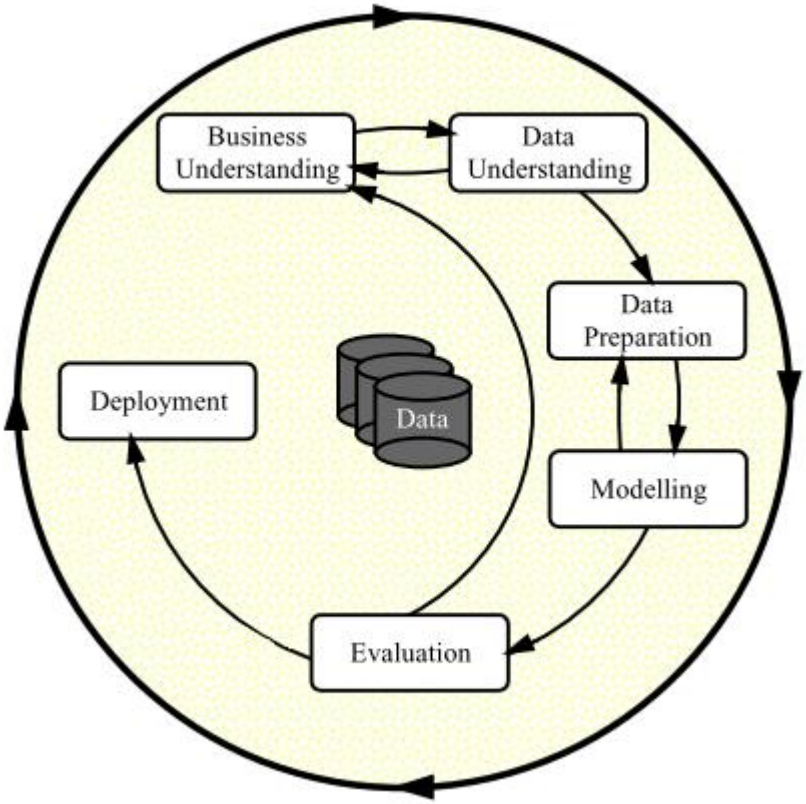

Figure 6. Phases of the Cross Industry Standard Process for Data Mining (from Wirth (2000))

Wirth (2000) describes the six phases of CRISP-DM as following:

1. The initial phase of the process, **business understanding**, focuses on understanding the objectives and requirements of the project from a business perspective. This knowledge is translated into a data mining problem definition, and a preliminary project plan designed to achieve the objectives.
2. The **data understanding** phase starts with an initial data collection and continues getting familiar with the data, identifying data quality problems, and discovering first insights into the data.

3. The third phase, **data preparation**, covers all the activities to construct the final dataset from the initial dataset.
4. In the **modelling phase**, modelling techniques are selected and applied to the prepared dataset from the data preparation phase. The parameters of the modelling techniques are then calibrated to optimal values to get the best results from the model.
5. The **evaluation phase** focuses on evaluating the model(s) that have been created in the modelling phase and reviewing all steps to be certain it properly achieves the business objectives.
6. In the **deployment phase**, the extracted knowledge needs to be organized and presented in a way that someone else can use it.

Each phase of the cycle, except the deployment phase, is went through stepwise, from business understanding up until the evaluation phase.

4.1.1 Business understanding

According to benchmark outcomes concerning the Dutch hospitals, a part of the hospital sector in the Netherlands has financial problems (van den Haak, 2017). The outcomes show that almost a quarter (±23%) of the hospitals is financially unhealthy. Furthermore, the WFZ (Guarantee Fund for the Health Care Sector) are monitoring fifteen of their affiliated eighty hospitals, because they have a high risk profile and financially unstable (Waarborgfonds, 2016).

With this in mind, the newly assigned government of the Netherlands hopes to save 1.9 billion euro structurally for the upcoming years in the curative care (Rutte et al., 2017). Curative care focuses on the cure and treatment of acute and chronic physical conditions. Therefore, a large part of the curative care is covered by the hospital sector. As a result of the announced annual economic savings by the new Dutch government, the hospitals in the Netherlands are getting less money and are expected to deliver at least the same quality of care.

Therefore, it is of high interest to identify and especially predict the financial health of the hospital sector. The rest of this chapter focuses on explaining the characteristics that are used as comparison in this research.

4.1.1.1 Model characteristics

The objective of the research is to create a variety of models to predict the financial health of hospitals and make a comparison to identify the best model(s). This approach has been adapted from (Geng et al., 2015), see Figure 7. The research aims to identify three different characteristics of the prediction model that predict best:

- Time frame
- Feature set
- Classifier model

For the time frame characteristic, this research hopes to identify the best prediction time frame for financial health of hospitals. The time frames that are used as comparison are $t-1$ up until $t-5$ to predict the financial health in year t. For example, data from the year 2009 ($t-1$) and 2005 ($t-5$) are used to predict the financial health in 2010 (t) and afterwards compared which set predicts best.

For the feature set characteristic, three different feature sets are used and compared. The first feature set consists of financial statements, the second of textual annual reports, and the third of customer satisfaction grades/patient ratings.

All health care institutions in the Netherlands must submit their annual report and financial statement to a national register that is managed by the CIBG, an organization affiliated to the Dutch Ministry of Health, Welfare and Sport. The annual report is a general overview of what has happened in the organization for the concerned year, including work methods, employee information, and organizational goals. Financial statements consist of a

balance sheet and an income statement, an overview of the financial situation of an organization for the previous year. In general, the financial statement is added to the annual report as an financial overview. Customer satisfaction for the hospitals in the Netherlands is measured as a grade from 1 up until 10. With 1 being completely unsatisfied and 10 being completely satisfied about the given care in the hospital.

The classifier model characteristic is a list of different classifier models for unsupervised machine learning algorithms. The following classifier models are applied in this research:

1. Logistic regression (LR)
2. Support vector machine (SVM)
3. Decision tree (DT)
4. Naïve Bayes (NB)
5. K-nearest-neighbor (k-NN)

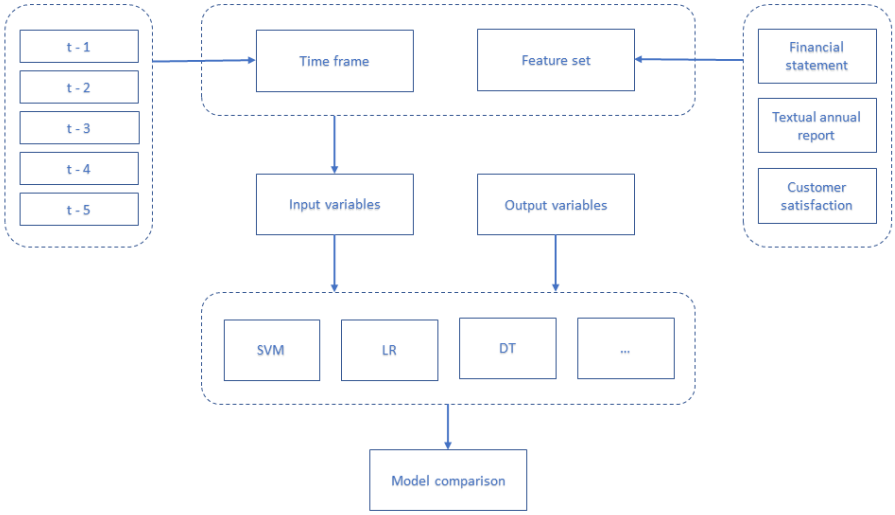

Figure 7. Overview showing the steps of the research (adapted from (Geng et al., 2015))

4.1.1.2 Financial indicators

The required data for every characteristic is publicly available, however there is no list available of financial (un)healthy hospitals. Therefore, we base this characteristic on a combination of financial indicators that are used in multiple benchmarks in the Netherlands to determine whether a hospital is financially healthy or unhealthy called the 'stress test' (BDO, 2017; EY, 2017). The stress test is based on six financial ratios: profitability, solvency, loan to value, debt service coverage ratio, interest coverage ratio, and the net debt / EBITDA (Earnings Before Interest, Taxes, Depreciation and Amortization) ratio. These ratios tell something about result and performance of a company, as does it about the financial position and structure, and the interest- and repayment capacity. The financial indicators with their corresponding formulas and standards are shown in Table 5.

Table 5. List of financial indicators to determine financial health

Category	Financial indicator	Formulae for calculation	Standard
Result and performance	Return	Net result / earnings	>1.5%
Financial position and structure	Solvency	Equity capital/ total capital	>20%
	Loan to value	Long term debt / tangible asset	<70%
Interest- and repayment capacity	DSCR	EBITDA / (interest expenses + amortization)	>1.3
	ICR	EBIT / interest expenses	>2
	Net debt / EBITDA	Net debt / EBITDA	<3.5

The standards of each financial indicator are applied to each available financial year of a hospital, calculating an outcome for each indicator. The minimum score for a hospital is zero (when they meet none of the financial ratio standards), whereas the maximum score is six (when they meet all six standards). To determine whether a hospital is financially healthy, a line is drawn between three and four positive outcomes. A hospital is seen as financially healthy when it meets a minimum of four financial standards, when it scores positive for only three or less it is seen as financially unhealthy. These outcomes are used as output variables for the unsupervised machine learning algorithms (see Figure 7) to train and test the dataset.

4.1.2 Data understanding

The final dataset which is used in this research consists of three separate datasets: 1. financial statement data, 2. annual reports from hospitals, and 3. customer satisfaction grades/patient ratings. Each dataset is discussed in the paragraphs below.

4.1.2.1 Financial statement data from DigiMV

The data from financial statements of hospitals is extracted from DigiMV, which is a digital social accountability platform for health care institutions. Health care institutions are obligated to submit a range of data that is pre-determined by the government, in this case by the Dutch Ministry of Health, Welfare and Sport. The purpose of DigiMV is that the Dutch society can see where public/tax money is spent on by each health care institution, and how their care, services and support are organized. As the general hospitals are health care institutions that depend on tax money, they also have the obligation to publish their performance data each year. These performance data are publicly available on the website www.jaarverantwoordingzorg.nl, which is managed by the CBIG.

4.1.2.2 Annual report

Another part of the social accountability for Dutch health care institutions is publishing their annual report. They have to submit their annual report together with the financial statement each year to an online national register that is managed by the CIBG. The national register is accessible for everyone on the website https://www.desan.nl/net/DoSearch/Search.aspx, where both the annual report and the financial statement can be downloaded as PDF files. All available annual reports from the year 2010 up until 2016 are collected and stored to use in the text analysis part of the research.

However, there are some limitations to the annual reports in the national register. Some annual reports are uploaded as a paper scan of the original document and not all reports are (still) available in the register. Therefore, the website of each individual hospital is visited and the annual report is searched. As a result, some other useful

annual reports were found, but also reports that are presented in web and/or video format. The reports that are presented as web and/or video are excluded from the research.

4.1.2.3 Patient ratings

The customer satisfaction grade/patient ratings dataset is not part of the annual social accountability for Dutch health care institutions. The dataset is retrieved manually from Zorgkaart Nederland (Health map from the Netherlands) by scraping the open data from their website. Every patient can submit a grade and review about their health care provider after which the submission is checked by Zorgkaart Nederland. The dataset consists of two variables per health care provider per year: the average grade and the amount of submissions.

Research has shown that 30 ratings are sufficient to get a clear picture of hospitals (Rijcke, Wallenburg, Wouters, & Bal, 2016). Therefore, only hospitals with more than 30 ratings per year are included in the final customer satisfaction dataset.

4.1.3 Data preparation

This step concerns the data preparation where the data is prepared for input in the machine learning models. Each step is explained in an individual paragraph.

The following table shows the amount of observations per dataset after preparing and cleaning the dataset (see Table 6). For example, for the x + 1 timeframe this means that an observation is only valid if there is data of the hospital in year x and x + 1, because x is the data which is used as input for the machine learning models and the data of x +1 is used to calculate the target variable, financial distress, of the hospital.

Table 6. number of observations per dataset after data preparation

# of observations for…	x + 1	x + 2	x + 3	x + 4	x + 5
Financial statement analysis	336	260	197	139	70
+ patient ratings	336	260	197	139	70
+ textual annual report	270	213	163	117	54

4.1.3.1 Financial statement data

For the data preparation of the financial statements several steps are taken before it can be used as input for the machine learning models.

1. Remove empty rows and columns;
2. Change every cell to the same number format;
3. Split dataset into three separate data frames: *balance sheet, income statement and a combination of both;*
4. Create new datasets for each individual timeframe: *x + 1, x + 2, …., x + 5*;
5. Normalisation and standardisation of datasets.

4.1.3.2 Annual reports data

The text mining part of the research follows a framework introduced by (Kobayashi et al., 2017), of which the first step is used in the data preparation step (see Figure 8). This step includes three different phases: text collection, text cleaning and text transformation. Each phase is discussed in the following sections. The second step of the framework, text mining operations, is not used in the way it is mentioned as the textual data from annual reports is not the only input for the unsupervised machine learning algorithms. The last step, post-processing, is discussed briefly, because the main intention of the research is not on extracting new insights from the explicit textual data.

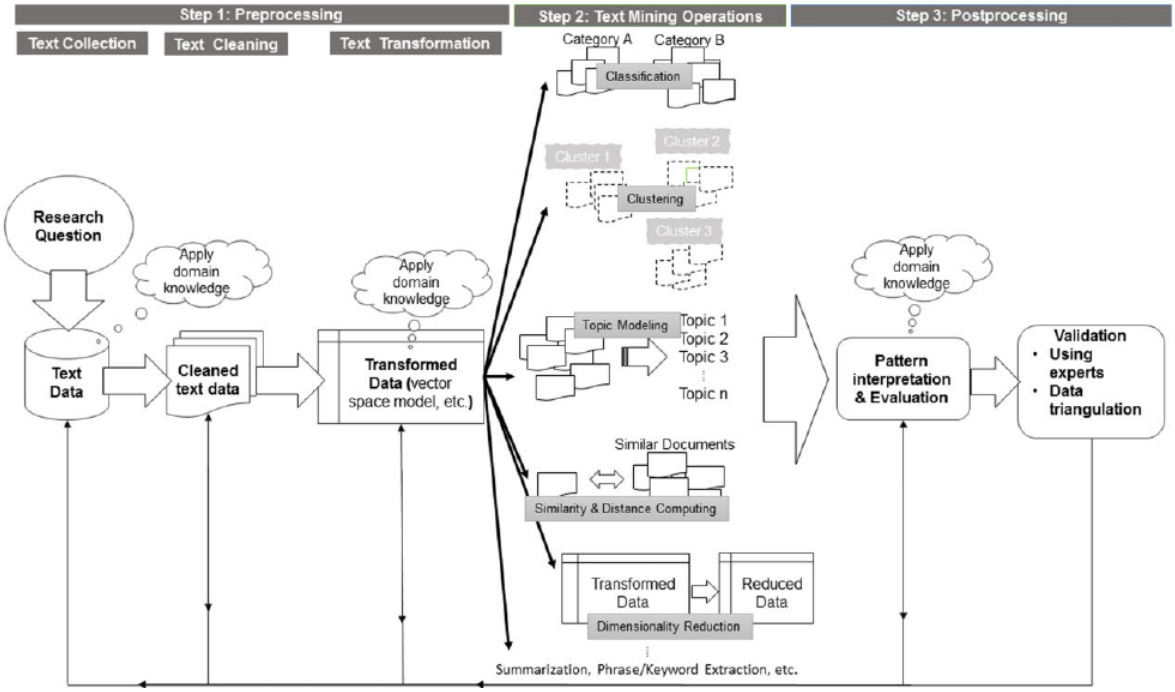

Figure 8. Text mining framework (from (Kobayashi et al., 2017))

4.1.3.2.1 Text collection

The first part of the pre-processing step from the text mining framework is the collection of text. The collection of text documents, called the corpus, for this research exists of annual reports of Dutch general hospitals. This section discusses the approach of how the corpus is collected.

As mentioned in the data understanding step of the CRISP-DM framework, the annual reports are collected by downloading them from the national register and hospitals' websites. In general, the reports are available as PDF files, but some reports are only available in other formats. These other formats are excluded from this research, with a consequence of reducing the final dataset.

The reports that are stored as readable PDF files, so no papers scans of the document, are used for the remaining pre-processing steps. The PDF file format is a good file format to present a file to an audience who are only allowed to read the document and not change it. These files are protected from others to change the files in contrary to other file formats like .txt and .docx. Additionally, when using the text from a PDF file for text mining, the text cannot be accessed from the file directly. The file must be converted to a file format which is readable for a programming language (i.e. Python/R), where the eventual text mining tasks are performed. Therefore, the annual reports that are stored in PDF files are converted to .txt files using the PyPDF2 package in Python. This package lets the user read, write and merge pdf files. For the conversion to text files the PDF files are handled as follows:

1. The file is loaded into Python;
2. Each page is read and saved using the PyPDF2 package;
3. The pages are combined and saved to one .txt file

The result of this conversion from PDF to .txt file is a suitable text file which can be used for text mining. However, the files contain a lot of data that is superfluous for the text mining operations, so the files are cleaned before using them in the analytics part of the research.

4.1.3.2.2 Text cleaning

The second step in the pre-processing process is the cleaning of text. This step includes three different tasks that are carried out: tokenization, stop-word removal and word stemming. Each of the individual tasks is elaborated on in the paragraphs below. For each task in the text cleaning phase the NLTK package, short for Natural Language ToolKit, is used (Bird, Loper, & Klein, 2009).

4.1.3.2.2.1 Tokenization

The first step in the cleaning process is tokenization, where a text document is split into a stream of words by removing punctuation, numbers, tags, and other symbols other than words (Hotho, Nürnberger, & Paaß, 2005; Kobayashi et al., 2017). For the tokenization of the texts in the corpus the tokenize function of the NLTK package for Python is used. A sample of how the tokenization process is coded can be seen in the code block below (Figure 9).

```
from nltk.tokenize import RegexpTokenizer

text = 'the text which is going to be tokenized'

text = text.lower()

tokenizer = RegexpTokenizer(r'\w+')

tokens = tokenizer.tokenize(text)
```

Figure 9. Example code of tokenization

One part of the tokenizing the text is to change everything into lowercase letters to ensure the text mining algorithm handles words that are different only by the capitalization of letters. Otherwise the algorithm would interpret '*Hospital*' and '*hospital*' as two different words, although they are initially the same.

4.1.3.2.2.2 Stop-word removal

Stop-words are the words that are very common in texts without dependency to a particular topic (e.g., conjunctions, prepositions, articles, etc.) (Uysal & Gunal, 2014). These words have no use for the text mining operation and can be removed from the text files beforehand. Stop-words are unique to the language, so for this research we use a list of Dutch stop-words which is included in the NLTK package for Python (see Table 14 in Appendix A). The result of this cleaning step is the same text files only without the Dutch stop-words from the list.

4.1.3.2.3 Word stemming

Stemming of words is a method to homogenize the representation of semantically similar words (Kobayashi et al., 2017) and therefore reduce the number of words that need to be taken into consideration (Gaustad & Bouma, 2002). Words can have multiple forms while having the same meaning. This leads to a large amount of words when doing analysis on texts, which in turn takes more time to do the actual analysis. Therefore, stemming is a useful tool to combine the different word forms with the same meaning into a form that covers them all. The underlying assumption for a fruitful usage of such a stemmer, is that morphological variants of words are semantically related (Kraaij & Pohlmann, 1994). For example, the words *write*, *writer*, *writing* and *writings* all have the same background in semantics. The purpose of stemming is to reduce these words into one 'stem' that covers all four words. In this case, the four words are reduced to the stem *writ*. An application field where this method is used are online search engines: a user wants and gets the same results when he enters either house or houses.

One of the founding fathers of stemming algorithms is M.F. Porter, who introduced a stemming algorithm for the English language only (Porter, 1980). Most other research in the field of stemming has been focused on the

English language as well. However, in the last decades stemming algorithms have been written for other languages (Willett, 2006). Porter has developed a high-level programming language for stemming algorithms called Snowball which can be applied to other languages (Porter, 2001). One of the languages where a stemming algorithm has been created for is the Dutch language, resulting in a Dutch version of Porter's stemming algorithm.

For the implementation in the research, the Dutch Snowball algorithm has been used which is part of the NLTK package in Python. Taking into account the limitations of the algorithm for the Dutch language, the stemming will be effective for Dutch text retrieval (Kraaij & Pohlmann, 1995).

The following code block is a sample of the main functionality in the algorithm that is applicable to the research:

```
from nltk.stem.snowball import SnowballStemmer
stemmer = SnowballStemmer("dutch")
print(stemmer.stem("dateren")
```

Figure 10. Stemming example 1

```
from nltk.stem.snowball import SnowballStemmer
stemmer = SnowballStemmer("dutch")
print(stemmer.stem("daterend")
```

Figure 11. Stemming example 2

The code in both Python code blocks above (Figure 10 and 11) imports the Snowball stemmer from the NLTK package, sets the language of the stemmer to Dutch and eventually stems the Dutch words *dateren* and *daterend*. When we run the code of both examples, the result is the same for both: *dater*.

However, preliminary data analysis of on the data showed that it did not make a large difference if stemming was used or not. Therefore, it has been chosen that stemming is not applied on the text to retain the meaning of each word.

4.1.3.2.4 Text transformation

The last step in the pre-processing process of text mining is text transformation. Text transformation is a quantification strategy in which text is transformed into mathematical structures (Kobayashi et al., 2017). The mathematical structure that is being used is a document-by-term matrix which is a vector representation of how many times all words occur in a document (Scott & Matwin, 1999).

For creating the document-by-term matrix, the *pandas* (McKinney, 2010) and *sklearn* (Pedregosa et al., 2011) packages for Python are used. The *sklearn* package has a function called *CountVectorizer*, which converts a collection of text documents to a matrix of token counts. It also has a function called *TfidfVectorizer* which evaluates how important a word is to a document in a corpus. *Tfidf* is an acronym for *term frequency-inverse document frequency* and used both frequencies to calculate the relevance of each word (H. C. Wu, Luk, Wong, & Kwok, 2008). For both functions, the result is a pandas data frame which can be used for later input in the learning algorithm of the text mining part. For the data analysis in this thesis only TF-IDF is used, and no word count.

4.1.3.3 Customer satisfaction data

The customer satisfaction data from Zorgkaart Nederland does not have to be prepared extensively as it is almost complete dataset. One thing that has to be done is to join the customer satisfaction dataset with the target variables to assure only the observations are left that are available in both datasets. Additionally, if the customer satisfaction dataset is used in combination with one or more of the other datasets, it has to be combined and prepared that each observation is available in each of the used datasets.

4.1.4 Data modelling

For the data modelling phase, several steps are taken before the data is put into the machine learning models as the research focuses on multiple combinations and applications of the datasets.

For the financial statement data, three different forms of the dataset are used as input data: balance sheet, income statement, and balance sheet combined with the income statement. Furthermore, each of these datasets are split up in two different versions again: the normal version of the dataset, and an optimized version of the dataset where columns that are a sum of other columns are removed from the dataset. The balance sheet dataset has 15 variables, the income statement 19 variables, and the combined therefore 34 variables. For the optimized version of the dataset, 12 variables are used from the balance sheet, 13 from the income statement, and therefore 25 from the combined set.

Furthermore, combinations of datasets are created with the balance sheet and/or income statement as starting point. So, the patient ratings and textual annual reports datasets are never used alone for the prediction model. In total these are twelve different datasets.

To cross check the results and see if there are big differences between them, three types of k-fold cross validation are used: 3-fold, 5-fold, and 10-fold. An overview of all the different options per category is shown in Figure 12. For each of the following machine learning models, the Python package sklearn is used.

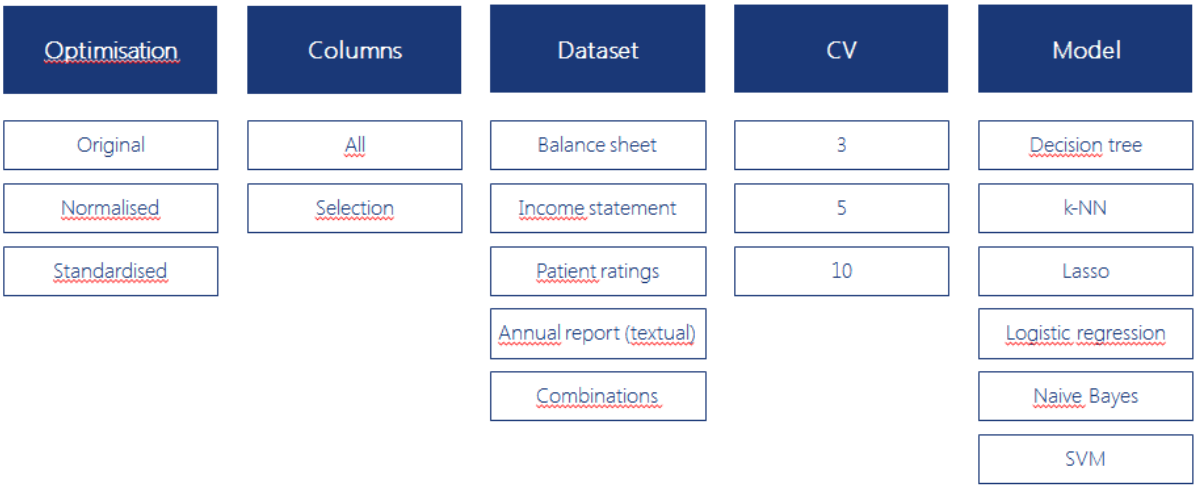

Figure 12. Overview of different variations per category in the model

4.1.5 Evaluation

For the evaluation of each model and their variations, already available methods from the sklearn package are used to obtain the evaluation metrics. The cross_val_score() function is used to split the dataset into training and test set, cross validate them, and score them based on the entered evaluation metric. All available scoring metrics are calculated for each model, but in this thesis the area under the curve (AUC) is used for each model to evaluate them with each other. The AUC is used as evaluation metric due to the fact that other metrics are not robust to

change of class distribution in the test and the area under the curve score being insensitive for this phenomenon (S. Wu & Flach, 2005). Each metric is initially composed by creating a confusion matrix (see Table 7) and calculating the evaluation metric based on it. Formulae 5.1 up until 5.9 are examples of what can be calculated from the confusion matrix, formula 5.9 being the formula to calculate the area under the curve.

Table 7. Confusion matrix to be used as evaluation matrix

		Predicted class	
		A	B
Actual class	A	True positive (TP)	False negative (FN)
	B	False positive (FP)	True negative (TN)

$$TP_{rate} = \frac{TP}{TP+FN} \quad \textbf{[5.1]}$$

$$TN_{rate} = \frac{TN}{TN+FP} \quad \textbf{[5.2]}$$

$$FP_{rate} = \frac{FP}{TN+FP} \quad \textbf{[5.3]}$$

$$FN_{rate} = \frac{FN}{TP+FN} \quad \textbf{[5.4]}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \textbf{[5.5]}$$

$$Precision = \frac{TP}{TP+FP} \quad \textbf{[5.6]}$$

$$Recall = \frac{TP}{TP+FN} \quad \textbf{[5.7]}$$

$$F_1 = 2 * \frac{precision*recall}{precision+recall} \quad \textbf{[5.8]}$$

$$AUC = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} 1_{p_i > p_j} \quad \textbf{[5.9]}$$

5. Results

This chapter focuses on the results of the conducted data analysis for sub-research question 4: which machine learning techniques perform best for predicting financial health? The different machine learning techniques are compared, as are the different datasets, and the most important variables for predicting financial distress at Dutch general hospitals are identified.

5.1 Overview analysis

The main focus of the data analysis were the machine learning techniques and the different datasets. However, next to these two main categories, the analysis also used different techniques to optimize the results. The following paragraphs explain the different categories shortly.

5.1.1 Supervised machine learning techniques

For the machine learning techniques category, six different techniques were used (see Table 8). These supervised machine learning techniques were identified in the literature section about classification with supervised machine learning techniques (see Chapter 3.3). The techniques are all used while keeping their default parameters, only their random state is specified when it is possible due to reproducibility purposes.

Table 8. Overview of used machine learning techniques, their abbreviations,

Technique	Abbreviation
Decision tree	DT
Logistic regression	LR
Support vector machine	SVM
k-Nearest-Neighbors	k-NN
Lasso	Lasso
Naïve Bayes	NB

5.1.2 Datasets

For the data analysis four datasets are used as input for the prediction of financial distress (see Table 9). Two of these datasets consists of financial data, the balance sheet and income statement, which are both the consolidated version. The patient ratings are customer satisfaction ratings that can be uploaded online by the patients voluntarily. The last dataset, the textual annual report, consists of TF-IDF scores for the fifty most important words in the corpus of annual reports. An overview of all variables per dataset is shown in Appendix A, table 18 and 19. Additionally, combined datasets are used as well, but always with either the balance sheet and/or income statement as a starting point. Therefore, an amount of 12 (combinations of) datasets are used and evaluated in the analysis.

Table 9. Overview of used datasets, their abbreviations, and the number of variables per dataset

Dataset	Abbreviation	Number of variables	Source
Balance sheet	BS	15	DigiMV
Income statement	IS	19	DigiMV
Patient ratings	PR	2	Zorgkaart Nederland
Textual annual report	tAR	50	CIBG/VWS

5.1.3 Additional categories for optimization

Apart from the two main categories, three other categories are used to optimize the analysis: optimization of input datasets, feature selection, and cross validation.

The optimization of data is done by giving all variables the same scale for the purpose of making each variable as (un)important as the other. For the optimization of the input datasets, the following three different techniques were used: (1) keeping the original dataset; (2) standardisation; and (3) normalisation.

Feature selection is done for the balance sheet and income statement as some variables are the sum of other variables in the dataset. The variables that consist of the sum are removed from the dataset, because these are superfluous as the relation should be recognised by the machine learning technique. Therefore, the datasets that contain the balance sheet and/or income statement have two versions each: one with all variables, and one with the superfluous variables removed.

The last category that is used to optimise the analysis is k-fold cross validation. For the k-fold cross validation the following types are used in this analysis: (1) 3-fold; (2) 5-fold; and (3) 10-fold.

5.1.4 Evaluation metric

The metric that is used for evaluation in this data analysis is the area under the roc-curve, in short: AUC-score. The area measures discrimination, which is the ability of the test to correctly classify those with either class (Tape, 2006). Additionally, it is used due to the fact that other metrics are not robust to change of class distribution in test set in comparison to the training set (S. Wu & Flach, 2005). However, the area-under-the-curve is insensitive for this phenomenon.

5.2 Model prediction

In this paragraph the results of the different prediction models are compared and evaluated. We start with comparing the results for predicting financial distress of Dutch general hospitals one year in the future (x + 1) as this is the largest dataset in this research. For every additional year for predicting the distress (so x + 2, x + 3, etc.), the dataset shrinks in size. For this, the curse of dimensionality have to be kept in mind, which is a problem caused by the exponential increase in volume associated with adding dimensions (Keogh & Mueen, 2011).

5.2.1 By dataset

The first result that has come out of the data analysis is that the balance sheet and income statement have the best average AUC scores of all possible combinations of datasets (see Figure 13, average AUC score per dataset). The balance sheet has a score of 0.66 and the income statement just a bit lower with 0.64. Interesting to see is that a combination of these two datasets scores lower (AUC score = 0.56) although it has the same variables as the two datasets apart from each other. When we look at the datasets where patient ratings are added to the initial dataset of BS and IS, we can see that in both cases the score has decreased (for BS from 0.66 to 0.58, for IS from 0.64 to 0.56). However, when the textual data (tAR) is added to the initial datasets, the scores decrease even more than the previous case. The score for the BS decreased from 0.66 to 0.54, and for the IS from 0.64 to 0.52.

Table 10. Differences in score for adding new datasets

Initial dataset	Additional dataset	Initial score	New score	Difference
Balance sheet	Patient ratings (PR)	0.66	0.58	-0.08
	Textual annual report (tAR)	0.66	0.54	-0.12
	PR + tAR	0.66	0.55	-0.11
Income statement	Patient ratings (PR)	0.64	0.56	-0.08
	Textual annual report (tAR)	0.64	0.52	-0.12
	PR + tAR	0.64	0.53	-0.11
Balance sheet	Income statement	0.66	0.56	-0.10

As can be seen in Table 10, the differences in adding the patient ratings data is -0.08 in both cases (adding to balance sheet and to income statement). Adding the textual data has an even larger difference with -0.12 in both cases. Interesting to see is that the difference in score for adding both the patient ratings and textual data is smaller than only adding textual data to the initial dataset (in both cases the difference is -0.11).

These first results are interesting and curious at the same time, because normally additional input variables would increase the AUC score and other evaluation metrics for machine learning techniques. As we look closer at the datasets, we can say the following about these results. The dataset can be described as small in comparison to datasets that are normally used in machine learning as the dataset of the balance sheet and income statement both only have 353 entry rows compared to thousands that are normally used to train the machine learning model. When other data is added the dataset gets smaller in size even more, to 261 entry rows when you combine all datasets (see Table 11). In return the amount of variables only increases when adding datasets to the initial datasets, probably causing a dimensionality problem and scoring lower as a result. Therefore, we can assume that the reason that the initial datasets (BS and IS) predict best individually in comparison to the other combinations of datasets is that they have the largest training set and the least variables. Which means that the assumption can be made as well that the effect of the dimensionality problem influences the initial datasets less than the other combined datasets.

Table 11. Overview of amount of entry rows and variables per dataset for x + 1

Dataset	Entry rows	Variables
Balance sheet (BS)	353	15
Income statement (IS)	353	19
BS + IS	353	34
BS + patient ratings (PR)	336	17
IS + patient ratings (PR)	336	21
BS + IS + PR	336	36
BS + textual annual report (tAR)	270	65
IS + textual annual report (tAR)	270	69
BS + IS + tAR	270	84
BS + PR + tAR	261	67
IS + PR + tAR	261	71
BS + IS + PR + tAR	261	86

5.2.2 By optimization

If we look closer at the average scores and split each dataset up by the optimization technique that is used, we can observe the following. In general, when the normalised and standardised datasets are used as input variables in this research, the average AUC score is lower than when the original dataset is used (see Figure 13). For both the normalised and standardised datasets, the average score is approximately 0.54 in comparison to an average score of 0.62 for the original dataset.

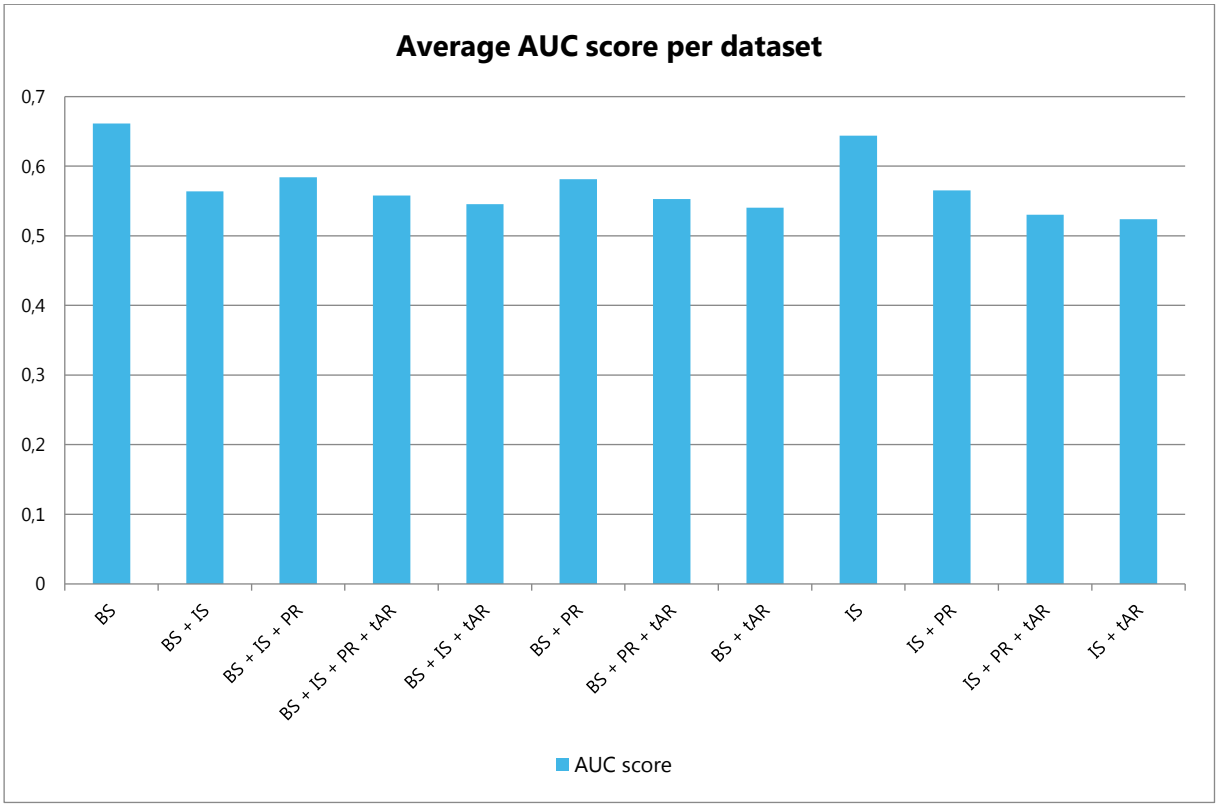

Figure 13. Average AUC score per dataset for x + 1

If we take a look at Figure 14, we can see that the datasets where a dataset other than the original dataset scores highest are the ones that scored highest as well on average: the balance sheet and income statement. Also, the income statement plus patient ratings has the best score when it is standardised. Furthermore, the same pattern is seen in this figure as in Figure 13 where the general average AUC scores are displayed: the datasets that include the textual annual reports score lowest, and the ones with patient ratings somewhere in between. The question why the original dataset scored higher on average than the normalised and standardised versions is hard to answer. A pretty solid assumption can be made by assuming that yet again the problem is in dimensionality. In general, normalised and standardised data predicts better than original data due to the fact that each variable has the same scale, where the original data has its own scale for every variable. One thing we can conclude from this results is that standardised data performs best in comparison to the other optimisations for the best predicting datasets: balance sheet and income statement.

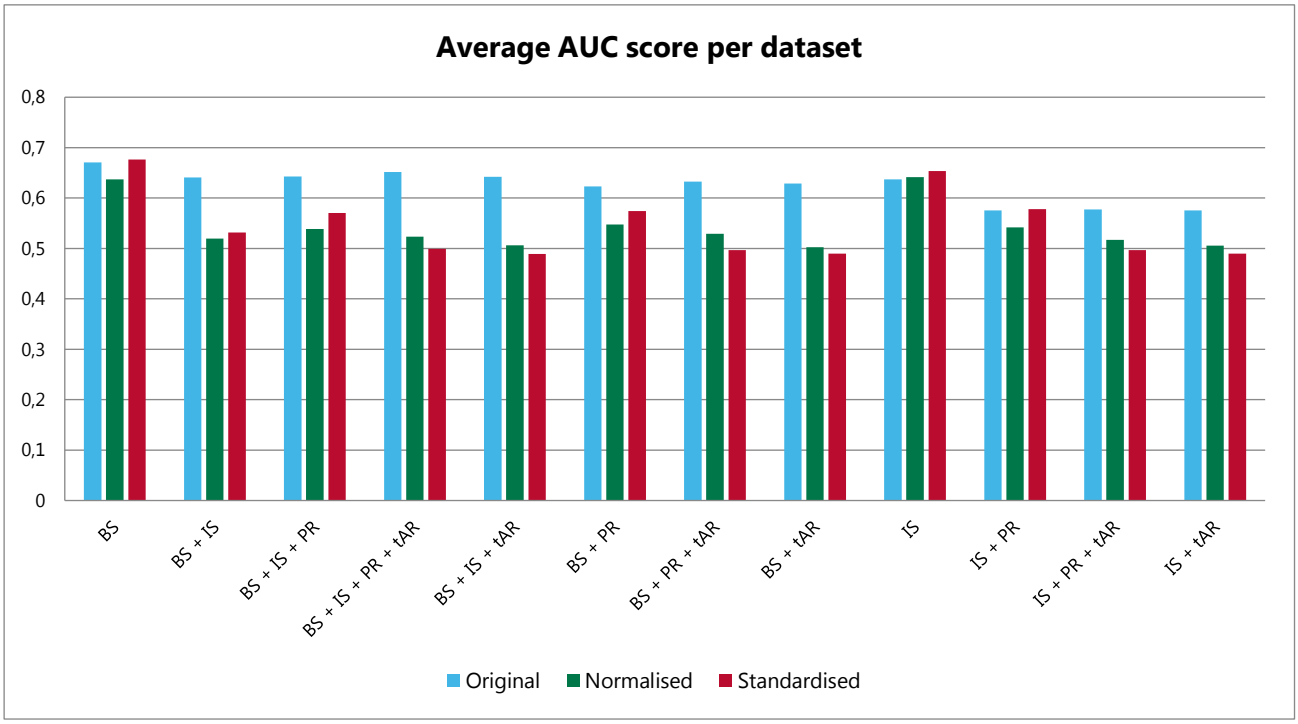

Figure 14. Average AUC score per dataset for x + 1, categorised by optimisation of dataset

5.2.3 By machine learning model

This section focuses on the comparison of average AUC score per machine learning model for predicting one year ahead in the future. An overview of these average scores is given in Figure 15, where we can see immediately that the scores vary from a low of 0.51 for the Naïve Bayes model to a high of 0.59 for the k-NN model. If we look how these model scores are divided over the different datasets that are used (see Figure 16), it can be confirmed that k-NN is the best performer on average. However, it does not have the best models in this research at all, instead most other models perform better when predicting with the top two datasets, balance sheet and income statement. The fact that its average score is highest is due to the fact that it performs better on the datasets that predict worse on average.

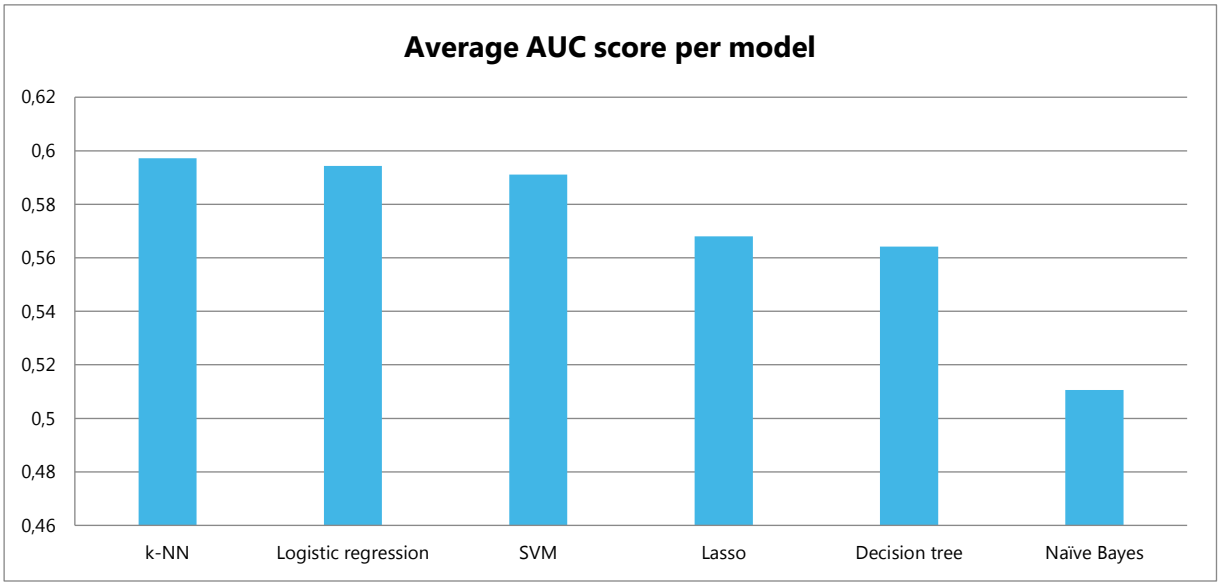

Figure 15. Average AUC score per model for x + 1

If a closer look is taken at Figure 16, it can be observed that most of the supervised machine learning models perform best again in combination with the balance sheet and income statement individually. The only exceptions are a decision tree model in combination with the balance sheet, income statement and patient ratings data, which performs better than the same model with only the balance sheet as input data. Additionally, most of the k-NN models have an AUC score around 0.60, with only the balance sheet data that performs really higher. The Naïve Bayes models perform worst for all combinations of datasets but one.

The best performing individual models are the lasso and logistic regression models for the balance sheet and income statement. Lasso performs best with both balance sheet and income statement, scoring an average AUC of 0.75 and 0.70 correspondingly. Logistic regression performs second-best with 0.70 and 0.69 for the balance sheet and income statement correspondingly. The support vector machine performs relatively good as well with 0.69 in combination with the balance sheet and 0.67 in combination with the income statement.

It is interesting to see that the lasso regression model performs best in combination with the individual best two datasets, but it can be explained relatively easy. It has been observed in the previous paragraphs that dimensionality of the data is a problem in this research, therefore predicting bad with most combinations of datasets. The models with the least amount of variables predict best, probably due to the fact that the dimensionality problem is least with these datasets. When you keep this in mind and add a lasso regression model, it makes sense that it performs better than other models. Lasso penalizes variables and can even set variables to a predictive power of 0, meaning the variable is not included in the prediction model. In the next paragraph, we take a closer look into the lasso models in combination with the balance sheet and income statement individually.

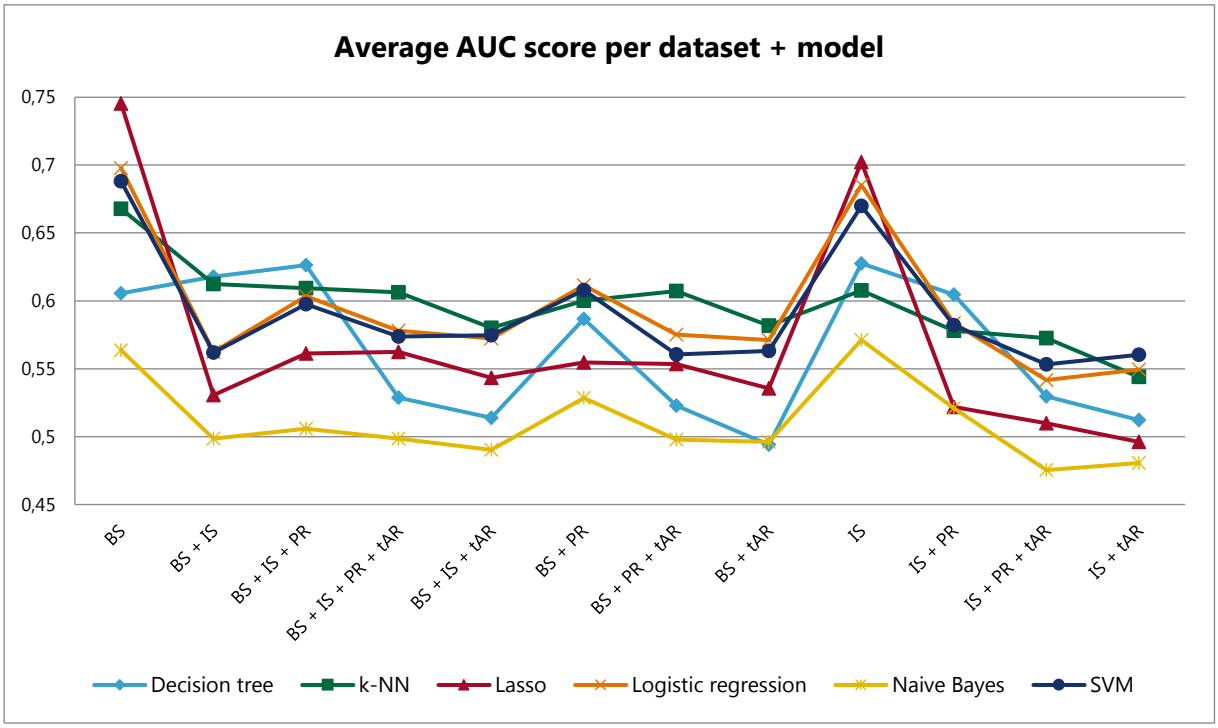

Figure 16. Average AUC score per dataset and model for x + 1

5.2.4 Best performing models

In the previous section it has been observed that lasso regression in combination with the balance sheet and income statement datasets individually perform best in comparison to the other variations in this research.

If we take a closer look at the lasso regression results (see Figure 17), we can confirm the findings from the last paragraph: lasso predicts best with the balance sheet and income statement. Figure 16 shows that it does not make a large difference in AUC score when different k-fold cross validation is used. For the balance sheet the score is approximately 0.74 and for the income statement between 0.69 and 0.71. As the figure shows, the score for the other datasets does not come above the 0.6 line.

Therefore, a closer look is taken into the most important variables of the two best performing models. At first, the best model of all is identified, which is with the balance sheet as dataset, not-optimized, a selection of the variables, and with 3 fold cross validation. The AUC score for this combination of parameters is 0.77. The most important variables, and only variables used in this model are (1) totally equity, (2) total current assets, (3) tangible fixed assets, and (4) long-term liabilities (see Figure 18). Only four of in total twelve variables are used to predict the financial distress in this model. The complete balance sheet contains fifteen variables, but due to the feature selection three variables are removed. Additionally, as lasso is used as machine learning model, variables are weighted by the model and can even be penalised to zero. So we can observe that the lasso model penalised eight variables to zero, meaning they are of no importance in this model. The variable with the highest positive effect is the total equity, which can indeed be explained because in general when you have a high total equity, you often have a lot of capital in your organisation relatively. On the other hand, long-term liabilities is the variable that has the highest negative coefficient. This can be explained relatively easy, as liabilities are something an organisation does not have within their organisation, but is a debt instead.

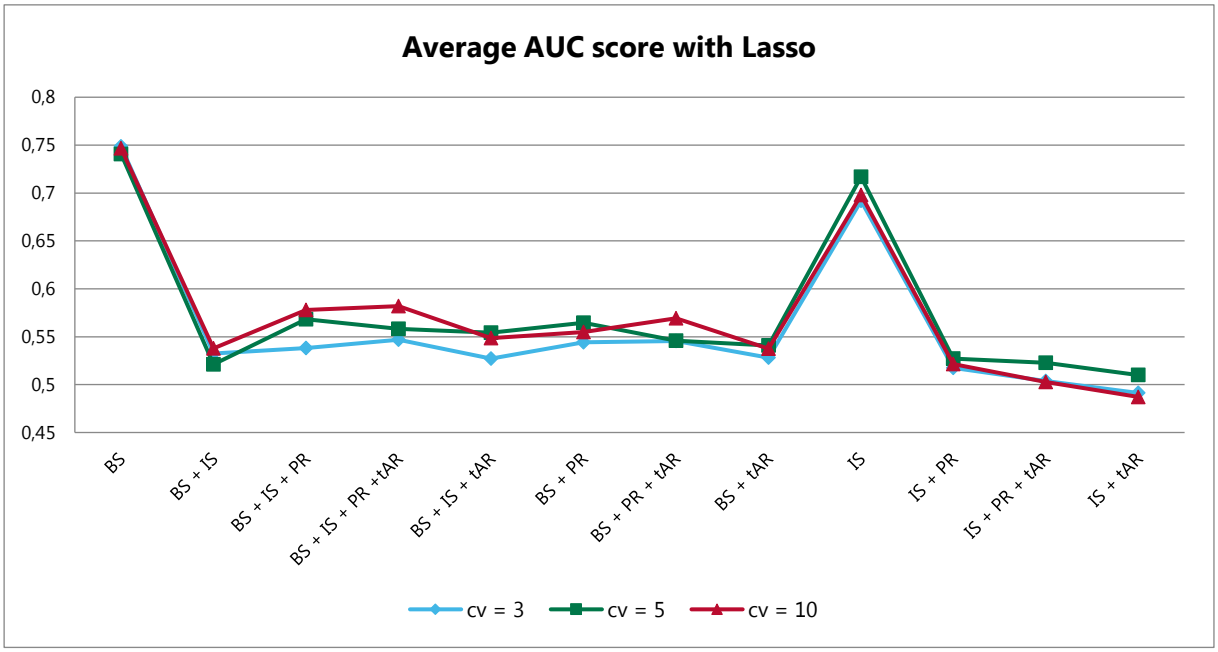

Figure 17. Average AUC score per dataset with Lasso for x + 1, categorised by k-fold cross-validation

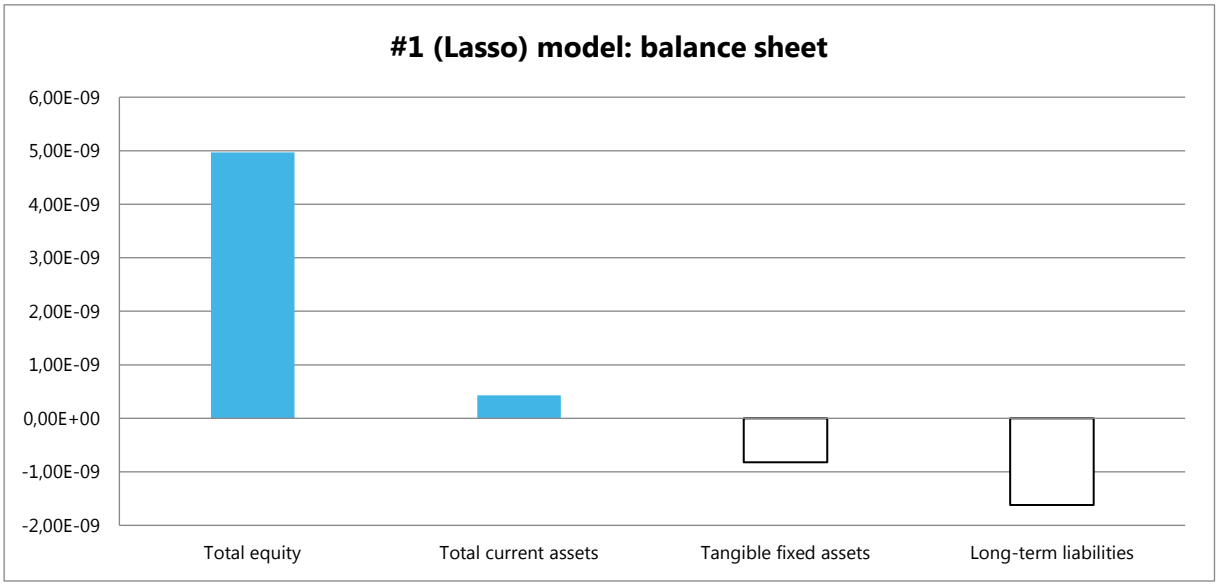

Figure 18. Coefficients of the best prediction model: lasso/balance sheet/not-optimised/selection of variables/3-fold cross validation.

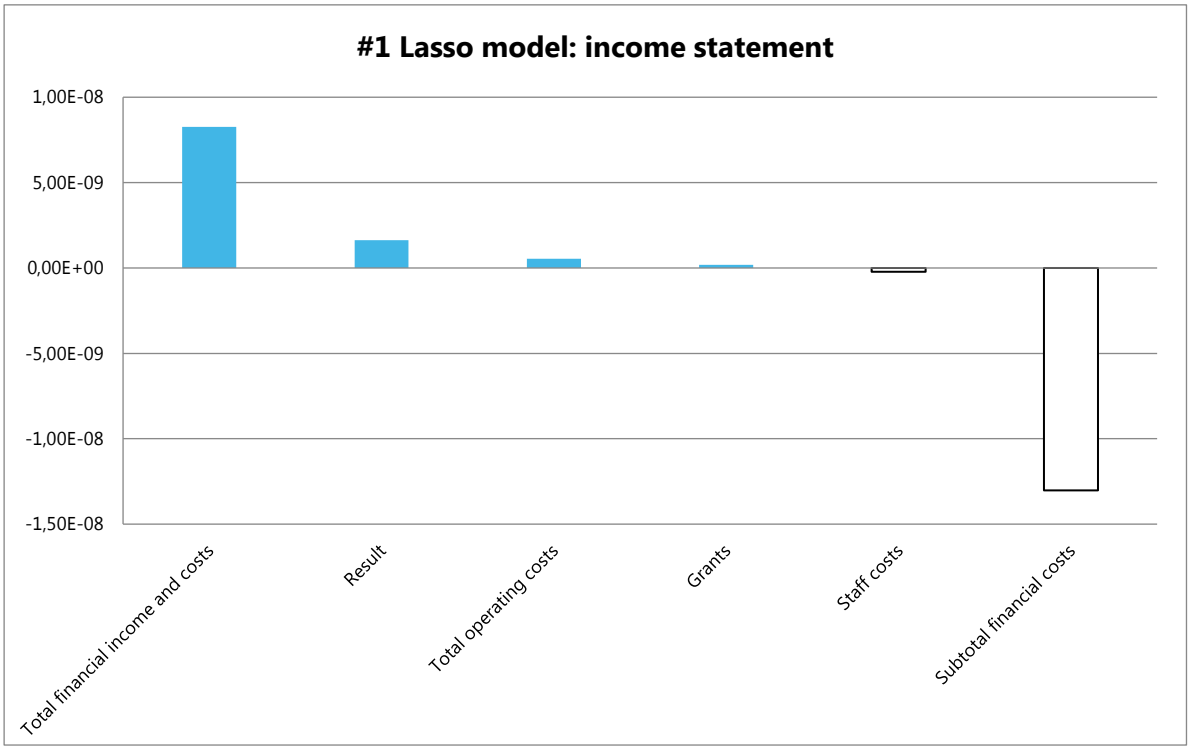

Figure 19. Coefficients of the best prediction model for income statement: lasso/ sheet/not-optimised/all variables/5-fold cross validation.

The best prediction model for the income statement was with a lasso regression model as well, scoring 0.73 for AUC. The income statement contains nineteen variables in total of which only six are used in the best performing income statement prediction. The highest positive coefficient in the model is the total financial income and costs, the highest negative coefficient being the subtotal financial costs (see Figure 19). Just like the best performing balance sheet model, the best performing income statement model includes only a fraction of the total amount of variables of the specific dataset. This means that the lasso model penalises thirteen of the nineteen variables of the income statement to zero.

5.2.5 Evaluation

The results of the previous paragraphs show that the best predicting models are lasso regression models with the balance sheet and income statement data individually, where the first predicts a little better in general. The best balance sheet model has an AUC score of 0.77, where the best income statement model has an AUC score of 0.73. Most variables are penalised to zero by the lasso regression models that perform best, leaving only four variables for the best predicting balance sheet model and six for the best predicting income statement model. This means that most of the variables in these models have no predicting power at all. It can be concluded that with a low amount of variables the predictions are better than with a lot of variables. This can be related to the dimensionality problem mentioned in the beginning of the chapter.

5.3 Financial ratios as input for prediction

This paragraph focuses on predicting financial distress with the six financial ratios from the stress test as input variables instead of using datasets with more variables like in the previous chapter. The decision to use the financial ratios as input variables came from the results of the main data analysis of this research as the best prediction model only included a small amount of variables. Therefore, the financial ratios of the stress test were used to evaluate if they predict the financial distress of hospitals better. The expectation was that it predicts better than the models from the main data analysis in this research as it has a small amount of variables, but variables that are important for financial experts as they are composited numbers from various other financial figures.

The data analysis has not been done all over again completely due to the time necessary to compose and evaluate the results. Therefore, it has been chosen to use the following parameters for all the previously used models: 10-fold cross validation, once standardized variables and once the original dataset. The 10-fold cross validation has been chosen due to the fact that it has the largest amount of folds of the ones used in the main data analysis and therefore is tested more often and on different folds. In the main data analysis part of this research the best models included the original dataset and therefore we wanted to see if this dataset predicted best again. However, the expectation is that standardised variables predict better as each variable is as important as the other ones initially. In the original dataset, variables that have high values will probably have more impact on the prediction than variables that have low values.

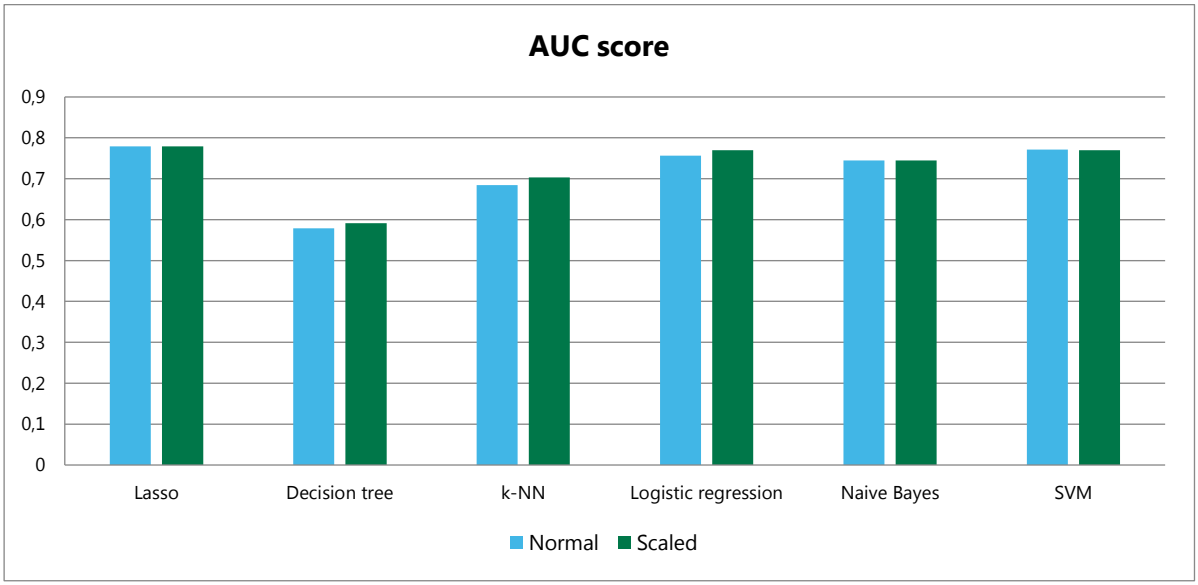

Figure 20. AUC score per model for x + 1 using financial ratios as input variables

As can be seen in Figure 20, the Lasso predicts best again of all used models with an AUC score of 0.78 for both the original/normal dataset and standardised dataset. However, it shows that most other models score higher compared to their scores in the main data analysis of this research. As can be seen in Figure 15, the average AUC score for each model was below 0.6 in the main data analysis. Using the six financial ratios of the stress test as input variables increased most scores to 0.7 or higher. The only model that scores under 0.6 for both original and standardised dataset is the decision tree model. The Naïve Bayes model has increased from just over 0.5 to almost 0.75. Also, the support vector machine model has increased from just under 0.6 to 0.77 for both original and standardised dataset and scores almost as good as the lasso model.

These results are already better than the results from the main analysis of the research, but it has to be put into perspective as we have optimised the model in this paragraph based on the results of the main data analysis. Therefore, it can be observed that the models with the six financial ratios of the stress test perform better than the models from the main analysis. Additionally, the results in Figure 20 show that the standardised datasets perform better than the original datasets in many cases.

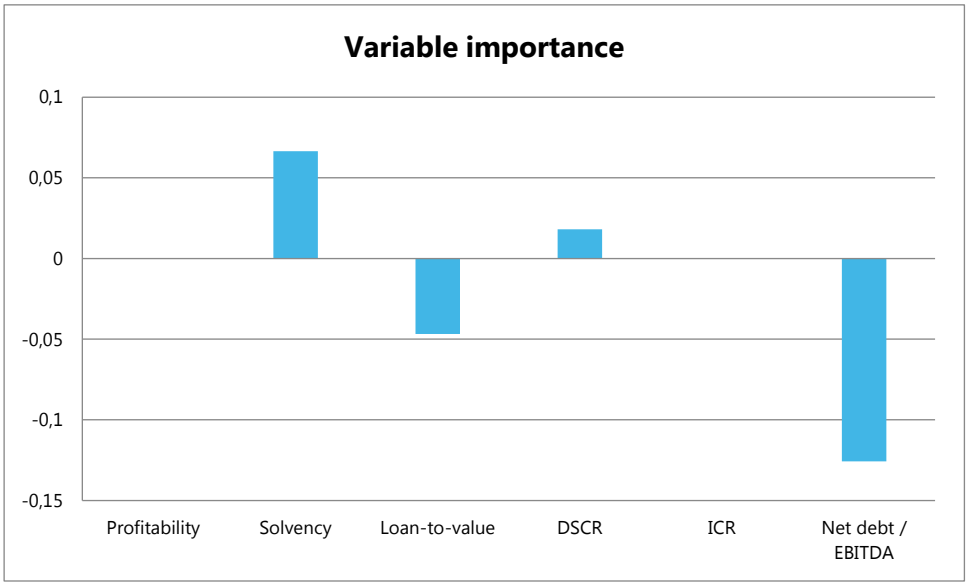

Figure 21. Variable importance of best model (lasso) in second analysis

As can be seen in Figure 21, the financial ratios of profitability and ICR are penalised to zero in the best model in the second analysis, using only financial ratios as input data. The Net debt / EBITDA ratio has the highest coefficient in this model, the remaining three definitely having lower coefficients each.

5.3.1 Financial ratios and patient ratings

As one of the hypotheses is that patient ratings improve the prediction of financial distress at hospitals, the patient ratings are added to the financial ratios to evaluate if they do improve the prediction with these models. Table 12 shows that the patient ratings only improve the AUC score for two of the six models with both the original and standardised dataset. Most are approximately the same when using solely financial ratios as when patient ratings are included as well, and in two cases the score decreases significantly (normal k-NN and SVM). The AUC scores for the dataset with patient ratings included is displayed in Figure 22.

Table 12. Overview of AUC scores for financial ratios only and financial ratios + patient ratings.

	Normal		Standardised	
	FR	**FR + PR**	**FR**	**FR + PR**
Lasso	***0,779***	0,766	***0,779***	0,776
Decision tree	0,579	***0,606***	***0,592***	0,583
k-NN	***0,685***	0,621	***0,703***	0,678
Logistic regression	0,756	***0,773***	0,770	***0,774***
Naive Bayes	0,745	0,745	***0,745***	0,743
SVM	***0,771***	0,664	0,770	***0,773***

If we compare these numbers to Table 10, which shows the difference in AUC score when other datasets are added to either the balance sheet or income statement, it can be observed that financial ratios with patient ratings perform better than another dataset with patient ratings included. For both the balance sheet and income statement, adding the patient ratings to the input variables meant that the score decreased 0.08. This is definitely not the case with financial ratios and the inclusion of patient ratings.

The most important variables in the best model with the addition of patient ratings is approximately the same as without them. The profitability and ICR are again penalised to zero and the net debt / EBITDA ratio has the highest coefficient. However, the patient rating and the amount of patient ratings are not penalised to zero, but have very small coefficients.

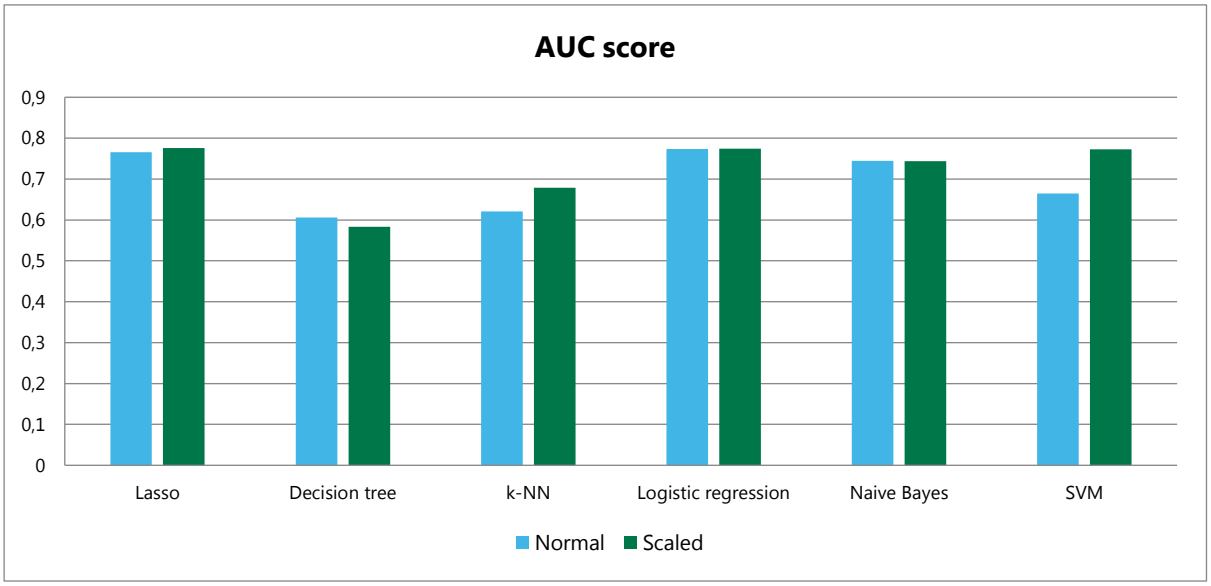

Figure 22. AUC score per model for x + 1 using financial ratios and patient ratings as input variables

5.4 Naive prediction

To evaluate if the predictions with machine learning models have an added value in comparison to a naive prediction, an alternative prediction is made based on the situation of financial distress of the previous year. A confusion matrix is created from this prediction method, displayed in Table 13. So, the matrix is created by looking at the situation of financial distress (0 or 1) in year x, then using this to evaluate if the situation is the same in year x + 1. Therefore, the variables in x + 1 are the true target variables, and the variables in x the predicted variables.

Table 13. Confusion matrix for naive prediction

		x	
		1	0
x + 1	1	117	54
	0	46	119

The confusion matrix is used to calculate the sensitivity and specificity of the prediction, which are the true positive rate and false positive rate correspondingly and are the axes of a ROC curve. The true positive rate (TPR) is calculated and scores 0.68, whereas the false positive rate (FPR) is 0.28 based on the formulae in paragraph 4.1.5 of the Method chapter. The scikit-learn package in Python is used to calculate the AUC score, which results in an AUC score of 0.70.

If this score is compared to the scores from the main data analysis, we can conclude that only the best models perform better than this naive prediction. These models are foremostly based on the balance sheet and income statement and use a lasso model. However, if the score is compared to the results where the six financial ratios of

the stress test are used as input variables, it can be observed that four of the six models which use the financial ratios as input perform better than the naive prediction.

6. Discussion and conclusions

Benchmark outcomes in the Netherlands among general hospitals showed that almost a quarter of them is financially unhealthy (van den Haak, 2017). Additionally, the newly assigned government of the Netherlands has written in the governmental agreement that they hope to save 1.9 billion euro structurally for the upcoming years in the curative care (Rutte et al., 2017). Therefore, the financial outlook of the general hospital sector in the Netherlands is not very positive, providing the same quality of care with less resources.

Evaluating the financial distress of hospitals is already done by benchmarking, but they focus on the past and the present only. These benchmarks often use financial figures, including ratios, to compare the hospitals with each other (Zeller et al., 1996). However, the available financial data is not yet used to predict the future financial distress of the hospitals with the help of machine learning techniques, in comparison with other sectors where they already use these techniques to predict financial distress and bankruptcy (Geng et al., 2015; Kim & Upneja, 2014; Lin et al., 2015; López Iturriaga & Sanz, 2015). These researches have in common that they all use financial data to predict financial health/distress.

Therefore, building on the earlier research on predicting financial health/distress, this research focused on the prediction of financial distress at Dutch general hospitals with the use of financial data. This kind of research had not been done in the Dutch hospital sector yet, therefore being a gap in the literature.

Looking at the available (open) data sources for Dutch general hospitals, there were two large datasets accessible that were used in this research: annual reports (including financial statements) from DigiMV and patient ratings from Zorgkaart Nederland. Previous research has shown that using textual data from annual reports in combination with financial data predicted better than predictions from financial analysts (Qiu et al., 2014) and that customer satisfaction/patient ratings have a positive impact on financial health in other sectors (Chi & Gursoy, 2009; Kyoon Yoo & Ah Park, 2007; van der Wiele et al., 2002). Therefore, the textual data from annual reports and patient ratings data were added to this research in addition to the financial data that is used more commonly in predicting financial distress in other sectors.

The main purpose of the research is to investigate how a combination of machine learning and text mining techniques can be used to predict financial distress in the Dutch general hospital sector. The resulted in the following main research question:

> *How can a combination of machine learning and text mining techniques be used to predict financial distress at Dutch general hospitals?*

In order to answer the main research question, multiple sub-questions were formulated and answered first. In the remainder of this chapter, the sub-questions are discussed in the order they are mentioned in the introduction of this research. Eventually, the main research question will be answered at the end of the chapter.

6.1 Measuring financial distress (of hospitals)

For answering the first sub-question, a literature review has been conducted to identify ways for measuring financial distress. The following sub-question has been formulated and answered:

> *How is financial distress (of hospitals) measured?*

A lot of research has been conducted in accounting with regard to predicting bankruptcy from information in companies' financial statements, starting in the mid-1900s with the ratio for evaluating credit-worthiness: the current ratio (Beaver, 1966). Failure prediction models, related to financial distress prediction models, can be divided into three different categories: classical statistical model, artificially intelligent expert system models, and

theoretical models (Adnan Aziz & Dar, 2006). This research has been focussing on the second: artificially intelligent expert system models.

In the last decade, a lot of research has been conducted on predicting business failure/bankruptcy/financial distress, where all of them used financial ratios as input for the machine learning models (Azayite & Achchab, 2016; Chen, 2011; Holmes et al., 2017; Koyuncugil & Ozgulbas, 2012).

In the Dutch hospital sector several organisations perform benchmarks, assessing the financial situation of the hospitals (BDO, 2017; CBS, 2013; EY, 2017; NVZ, 2017). These benchmarks all make use of financial ratios, with the ratios of the stress test being used extensively. This stress test is used in the banking sector as well and contains six financial ratios: (1) profitability, (2) solvency, (3) loan-to-value, (4) DSCR, (5) ICR, and (6) the Net debt / EBITDA. These ratios tell something about result and performance of a company, as does it about the financial position and structure, and the interest- and repayment capacity.

Therefore, the six financial ratios from the stress test were used to determine the financial health of the hospitals in this research. As research did not show when a company or hospital is either financially healthy or in financial distress, an arbitrary line has been drawn to divide them. This line has been drawn in the middle of the six ratios, an assumption that has been made to continue the research. This is an assumption that has been made and has to be taken into account when evaluating the results from the predictive modelling.

6.2 Text mining on annual reports

The second sub-question focused on the extent in which text mining has been used before within the topic of predicting financial distress by conducting a literature review. The following sub-question has been formulated and answered:

To what extent has text mining been used for extracting information from annual reports?

Text analysis in the financial domain has grown in popularity since textual reports of companies and other institutions have been made publicly available online on the internet. Before, the financial domain had been focusing on numerical data.

Text mining of annual reports has been used for the prediction of financial health of companies. Research has shown that it is feasible to use text classification on annual reports to predict the financial health of a company (Qiu et al., 2006). In further research, they have shown that their prediction with financial ratios and annual reports using machine learning predicted better than the prediction of financial analysts (Qiu et al., 2014). Text mining annual reports for prediction models in the financial domain has also been used for identifying financial statement fraud (Goel & Gangolly, 2012; Hajek & Henriques, 2017; Sadasivam & Lakshme, 2016), bug report classification (Zhou et al., 2016), stock price movements (Doucette & Cohen, 2015), and future accounting and market performance (Balakrishnan et al., 2010).

Therefore, it can be concluded that text mining has been used extensively for extracting information from annual reports in the financial domain. Not only for financial distress prediction, but for various other types of prediction in the financial domain as well. Additionally, research has shown that it is feasible to use and can be a better predictor (in combination with financial data) than financial analysts' predictions.

6.3 Textual data as input for machine learning

The third sub-question focused on how textual data can be used as input for machine learning techniques. The following sub-question has been formulated and answered by conducting a literature review:

How can textual data, next to financial data, be used as input for machine learning techniques?

Extracting useful insights from unstructured textual data is more difficult than with other forms of data and is seen as the most challenging research aspect for fundamental data (Khadjeh Nassirtoussi et al., 2014). Whereas financial numbers and ratios contain information about the past, linguistic structure and text style may indicate how well an organization will do (Kloptchenko et al., 2004).

An overview of the various text mining phases is given in Figure 8, which is a text mining framework. The three main phases of the framework are pre-processing, text mining operations, and post-processing (Kobayashi et al., 2017). The focus of this sub-question was on the pre-processing phase of this framework, as this is the phase before the data is used in a machine learning model. This phase is split up in three different phases again: collection, cleaning, and transformation.

The most interesting phases of this are cleaning and transformation, because these phases clean and transform the data until it is ready for input in a model. In the text cleaning phase, the text is tokenised, stop words are removed, and stemming can be applied (Gaustad & Bouma, 2002; Hotho et al., 2005; Kobayashi et al., 2017; Uysal & Gunal, 2014). In the transformation phase, the text is transformed into mathematical structures, for example a document-by-term matrix (Scott & Matwin, 1999). This can either be a word count or a TF-IDF vectorizer, the latter evaluates how important a word is to a document in a corpus. It is an acronym for *term frequency-inverse document frequency* and uses both frequencies to calculate the relevance of each word (H. C. Wu et al., 2008).

It can be concluded that textual data can be used as input for machine learning techniques. However, the textual data has to be cleaned first, and later transformed to a mathematical structure called document-by-term matrix before it can be used as input.

6.4 Performance of machine learning techniques for predicting financial distress

The last sub-question focused on which machine learning techniques perform best for predicting financial distress by conducting a small literature review to identify suitable machine learning models and conducting a comprehensive data analysis afterwards. The following sub-question has been formulated and is answered in this paragraph:

Which machine learning techniques perform best for predicting financial distress?

Supervised machine learning techniques are used in this research as the output/target data is known. Supervised machine learning is split into two different categories: regression and classification. Classification is used when the output variable is a class or category. Each entry of a dataset is then e.g. divided into one of two classes: 'good' or 'bad'. As this is classification with two classes, it is called binary or binomial classification (Lin, F. et al., 2014).

The supervised machine learning techniques that have been used in this research is a selection of the techniques used in (Kos et al., 2017): logistic regression, support vector machine, decision tree, naïve Bayes, and k-nearest-neighbor. A lasso model has been added to the research as well as it is a popular method for both variable selection and regularization (Meier et al., 2008; Tibshirani, 1996).

The main data analysis of this research has shown that a lasso model in combination with the balance sheet as input data predicts financial distress of Dutch general hospitals best with an AUC score of 0.77. According to (Tape, 2006), this is evaluated as a fair score. However, this is only one model and the average AUC score for

predicting x + 1 was lower than 0.6 for each machine learning model and for most of the datasets used in this research. Below 0.6 is evaluated as a fail, and between 0.6 and 0.7 as poor.

As most of the best predicting models in this research used lasso and penalised most of the variables of the dataset to zero, the data in this research probably has a dimensionality problem. Adding additional variables like the patient ratings and textual data from the annual report made the model score an even lower AUC score in general. Therefore, an additional exploratory analysis has been conducted using only the financial ratios of the hospitals as input data. The results of this additional analysis were better or just as good as in the initial research for almost all machine learning models. Adding the patient ratings to the input data increased the AUC score for some of the models, but lowered the score for others.

A alternative naive prediction has been added to the analysis as well to show the added value of the machine learning models in comparison to such a naïve prediction. This prediction used the data from year x as predicted values and year x + 1 as true values, investigating if just looking at the previous year is a better prediction model than the use of machine learning models. The AUC score for this approach was 0.70, so it can be concluded that the best predicting models of the main data analysis and most of the additional analysis perform better than this naïve way of predicting financial distress.

It can be concluded as well that the predictions in the main data analysis of this research were poor in general, with some exceptions where the models scored fair. The main cause of these low scores is probably a dimensionality problem of the data in this research. The additional data analysis showed that the models using the financial ratios of the stress test as input data performed better in most cases than the models in the main data analysis. Most models scored between 0.7 and 0.8 which can be evaluated as a fair result. However, these results have to be put into perspective as the models are optimised based on the results of the main data analysis.

6.5 Main research question

The main research question of this research was:

> *How can a combination of machine learning and text mining techniques be used to predict financial distress at Dutch general hospitals?*

The literature review has shown that it is feasible to use both textual data and financial data to predict financial distress and sometimes even performs better than the predictions of financial analysts. For the prediction of financial distress, financial ratios are often used as input variables for the machine learning model. Adding textual data from annual reports to prediction models is done on regular basis as well, also for other predictions in the financial domain. However, if textual data is used it has to be pre-processed first by cleaning and transforming the data into mathematical structures, so it can be used as input data for machine learning.

The main data analysis showed that in this research a lasso model in combination with the balance sheet as input data predicted the financial distress of Dutch general hospitals best. Adding patient ratings and/or textual data from annual reports lowered the scores in most cases, probably due to a dimensionality problem in the data. The average results can be called poor for each model in the main data analysis as well.

Therefore, further exploratory data analysis has been done with the financial ratios as input data, which showed better results for most of the machine learning models. Adding patient ratings to this dataset showed a better result in some cases, but also worse results in others.

From literature it can be concluded that a combination of machine learning and text mining techniques can be used for predicting financial distress. However, the results from the data analysis in this research showed that the best prediction models only had a fair score (an AUC score between 0.7 and 0.8) and the average scores were

poor (0.6 to 0.7) to a complete fail (0.5 to 0.6). The main cause of these low scores is probably the small size of the dataset in combination with a lot of variables, causing a dimensionality in most cases.

6.6 Further research and discussion

As this research had a dimensionality problem it is of importance that further research on predicting financial distress at Dutch general hospitals is conducted on a larger dataset. However, the problem is that there are not many hospitals in the Netherlands which means that the data has to come either from back in the past, a similar sector like the long-term care, or hospitals from abroad.

For a further evaluation of the outcomes of this research, meetings have been hold with two financial managers of hospitals in the Netherlands (Expert 1 of Tergooi, Hilversum & Expert 2 of CWZ, Nijmegen). They both showed great interest in this research as their departments spend a lot of time predicting their financial situation in the near future (1-4 years ahead). Both use historical data for their prediction, but not only data from the financial statement. Other data they look at are demographics of their region, staff costs, condition of their real estate, the agreements between hospital sector/patient federation/health insurers/etc. in the 'Hoofdlijnenakkoord' and other hospital related data. Patient ratings and textual data are not yet used in their predictions, neither are machine learning techniques. Although they think the topic can be of great importance in the near future, they only see it as a check on their own predictions at the moment. Therefore, future research can look at these data as well as the hospital financial managers both use them for their own predictions now.

7. References

Adankon, M. M., & Cheriet, M. (2009). Support Vector Machine. In S. Z. Li & A. Jain (Eds.), *Encyclopedia of Biometrics* (pp. 1303–1308). Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-73003-5_299

Adnan Aziz, M., & Dar, H. A. (2006). Predicting corporate bankruptcy: where we stand? *Corporate Governance: The International Journal of Business in Society*, *6*(1), 18–33. https://doi.org/10.1108/14720700610649436

Azayite, F. Z., & Achchab, S. (2016). Hybrid Discriminant Neural Networks for Bankruptcy Prediction and Risk Scoring. *Procedia Computer Science*, *83*(Ant), 670–674. https://doi.org/10.1016/j.procs.2016.04.149

Balakrishnan, R., Qiu, X. Y., & Srinivasan, P. (2010). On the predictive ability of narrative disclosures in annual reports. *European Journal of Operational Research*, *202*(3), 789–801. https://doi.org/10.1016/j.ejor.2009.06.023

BDO. (2017). *Alles onder controle ? Opbrengsten stijgen , marges blijvend onder druk inzichten nieuwe*.

Beaver. (1966). Financial Ratios As Predictors of Failure Authors (s): William H . Beaver Source : Journal of Accounting Research , Vol . 4 , Empirical Research in Accounting : Selected Published by : Wiley on behalf of Accounting Research Center , Booth School of Busi, *4*(1966), 71–111. Retrieved from http://www.jstor.org/stable/2490171

Beaver, W., McNichols, M., & Rhie, J. W. (2005). Have financial statements become less informative? Evidence from the ability of financial ratios to predict bankruptcy. *Review of Accounting Studies*, *10*(1), 93–122. https://doi.org/10.1007/s11142-004-6341-9

Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics*, *77*(2), 81–97. https://doi.org/10.1016/j.ijmedinf.2006.11.006

Bhatia, N., & Author, C. (2010). Survey of Nearest Neighbor Techniques. *IJCSIS) International Journal of Computer Science and Information Security*, *8*(2), 302–305. https://doi.org/10.1377/hlthaff.2014.1186

Bird, S., Loper, E., & Klein, E. (2009). *Natural Language Processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media Inc.

Blockeel, H., & Leuven, K. U. (2017). *Encyclopedia of Machine Learning and Data Mining*. https://doi.org/10.1007/978-1-4899-7687-1

CBS. (2013). *Financiële kengetallen zorginstellingen 2012*. *Centraal Bureau voor de Statistiek*.

Chen, M. Y. (2011). Predicting corporate financial distress based on integration of decision tree classification and logistic regression. *Expert Systems with Applications*, *38*(9), 11261–11272. https://doi.org/10.1016/j.eswa.2011.02.173

Chi, C. G., & Gursoy, D. (2009). Employee satisfaction, customer satisfaction, and financial performance: An empirical examination. *International Journal of Hospitality Management*, *28*(2), 245–253. https://doi.org/10.1016/j.ijhm.2008.08.003

Clatworthy, M. A., & Jones, M. J. (2006). Differential patterns of textual characteristics and company performance in the chairman's statement. *Accounting, Auditing & Accountability Journal*, *19*(4), 493–511. https://doi.org/10.1108/09513570610679100

Cristiannini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press.

Domingos, P., & Pazzani, M. (1997). On the Optimality of the Simple Bayesian Classifier under Zero-One Los. *Machine Learning*, *29*(1), 103–130. https://doi.org/10.1023/A:1007413511361

Doucette, J. A., & Cohen, R. (2015). Content of Annual Reports as a Predictor for Long Term Stock Price Movements, 416–421.

EY. (2017). Barometer Nederlandse Gezondheidszorg 2017, 16. Retrieved from http://www.ey.com/Publication/vwLUAssets/EY-barometer-nederlandse-gezondheidszorg-2017/$FILE/EY-barometer-nederlandse-gezondheidszorg-2017.pdf

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37. https://doi.org/10.1609/aimag.v17i3.1230

financial distress. (n.d.). In *Oxford Reference*. Retrieved from http://www.oxfordreference.com/view/10.1093/oi/authority.20110803095818367

Flach, P. (2012). *Machine learning: the art and science of algorithms that make sense of data*. Cambridge University Press.

Gaustad, T., & Bouma, G. (2002). Accurate Stemming of Dutch for Text Classification. *Language and Computers*, 1–14.

Geng, R., Bose, I., & Chen, X. (2015). Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research*, *241*(1), 236–247. https://doi.org/10.1016/j.ejor.2014.08.016

Goel, S., & Gangolly, J. (2012). BEYOND THE NUMBERS : MINING THE ANNUAL REPORTS FOR HIDDEN CUES INDICATIVE OF FINANCIAL STATEMENT FRAUD, *89*, 75–89. https://doi.org/10.1002/isaf

Hajek, P., & Henriques, R. (2017). Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods. *Knowledge-Based Systems*, *128*, 139–152. https://doi.org/10.1016/j.knosys.2017.05.001

Hinde, J. (2011). Logistic Normal Distribution. *International Encyclopedia of Statistical Science*, 754–755. https://doi.org/10.1007/978-3-642-04898-2_342

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, *1*, 278–282.

Holmes, G. M., Kaufman, B. G., & Pink, G. H. (2017). Predicting Financial Distress and Closure in Rural Hospitals. *Journal of Rural Health*, *33*(3), 239–249. https://doi.org/10.1111/jrh.12187

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A Brief Survey of Text Mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, *20*, 19–62. https://doi.org/10.1111/j.1365-2621.1978.tb09773.x

Huang, S. M., Tsai, C. F., Yen, D. C., & Cheng, Y. L. (2008). A hybrid financial analysis model for business failure prediction. *Expert Systems with Applications*, *35*(3), 1034–1040. https://doi.org/10.1016/j.eswa.2007.08.040

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, *31*(3), 264–323. https://doi.org/10.1145/331499.331504

Jiang, S., Pang, G., Wu, M., & Kuang, L. (2012). An improved K-nearest-neighbor algorithm for text categorization. *Expert Systems with Applications*, *39*(1), 1503–1509. https://doi.org/10.1016/j.eswa.2011.08.040

Kao, A., & Poteet, S. R. (Eds. . (2007). *Natural language processing and text mining*. New York, NY: Springer Science & Business Media.

Keisidou, E., Sarigiannidis, L., Maditinos, D. I., & Thalassinos, E. I. (2013). Customer satisfaction, loyalty and financial performance. *International Journal of Bank Marketing*, *31*(4), 259–288. https://doi.org/10.1108/IJBM-11-2012-0114

Keogh, E. ., & Mueen, A. (2011). Curse of dimensionality. In *Encyclopedia of machine learning* (pp. 257–258). Springer, Boston, MA.

Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, *41*(16), 7653–7670. https://doi.org/10.1016/j.eswa.2014.06.009

Kim, S. Y., & Upneja, A. (2014). Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models. *Economic Modelling*, *36*, 354–362. https://doi.org/10.1016/j.econmod.2013.10.005

Kloptchenko, A., Eklund, T., Karlsson, J., Back, B., Vanharanta, H., & Visa, A. (2004). Combining data and text mining techniques for analysing financial reports. *Intelligent Systems in Accounting, Finance & Management*, *12*(1), 29–41. https://doi.org/10.1002/isaf.239

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2017). *Text Mining in Organizational Research*. *Organizational Research Methods*. https://doi.org/10.1177/1094428117722619

Kos, W., Schraagen, M., Brinkhuis, M., & Bex, F. (2017). *Classification in a Skewed Online Trade Fraud Complaint Corpus*. (B. Verheij & M. Wiering, Eds.), *Preproceedings of the 29th Benelux Conference on Artificial Intelligence: BNAIC 2017*. Groningen, The Netherlands.

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, *31*, 249–268. https://doi.org/10.1115/1.1559160

Koyuncugil, A. S., & Ozgulbas, N. (2012). Early warning system for financially distressed hospitals via data mining application. *Journal of Medical Systems*, *36*(4), 2271–2287. https://doi.org/10.1007/s10916-011-9694-1

Kraaij, W., & Pohlmann, R. (1994). Porter`s stemming algorithm for Dutch. *Informatiewetenschap 1994: Wetenschapelijke Bijdragen Aan de Derde STINFON Conferentie*, 167–180. https://doi.org/10.1.1.41.4271

Kraaij, W., & Pohlmann, R. (1995). Evaluation of a Dutch stemming algorithm. *The New Review of Document and Text …*, 1–21. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.41.5707&rep=rep1&type=pdf

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. https://doi.org/10.1007/978-1-4614-6849-3

Kumar, B. S., & Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, *114*, 128–147. https://doi.org/10.1016/j.knosys.2016.10.003

Kyoon Yoo, D., & Ah Park, J. (2007). Perceived service quality. *International Journal of Quality & Reliability Management*, *24*(9), 908–926. https://doi.org/10.1108/02656710710826180

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(December), 18–22. https://doi.org/10.1177/154405910408300516

Lin, F., Liang, D., Yeh, C.C., & Huang, J. C. (2014). Novel feature selection methods to financial distress prediction. *Expert Systems with Applications*, *41*(5), 2472–2483. https://doi.org/10.1016/j.eswa.2013.09.047

Lin, C. C., Chiu, A. A., Huang, S. Y., & Yen, D. C. (2015). Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments. *Knowledge-Based Systems*, *89*, 459–470. https://doi.org/10.1016/j.knosys.2015.08.011

López Iturriaga, F. J., & Sanz, I. P. (2015). Bankruptcy visualization and prediction using neural networks: A study of U.S. commercial banks. *Expert Systems with Applications*, *42*(6), 2857–2868. https://doi.org/10.1016/j.eswa.2014.11.025

Magnusson, C., Arppe, A., Eklund, T., Back, B., Vanharanta, H., & Visa, A. (2005). The language of quarterly reports as an indicator of change in the company's financial status. *Information and Management*, *42*(4), 561–574. https://doi.org/10.1016/j.im.2004.02.008

Mariscal, G., Marbán, Ó., & Fernández, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *Knowledge Engineering Review*, *25*(2), 137–166. https://doi.org/10.1017/S0269888910000032

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, *1697900*(Scipy), 51–56. Retrieved from http://conference.scipy.org/proceedings/scipy2010/mckinney.html

Meier, L., van de Geer, S., & Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society Series B*, *70*(1), 53–71. https://doi.org/10.1111/j.1467-9868.2007.00627.x

Morey, J., Scherzer, G., & Varshney, S. (2004). Predicting Financial Distress And Bankruptcy For Hospitals. *Journal Of Business & Economics Research*, *2*(9), 89–96.

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics*, *18*(6), 275–285. https://doi.org/10.1002/cem.873

NVZ. (2013). *Zorg Loont, Brancherapport algemene ziekenhuizen 2013*.

NVZ. (2017). *Brancherapport Ziekenhuiszorg 2017*.

Patel, F. N., & Soni, N. R. (2012). Text mining: A Brief survey. *International Journal of Advanced Computer Research*, *2*(6), 243–248. Retrieved from http://www.theaccents.org/ijacr/papers/conference/icett2012/43.pdf

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://doi.org/10.1007/s13398-014-0173-7.2

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137. https://doi.org/10.1108/eb046814

Porter, M. F. (2001). Snowball: A language for stemming algorithms. Retrieved January 11, 2018, from http://snowball.tartarus.org/texts/introduction.html

Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Measure To Roc, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63. https://doi.org/10.1.1.214.9232

Price, C. A., Cameron, A. E., & Price, D. L. (2005). Distress detectors - measures for predicting financial trouble in hospitals. *Healthcare Financial Management*, *59*(8), 74–80.

Qiu, X. Y., Srinivasan, P., & Hu, Y. (2014). Supervised Learning Models to Predict Firm Performance With Annual Reports : An Empirical Study, *65*(178), 400–413. https://doi.org/10.1002/asi

Qiu, X. Y., Srinivasan, P., & Street, N. (2006). Exploring the Forecasting Potential of Company Annual Reports. *Proceedings of the American Society for Information Science and Technology*, *43*(1), 1–15. https://doi.org/10.1002/meet.14504301168

Quinlan, J. R. (2014). *C4.5: Programs for machine learning*. Elsevier.

Rijcke, S. de, Wallenburg, I., Wouters, P., & Bal, R. (2016). *Comparing Comparisons: On Rankings and Accounting in Hospitals and Universities*. *Statewide Agricultural Land Use Baseline 2015* (Vol. 1). https://doi.org/10.1017/CBO9781107415324.004

Rutte, M., van Haersma Buma, S., Pechtold, A., & Segers, G. J. (2017). *Vertrouwen in de toekomst - Regeerakkoord 2017 – 2021*. Retrieved from https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/publicaties/2017/10/10/regeerakkoord-2017-vertrouwen-in-de-toekomst/regeerakkoord-2017-vertrouwen-in-de-toekomst.pdf

Sadasivam, G. S., & Lakshme, S. M. (2016). Corporate governance fraud detection from annual reports using big data analytics Mutyala Subrahmanyam and Dasaraju Himachalam Bhanu Prasad Pinnamaneni, *3*(1).

Schraagen, M., Brinkhuis, M., & Bex, F. (2017). Evaluation of named entity recognition in Dutch online criminal complaints. *Computational Linguistics in the Netherlands Journal*, *7*, 3–16.

Scott, S., & Matwin, S. (1999). Feature engineering for text classification. *Machine Learning-International Workshop*, *6*, 1–13. https://doi.org/10.1016/j.jbi.2012.04.010

Shirata, C. Y., Takeuchi, H., Ogino, S., & Watanabe, H. (2011). Extracting Key Phrases as Predictors of Corporate Bankruptcy: Empirical Analysis of Annual Reports by Text Mining. *Journal of Emerging Technologies in Accounting*, *8*(1), 31–44. https://doi.org/10.2308/jeta-10182

Tape, T. G. (2006). Interpreting diagnostic tests.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing and Management*, *50*(1), 104–112. https://doi.org/10.1016/j.ipm.2013.08.006

van den Haak, C. (2017). *Alles onder controle ? Opbrengsten stijgen , marges blijvend onder druk inzichten nieuwe*.

van der Wiele, T., Boselie, P., & Hesselink, M. (2002). Empirical evidence for the relationship between customer satisfaction and business performance. *Managing Service Quality: An International Journal*, *12*(3), 184–193. https://doi.org/10.1108/09604520210429259

Waarborgfonds, S. (2016). Jaarverslag.

Watkins, A. L. (2000). Hospital financial ratio classification patterns revisited: Upon considering nonfinancial information. *Journal of Accounting and Public Policy*, *19*(1), 73–95. https://doi.org/10.1016/S0278-4254(99)00025-3

Webb, Y. yang and G. (2003). On Why Discretization Works for Naive-Bayes Classifiers. *Lect Notes Comput Sci 2903*, 440–452. https://doi.org/10.1007/b94701

Wieringa, R. (2014). *Design Science Methodology for Information Systems and Software Engineering*. *Springer Berlin Heidelberg*. https://doi.org/10.1145/1810295.1810446

Willett, P. (2006). The Porter stemming algorithm: then and now. *Program*, *40*(3), 219–223. https://doi.org/10.1108/00330330610681295

Wirth, R. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, (24959), 29–39. https://doi.org/10.1.1.198.5133

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., & Wesslén, A. (2012). *Experimentation in software engineering*. *Experimentation in Software Engineering* (Vol. 9783642290). https://doi.org/10.1007/978-3-642-29044-2

Wu, H. C., Luk, R. W. P., Wong, K. F., & Kwok, K. L. (2008). Interpreting TF-IDF term weights as making relevance decisions. *ACM Transactions on Information Systems*, *26*(3), 1–37. https://doi.org/10.1145/1361684.1361686

Wu, S., & Flach, P. (2005). A scored AUC Metric for Classifier Evaluation and Selection. In *Second Workshop on ROC Analysis in ML*. Bonn, Germany.

Wu, W. W. (2010). Beyond business failure prediction. *Expert Systems with Applications*, *37*(3), 2371–2376. https://doi.org/10.1016/j.eswa.2009.07.056

Zeller, T. L., Stanko, B. B., & Cleverley, W. O. (1996). A revised classification pattern of hospital financial ratios. *Journal of Accounting and Public Policy*, *15*(2), 161–182. https://doi.org/10.1016/0278-4254(96)00014-2

Zhou, Y., Tong, Y., Gu, R., & Gall, H. (2016). Combining text mining and data mining for bug report classification. *Journal of Software: Evolution and Process*, *28*(3), 150–176. https://doi.org/10.1002/smr.1770

Zhu, X. (2007). Semi-Supervised Learning Literature Survey. *Sciences-New York*, 1–59. https://doi.org/10.1.1.146.2352

Appendix A: other tables

Table 14. List of Dutch stop words from NLTK

List of Dutch stop words				
de	er	mij	omdat	hier
en	maar	uit	iets	wie
van	om	der	worden	werd
ik	hem	daar	toch	altijd
te	dan	haar	al	doch
dat	zou	naar	waren	wordt
die	of	heb	veel	wezen
in	wat	hoe	meer	kunnen
een	mijn	heeft	doen	ons
hij	men	hebben	toen	zelf
het	dit	deze	moet	tegen
niet	zo	u	ben	na
zijn	door	want	zonder	reeds
is	over	nog	kan	wil
was	ze	zal	hun	kon
op	zich	me	dus	niets
aan	bij	zij	alles	uw
met	ook	nu	onder	iemand
als	tot	ge	ja	geweest
voor	je	geen	eens	andere
had				

Table 15. Financial indicators used by (Geng et al., 2015)

Types	Symbols	Formulae for calculation
Solvency	TL / TA	Total liabilities / total assets
	CA / CL	Current assets / current liabilities
	(CA – I) / CL	(Current assets – inventory) / current liabilities
	TL / TSE	Total liabilities / total shareholders' equity
	CL / TA	Current liabilities / total assets
	NOCF / CL	Net operating cash flow / current liabilities
	EBIT / IE	Earnings before interest and tax / interest expense
Profitability	(SR – SC) / SR	(Sales revenue – sales cost) / sales revenue
	NP / SR	Net profit / sales revenue
	EBIT / ATA	Earnings before income tax / average total assets
	NP / ATA	Net profit / average total assets
	NP / ACA	Net profit / average current assets
	NP / AFA	Net profit / average fixed assets
	NP / ASE	Net profit / average shareholders' equity
Operational capabilities	MBI / ATA	Main business income / average total assets
	SR/ ACA	Sales revenue / average current assets
	SR / AFA	Sales revenue / average fixed assets
	MBC / AI	Main business costs / average inventory
	MBI / ABAR	Main business income / average balance of accounts receivable
	CS / APA	Cost of sales / average payable accounts
Business development capacity	MBI(t) / MBI(t-1)	Main business income of this year / main business income of last year
	TA(t) / TA(t-1)	Total assets of this year / total assets of last year
	NP(t) / NP(t-1)	Net profit of this year / net profit of last year

Types	Symbols	Formulae for calculation
Structural soundness	CA / TA	Current assets / total assets
	FA / TA	Fixed assets / total assets
	SE / FA	Shareholders' equity / fixed assets
	CL / TL	Current liabilities / total liabilities
Capital expansion capacity	NP / NOS	Net profit / number of ordinary shares at the end of year
	NA / NOS	Net assets / number of ordinary shares at the end of year
	NICCE / NOS	Net increase in cash and cash equivalents / number or ordinary shares at the end of year
	CR / NOS	Capital reserves / number of ordinary shares at the end of year

Table 16. Financial indicators used by CBS

Kengetal	Berekening
Resultaat gewone bedrijfsvoering (EBT)	Resultaat voor belastingen / totale bedrijfsopbrengsten * 100%
Financiële baten en lasten	Financieel resultaat / totale bedrijfsopbrengsten * 100%
Winst voor interest en belastingen (EBIT)	Resultaat voor belasting – financieel resultaat / totale bedrijfsopbrengsten * 100%
Winst voor interest, belastingen en afschrijvingen (EBITDA)	(Resultaat voor belasting – financieel resultaat – afschrijvingen op vaste activa) / totale bedrijfsopbrengsten * 100%
Rendement op geïnvesteerd vermogen	Operationele marge * omloopsnelheid kapitaal
Operationele marge	100% - (Operationele kosten + afschrijvingen)
Operationele kosten	Kosten personeel in loondienst + kosten personeel niet in loondienst + overige kosten
Kosten personeel in loondienst	(Totaal arbeidskosten + overige personeelskosten) / totale bedrijfsopbrengsten * 100%
Kosten personeel niet in loondienst	Kosten uitzendkrachten en overige inleen / totale bedrijfsopbrengsten * 100%
Overige kosten	Totaal niet eerder genoemde bedrijfskosten / totale bedrijfsopbrengsten * 100%
Hotelmatige kosten	Voeding- en hotelmatige kosten / totale bedrijfsopbrengsten * 100%
Algemene kosten	Algemene kosten / totale bedrijfsopbrengsten * 100%
Client- en bewonergebonden kosten	Client- en bewonergebonden kosten / totale bedrijfsopbrengsten * 100%
Onderhoud- en energiekosten	Onderhoud- en energiekosten / totale bedrijfsopbrengsten * 100%
Niet eerder genoemde kosten	(Huur en operationele leasing kapitaalgoederen + andere bedrijfskosten) / totale bedrijfsopbrengsten * 100%
Afschrijving	Afschrijvingen op vaste activa / totale bedrijfsopbrengsten * 100%
Omloopsnelheid kapitaal	Totale bedrijfsopbrengsten / totaal activa

Kengetal	Berekening
Vaste activa	(Immateriële vaste activa + materiele vaste activa + financiële vaste activa) / totale bedrijfsopbrengsten * 100%
Vlottende activa	(Voorraden + onderhanden werk uhv DBCs + kortlopende vorderingen + financieringstekort + effecten + liquide middelen) / totale bedrijfsopbrengsten * 100%
Weerstandsvermogen	Eigen vermogen / totale bedrijfsopbrengsten * 100%
Solvabiliteit	Eigen vermogen / totaal activa * 100%
Quickratio	(kortlopende vorderingen + effecten + liquide middelen) / kortlopende schulden * 100%
Current ratio	(voorraden + onderhanden werk uhv DBCs + kortlopende vorderingen + effecten + liquide middelen) / kortlopende schulden * 100%
Rentabiliteit	Resultaat voor belastingen / eigen vermogen * 100%

Table 17. Financial ratios used by (Lin, F. et al., 2014)

Category	Ratio	Category	Ratio
Liquidity	Current ratio	**Operational**	Payable turnover ratio
	Quick ratio		Total assets turnover
	Quick assets / total assets		Receivable turnover ratio
	Current assets / total assets		Fixed assets turnover
	Working capital / total assets	**Profitability**	Operating income after tax / equity
	Working capital / sales		Operating income after tax per share
	No-credit interval		1 if net income was negative for the last two years, otherwise, 0
Solvency	Interest expenses / equity		Pre-tax income per share
	Market value equity / book value of total debt		Retained earnings/total assets
	Cost of interest – bearing debt		Operating income before tax/total assets
	Interest expense / revenue		Operating income after tax/total assets
Growth	Total equity growth		Operation income per employee
	Total assets growth		Gross profit/net sales
	Ordinary income growth		Realized gross profit/net sales
	Return on total asset growth		Sales per employee
	Net income growth		Net income/total assets
	Sales growth		Net income/equity
Cash-flow	Cash flow / total assets	**Capital structure**	Equity/total assets
	Cash flow / total liabilities		Fixed assets per employee
	Cash flow / equity		Liabilities/total assets
	Cash re-investment ratio		One if total liabilities exceeds total assets, zero otherwise
	Funds provided by operations / total liabilities	**Other**	Size

Table 18. Overview of all balance sheet variables

Immateriële vaste activa
Materiële vaste activa
Financiële vaste activa
Financiële vaste activa
Voorraden
Onderhanden werk uit hoofde van DBC's / DBC-zorgproducten
Effecten
Liquide middelen

Totaal vlottende activa
Totaal activa (vaste activa + vlottende activa)
Kapitaal
Totaal eigen vermogen
Langlopende schulden (nog voor meer dan een jaar)
Totaal overige passiva
Totaal passiva (eigen vermogen + overige passiva)

Table 19. Overview of all income statement variables

Subsidies (exclusief Wmo en jeugdwet)
Overige bedrijfsopbrengsten
Som der bedrijfsopbrengsten
Personeelskosten
Afschrijvingen op immateriële en materiële vaste activa
Bijzondere waardeverminderingen van vaste activa
Overige bedrijfskosten
Som der bedrijfslasten
Bedrijfsresultaat
Subtotaal financiële baten
Subtotaal financiële lasten
Totaal financiële baten en lasten
Resultaat uit gewone bedrijfsvoering
Belastingen resultaat uit gewone bedrijfsvoering
Resultaat uit gewone bedrijfsvoering na belastingen
Resultaat uit gewone bedrijfsvoering na belastingen
Buitengewone lasten
Buitengewoon resultaat
Resultaat

Appendix B: code and other documents

The code and other documents are saved in external files, added to the zipped folder in which this thesis is submitted.