Corruption Risk Detection in Dutch Healthcare: A Red-Flag Approach

I.A.C. Peeters

Supervisors:
Dipl. Ing. Jakub Sieber, PhD
Dr. Ing. Niels Pulles

January 2025

UNIVERSITY OF TWENTE.

Contents

LIST OF FIGURES AND TABLES 3

ABSTRACT 4

1) INTRODUCTION: THE NEED FOR PREDICTING HEALTHCARE CORRUPTION RISK THROUGH RED FLAGS 5

1.1 HEALTHCARE ACCOUNTS FOR THE BIGGEST SHARE OF THE DUTCH GDP 5
1.2 DEFINING CORRUPTION IN THE HEALTHCARE SECTOR AND ESTABLISHING THE CONTEXT OF DUTCH HEALTHCARE 6
1.3 BY PREDICTING POSSIBLE CORRUPTION, TENDERS CAN BE MORE EASILY IDENTIFIED, AND CORRUPTION CAN BE MITIGATED FASTER 6
1.4 DEFINING THE RESEARCH QUESTION 8

2) LITERATURE OVERVIEW: RED FLAGS DISCOVERED AND DETECTED USING MACHINE LEARNING METHODS 10

2.1 VISUALLY OUTLINING THE LITERATURE OVERVIEW 10
2.2 CORRUPTION IS FEASIBLE IN THE TENDERING PROCESS BUT USING RED FLAG INDICES, TENDERS CAN JUSTIFIABLY BE INVESTIGATED FOR CORRUPTION 10
2.3 FORMS OF CORRUPTION SPECIFIC TO THE HEALTHCARE DELIVERY PROCESS BETWEEN VARIOUS FUNCTIONS 13
2.4 DESCRIBING REAL-LIFE CORRUPTION CASES IN EUROPE AND MEASURES TO ADDRESS PRESENT AND FUTURE CORRUPTION 14
2.5 GOVERNMENTAL ENTITIES AND SCHOLARS CREATE VARIOUS TYPES OF RED FLAGS AND INDICES TO CONTROL CORRUPTION 15
2.6 THE RANDOM FOREST ALGORITHM IS THE BEST-PERFORMING MACHINE LEARNING METHOD FOR DETECTING COLLUSION 17

3) METHODOLOGY: HOW TO APPROACH THE RESEARCH 20

3.1 CRISP-DM FRAMEWORK IS MOST USED FOR DATA MINING 20
3.2 DATA SAMPLE DESCRIPTION 21
3.3 TWELVE RED FLAGS ARE BEING USED FOR THE MODEL 21
3.4 USING THE RANDOM FOREST ALGORITHM TO SOLVE A CLASSIFICATION PROBLEM 24
3.5 RANDOM FOREST ALGORITHM 25
3.6 MICE IS USED TO IMPUTE LARGE AMOUNTS OF MISSING DATA 27

4) RESULTS 29

4.1 MISSING VALUES 29
4.2 DESCRIPTIVE STATISTICS 30
4.2.1 Original 30
4.2.2 Imputation 30
4.3 COMPARISONS BETWEEN RANDOM FOREST MODELS 33
4.3.1 Confusion Matrix 34
4.3.2 Performance metrics 35
4.4 COMPARISONS BETWEEN SHAP BEESWARM PLOTS 36
4.4.1 Model 2 36
4.4.2 Model 4 39
4.4.3 Summary of findings from the SHAP beeswarm plots 40

5) DISCUSSION 42

5.1 KEY FINDINGS 42
5.2 OVERFITTING 43
5.3 THEORETICAL IMPLICATIONS 43
5.4 PRACTICAL IMPLICATIONS 44
5.5 CONCLUSION 44
5.6 LIMITATIONS AND FUTURE RESEARCH 45

REFERENCES 48

6) APPENDICES 53

6.1 APPENDIX I – RED FLAGS 53
6.2 APPENDIX II – FLAGS USED IN MODEL 59
6.3 APPENDIX III – R-CODE RANDOM FOREST 67

List of Figures and Tables

Figure 1 - Conceptual Framework 9
Figure 2 - Outline Literature Overview 10
Figure 3 - OECD's procurement process (Schulz & Kourkoulas, 2023) 11
Figure 4 - General Tendering Process (Mohemad et al., 2011) 12
Figure 5 - Corrupt practices through actors' relationships in the health system (Vian, 2008) 13
Figure 6 - CRISP-DM phases (Schröer et al., 2021) 20
Figure 7 - Trade-off interpretability and flexibility on different statistical methods (James et al., 2021) 25
Figure 8 - Random Forest algorithm steps(Biau & Scornet, 2016) 26
Figure 9 - Formula Random Forest (Github, n.d.) 27
Figure 10 – Process 27
Figure 11 - Overview models 33
Figure 12 - SHAP beeswarm plot Model 2 37
Figure 13 - SHAP beeswarm plot Model 4 39

Table 1 - Machine learning measures (Sisodia & Sisodia, 2018) 18
Table 2 - Information columns 21
Table 3 - Implemented red flags TenderNed 23
Table 4 - Missing values per variable 29
Table 5 - Frequency awarded tenders by contracting authority 30
Table 6 - Frequency of awards to contractors 30
Table 7 - Descriptive statistics original data 30
Table 8 - Frequency table RF1 31
Table 9 - Descriptive statistics imputed data 32
Table 10 - Frequency table RF3 33
Table 11 - Random Forest output Model 2 35
Table 12 - Random Forest output Model 4 35
Table 13 - Model 2 metrics 36
Table 14 - Model 4 metrics 36

Abstract

Public procurement is the process by which public authorities purchase goods and/ or services from the private sector. Procurement accounts for 42% of the Dutch GDP and healthcare has the biggest share. Unfortunately, around 0.76% or €4.4 billion of the Dutch GDP is lost to corruption and has serious consequences. The World Bank, European Union and anti-fraud organisations have developed indices to detect corrupt practices using red flags. Red flags are indicators of corruption but not a guarantee. The research paper aimed to predict the level of possible corruption in Dutch healthcare tenders through machine learning with past-defined red flags. The red flags were extracted from the indices and aligned with the available data from the Dutch government. A Random Forest algorithm was chosen based on being the best-performing and most effective machine learning method for accurately detecting corruption. The algorithm made predictions of whether a tender would be considered corrupt based on levels. Four different types of predictive models were created, but two were analysed in detail because they assumed corruption would be a scale variable rather than a dichotomous one. The recall, precision, accuracy, and F1-score measures were calculated to assess and evaluate the models. It is possible to predict the level of potential corruption of tenders in Dutch healthcare through machine learning with well-defined levels. The predictions are more accurate when predictors are coded numerically rather than dichotomously. Furthermore, a SHAP beeswarm plot was used to examine the influence of the predictors. The advertisement period and contract execution period were the most influential. If they are present the level of possible corruption would increase by 0.5 points.

1) Introduction: The need for predicting healthcare corruption risk through red flags

1.1 Healthcare accounts for the biggest share of the Dutch GDP

Procurement can be divided into public and private procurement. Public procurement is defined as "*the process by which public authorities purchase goods and/ or services from the private sector"* (Dotoli et al., 2020). Private procurement is linked to organisations and is freer in how they operate compared to governments. The difference between public-private partnerships and traditional public procurement is the efficiency of private providers and the lack of flexibility from private providers in fast-changing sectors like healthcare and technology (Ross & Yan, 2015). The public authorities are responsible for the daily operations of sectors such as the military, health, education, infrastructure, public transport, water management, law enforcement and other connections to all public services and goods in which management is a governed body. In The Netherlands, public procurement accounts for 42% of GDP in 2021, mostly going towards health (35%) and social protection (24%), which is substantially more than in other European countries (OECD, 2023). One could contend that public procurement influences society and its citizens greatly by boosting jobs, investing in future generations, and being resource- and energy-efficient (European Commission, n.d.). Therefore, European countries are lawfully obliged to carry their procurement processes out with integrity, transparency, oversight and control (OECD, 2016).

A recent shift or trend is the use of e-procurement by governments and firms due to technological developments. E-procurement systems are designed to facilitate the procurement process between two parties through digital platforms. Malaysia implemented an e-procurement system in the governmental sector (Aman & Kasimin, 2011), the European Union made e-procurement a key component in their new EU Directives to simplify procedures of public purchases and increase flexibility (Bobowski & Gola, 2018), and the e-procurement technology was implemented in the South African construction industry (Ibem & Laryea, 2015). As demonstrated, developing and developed countries are implementing e-procurement systems. However, countries with agile and efficient organisational structures are more adjustable towards e-procurement tools such as electronic forms (Seri et al., 2014). The Netherlands implemented an e-procurement system in 2005 before the European Union made it mandatory, representing 8% of the total procurement (Mélon & Spruk, 2020). 8% of all procurement-related purchases are made through an e-procurement system.

1.2 Defining corruption in the healthcare sector and establishing the context of Dutch healthcare

Corruption is defined as: *"the misappropriation of authority, resources, trust or power for private or institutional gain"* and around 0.76% or €4.4 billion of the Dutch GDP is lost to corruption by embezzlement, bribery, fraud, favouritism, nepotism, and misappropriation (Bartlett Quintanilla et al., 2018; Mackey & Cuomo, 2020). Intergovernmental organisations, non-profit organisations and even scholars are developing measures to detect corruption by setting up investigative offices, corruption indices utilising red flags, and tools to validate analyses. Red flag indicators of corruption are defined as: "*an accumulation of traces that may point to the presence of corrupt activities aimed at helping practitioners, investigators, and policymakers in estimating the probability of corruption of a certain procurement case and lay the foundation of a new evidence-based approach to fighting corruption"* (Decarolis & Giorgiantonio, 2022). A better understanding of the red flags can lead to proactive measures that ensure a more transparent and accountable procurement process.

As previously mentioned, the Dutch government spends most of its public funding towards the healthcare sector. The healthcare sector is comprised of many different sub-sectors such as health insurance, mental health, hospitals, childcare, general practitioners, dentists, elderly care, etc. Healthcare encompasses every institution and worker who is providing care or is a supplier of healthcare equipment. The Netherlands scores high on the Digital Non-Corrupt Health Index (Transparency International, 2023a), meaning they have robust and resilient healthcare systems and lower levels of perceived corruption compared to other countries. However, such a statement doesn't mean they are still vulnerable to corrupt practices. Furthermore, the score of The Netherlands has slowly been declining since 2015. The decline could be due to inadequate political integrity measures and a lack of sufficient lobbying oversight (Transparency International, 2023b).

1.3 By predicting possible corruption, tenders can be more easily identified, and corruption can be mitigated faster

The prevalence of corruption globally hinders economic growth, undermines democratic principles, compromises social justice, dwindles public trust, interferes with political decision-making, and jeopardises private investment intentions (Lima & Delen, 2020).

Possible corruption is detected using red flags. Red flags indicate possible corruption and are based on characteristics of previous corrupt cases. For example, during the COVID-19 pandemic, there were procurement cases which were identified as corrupt due to fewer bidders, shorter advertisement periods, and less transparent procedures (Thomann et al., 2024). These three characteristics are now used as red flags to detect future (procurement) cases for possible corruption.

The World Bank, the European Union, governments, the Organisation for Economic Co-operation and Development (OECD), the European Anti-fraud Office (OLAF), and Transparency International have made indices of red flags to detect and prevent possible corruption. Furthermore, previous academic papers have utilised these indices to report on red flags in public procurement, how to detect them and their accuracy, such as through the Corruption Risk Index (CRI) (Decarolis & Giorgiantonio, 2022; Fazekas & Kocsis, 2020; Fazekas et al., 2016; Ferwerda et al., 2017; Tátrai & Németh, 2018). The indices are further discussed in Section 2.5. However, (Fazekas et al., 2016) argue that indices to measure corruption available from the OECD, OLAF, and The World Bank are possibly outdated due to the divide between perception and reality because of media coverage of high-profile corruption cases, citizens and experts have no direct experience of corruption, representativeness bias by non-representative surveys, reflexivity bias by influenced respondents, and non-response or false response to sensitive questions.

So, red flags are indicators of corruption, but it is not guaranteed. An investigation may be necessary to investigate possible corruption and prevent (major) scandals when one or several red flags are triggered. However, reviewing each case by hand to assess whether red flags were triggered is inefficient. Machine learning methods are a good tool for detecting corruption more effectively. They can make decisions and predict outcomes based on the red flags. Machine learning methods can be divided into supervised (regression and classification) and unsupervised learning (clustering and association), and are useful for distinguishing classification patterns, visualisation, clustering, exploration, predicting or estimating outcomes, and detecting, investigating, or testing certain variables (Badillo et al., 2020). Instances of using machine learning methods are predicting political connections with supervised machine learning techniques to combat conflicts of interest (Titl et al., 2024), predicting jail-or-release decisions by training an algorithm to form a prediction function (Kleinberg et al., 2017), predicting mortality in older adults through lasso and Random Forest analysis (Puterman et al., 2020), detecting new red flags through automatic extraction of evidence of fraud (Lima et al., 2023). The most used machine learning techniques for predictive modelling techniques are logistic regression models, Naive Bayes, k-nearest neighbours, decision trees, neural networks, vectors- machine and Random Forests (Mackenzie, 2015).

1.4 Defining the research question

This thesis will create a predictive model through machine learning by examining red flags for compatibility with tender cases between 2016 and 2023 in the Dutch healthcare sector to measure the risk of possible corruption. The data is derived from TenderNed, an online platform for tenders, and includes tenders of academic medical centres, healthcare institutes, ministries or other national or federal authorities with regional or local subdivision, public law institutions, and national, regional or local authorities, agencies or offices. However, no hospitals are part of this dataset.

To ensure the predictive model is coded correctly and the model is accurately predicting the outcomes, four different models will be created. The difference in versions is data manipulation.

Version 1 contains dichotomous variables of the red flags, where corruption is dichotomous (0-1). Version 2 contains dichotomous variables of red flags, where corruption is at levels 1 through 6. Version 3 contains numerical or categorical variables of red flags, where corruption is dichotomous (0-1). Version 4 contains numerical or categorical variables of red flags, where corruption is at levels 1 through 6.

This approach has not yet been applied to the healthcare sector in the Netherlands. This study focuses on the healthcare sector because it is the largest sector in which the Dutch government allocates funds. Trends like the rising elderly population, a pandemic, expensive technology and equipment, and complex diseases put more pressure on the healthcare sector and government, which makes it more susceptible to corruption (Bîzoi & Bîzoi, 2023).

The conceptual framework displays the independent variable as the composite of red flags and the dependent variable as the risk of possible corruption, see Figure 1. As mentioned previously, the literature reports on detecting red flags and developing measurements of corruption. A similar framework was used in (Fazekas & Kocsis, 2020), where they test the validity of a self-constructed corruption risk index and single bidding as proxy indicators for measuring high-level corruption and the framework of (Fazekas & Kocsis, 2020) posits looking into regions, sectors, organisations or individuals’ behaviour to advance the field.

The following research question will therefore be answered.

RQ: Can past-defined red flags predict the level of possible corruption in Dutch healthcare tenders through machine learning?

To answer the main research question, the main research question is split into sub-questions, which are as follows:

1. How accurate are machine learning models while making the prediction?
2. How should the machine learning predictive model operate?

Figure 1 - Conceptual Framework

The remainder of this paper is structured as follows: section two summarises the literature overview about existing red flags and a machine learning predictive model. Section three announces the methodology. Section four discusses the results and answers the research questions. Section five concludes the research, discusses the outcomes, shows the research limitations, and identifies research gaps for future research.

2) Literature Overview: Red flags discovered and detected using machine learning methods

2.1 Visually outlining the literature overview

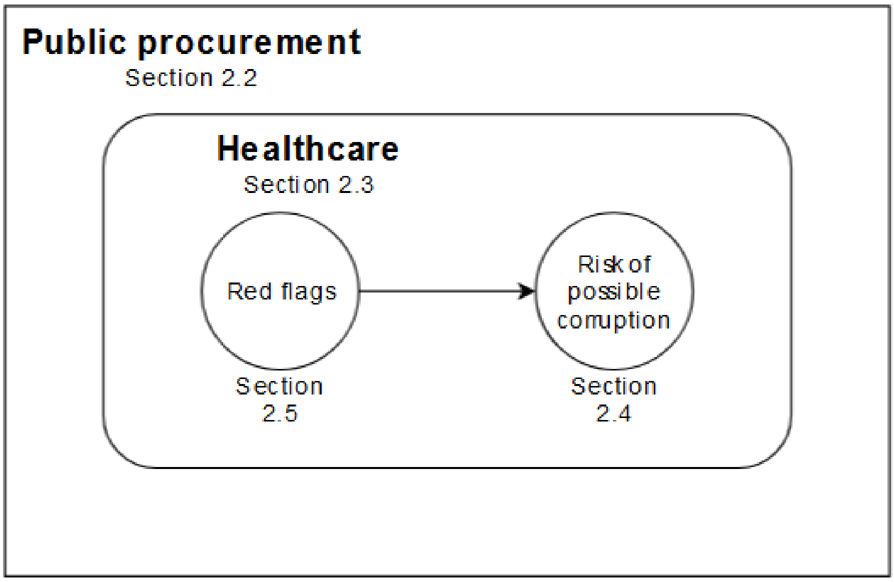

Figure 2 - Outline Literature Overview

As mentioned previously, public procurement can occur in governments and businesses. In this case, the research is focused on Dutch healthcare making up 42% of the GDP. The thesis focuses on the relationship between red flags and corruption through a machine learning model within the healthcare sector, see Figure 2.

This section first explains the tendering process, while keeping the general concept of procurement in mind. Second, the healthcare sector in The Netherlands and healthcare corruption will be discussed. Thirdly, examples of healthcare corruption will be presented. Next, the existing red flags and indices will be discussed. Lastly, the machine learning variable will be explained.

2.2 Corruption is feasible in the tendering process but using red flag indices, tenders can justifiably be investigated for corruption

In previous studies, the procurement process was usually split into a few phases or steps. However, these phases were all different from each other. So, the most relevant process is the process from the European Union because the setting of the research question is in The Netherlands, which is part of the European Union (see Figure 3).

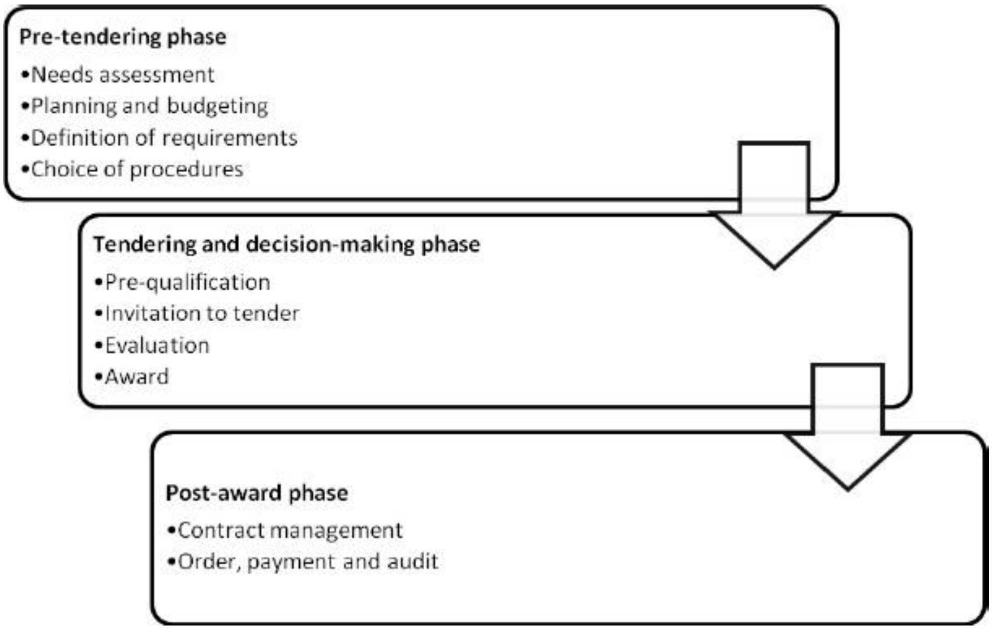

Figure 3 - OECD's procurement process (Schulz & Kourkoulas, 2023)

A tender process occurs when a company or government (entity) announces its need for goods or services. Companies interested in fulfilling the project must put a tender document in an online portal or to a mediator, who communicates between the buyer and supplier. The type of availability for the providers of the tenders can differ from all providers (open tender), a few selected providers (restricted tender) to negotiating by the mediator with the provider (negotiated tender) (Mohemad et al., 2011). Suppliers must adhere to several factors or weights, decided by the owner. First, (pre)qualifications are therefore made by the contracting authority. After the collection period, the company can see all the quotations and choose who to do business with based on evaluation criteria. A scoring rule auction mechanism, where the bidders are evaluated by a linear function on quality and price, is usually used to select the bidder with the highest score (Camboni & Valbonesi, 2021). The contracting authority will award the contract to the most suitable provider. Figure 4 shows the tendering and decision-making phase of the procurement process in more detail through a flowchart separated by the different actors.

However, the contracting authority, in our case government and municipalities are vulnerable to several forms of corruption during public procurement. Corruption involves people who act on opportunities and their incentives of perceived value, low risk of sanction, and low degree of professional integrity (Heggstad & Froystad, 2011). Corruption can be caused structurally (political, historical, and cultural) or individualistically (individuals, companies, groups).

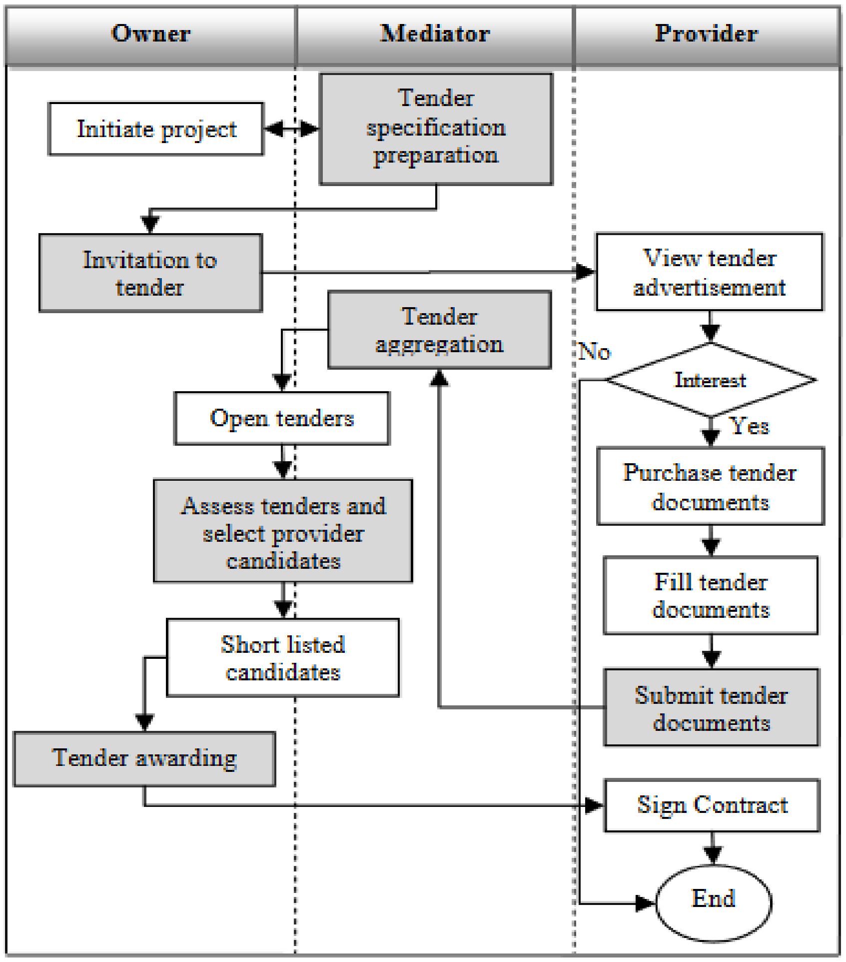

Figure 4 - General Tendering Process (Mohemad et al., 2011)

Bribery is one of the most common types of corruption characterised by offering money to someone in exchange for something valuable with a low risk of being caught, for example, information, preferential treatment or influence (Heggstad & Froystad, 2011). Bid rigging by public officials or contractors is another common type of corruption. Bid rigging is manipulating the outcome of a bid by offering something valuable in exchange. It can be done in a multitude of ways such as: excluding qualified bidders, rigged specifications, unbalanced bidding, unjustified direct awards, manipulation of bids, undeclared conflict of interest, complementary bidding, bid suppression, and market division (Heggstad & Froystad, 2011). Other forms of corruption are taking someone's property allocated to someone else (embezzlement); misuse of power (extortion); informal networks; favouritism towards family and friends (nepotism); political support in exchange for favours (patronage); and conflict of interest (Heggstad & Froystad, 2011). By using the indices and the associated red flags, the owner can recognise possible attempts at corruption during the process or afterwards. Similarly, investigators of corruption cases can use red flag indices as support to investigate possible corrupt tenders.

2.3 Forms of corruption specific to the healthcare delivery process between various functions

Corruption takes on different types of forms in the healthcare sector. Section 1.2 mentioned bribery, misuse of level positions, embezzlement of medicines and medical devices, and improper marketing relations as different types of corrupt practices. When examining corrupt practices within the healthcare delivery process and the interactions between the various actors and functions, (Vian, 2008) found bribery; collusion; under-the-table payments for care; bid rigging; absenteeism; unethical drug promotion; inappropriate ordering of tests and procedures to increase financial gain; lack of accountability; biased application; nepotism; unnecessary referrals; political influence; and use of government resources for private practice as corrupt practices (see Figure 5). The corrupt practices can be investigated based on red flags, and risks causing corruption to prevent a next corruption scandal. The next section presents real-life cases of healthcare corruption in Europe.

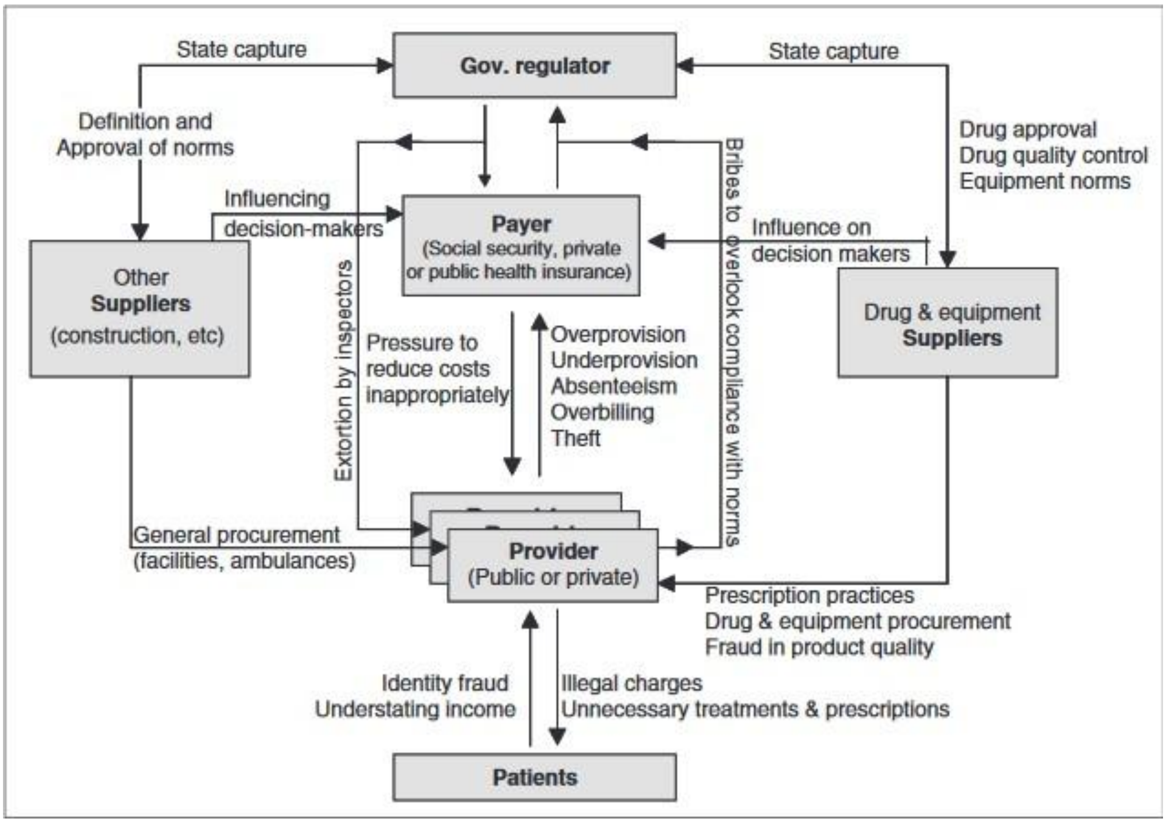

Figure 5 - Corrupt practices through actors' relationships in the health system (Vian, 2008)

2.4 Describing real-life corruption cases in Europe and measures to address present and future corruption

During the COVID-19 pandemic, countries needed medical equipment in large quantities and within a short time frame, resulting in more relaxed regulations. The infamous case of facemasks bought from Sywert van Lienden comes to mind in The Netherlands. Due to the lack of facemasks, the government bought 40 million from Van Lienden. The Ministry of Health was aware Van Lienden would make a profit, but they were misled that he would make a big profit of 9 million euros (Klaassen, 2022). Unfortunately, The Netherlands was not the only country victim of corruption during the COVID-19 pandemic. In Germany and Italy, corruption was more prevalent due to fewer bidders, shorter advertisement periods, and less transparent procedures (Thomann et al., 2024). In Romania, a criminal network that manipulated public procurement to benefit predetermined suppliers during the pandemic was investigated resulting in damages of over 8 million euros in the EU budget (European Public Prosecutor's Office, 2024b).

Even though a higher risk of possible corruption during COVID-19, cases such as the mental institution in Emmen in 2016 and corrupt cardiologists in Zwolle also occurred in The Netherlands. The mental institution in Emmen was accused of declaring higher amounts of care than was provided, invoicing temporarily imprisoned clients who no longer received care, and laundering funds totalling €100,000 (NOS, 2016). A more recent example is the case of the Isala Hospital in Zwolle. Five employees of one of the biggest hospitals in The Netherlands were accused of accepting bribes after two cardiologists were fired the week before due to bribing from a supplier of medical products in exchange for millions of euros to receive preferential treatment from the cardiologists for years (NOS, 2022).

Crossing the border, other European countries are also vulnerable to corruption in the healthcare sector. The Austrian office of Siemens cooperated with Austrian authorities to investigate accusations of inflated invoices for the delivery of building technology totalling less than 10 million euros, resulting in five people getting arrested (Schwarz-Goerlich et al., 2023). In England, a woman was accused of faking nursing qualifications and experience in the NHS (Hume & Ferda, 2024). In Malta, the former prime minister and several former and current officials were accused of money laundering and selling public hospitals to a private company by awarding a 30-year, multi-billion-euro contract to a firm without health sector experience (Chadwick, 2024). In the Czech Republic, employees of two hospitals were accused of bribery, using forged invoices for fictitious services, and participating in a suspected criminal organisation that manipulated the public procurement of medical supplies by tailoring contracts to preferred suppliers and transmitting confidential information to them resulting in damages in the EU budget of close to 1 million euros (European Public Prosecutor's Office, 2024a).

Another sub-sector of healthcare has been studied of pharmaceutical corruption in medicine. In particular, the technological methods of detecting and combating corruption to ensure a more transparent process through a literature review on electronic data interchange, e-procurement systems, and blockchain in the pharmaceutical supply chain and procurement field. Mackey and Cuomo (2020) found e-procurement driven by cost-saving negatively affects combatting corruption, but when focussed on transparency it has the opposite effect.

2.5 Governmental entities and scholars create various types of red flags and indices to control corruption

As said in Section 1.3, red flags indicate possible corruption and are based on characteristics of previous corrupt cases. Appendix I lists all the red flags mentioned in this section. All the red flags will be focused on and examined for compatibility with the dataset. After the examination, there will be a final number of red flags which is going to be used further in the process.

The European Commission has set up an Anti-Fraud Office (OLAF) to investigate corruption involving the EU funds, staff, and institutions, and recommend stopping fraud at a national level (OLAF, 2022). OLAF identified 27 red flags based on 192 cases through interviews, case studies, and testing. The tool created by OLAF is based on the procedural phase and, therefore, receives criticism for its weak ability to tailor towards individual characteristics and failure to distinguish corrupt public procurement from the rest (Fazekas et al., 2016; Tátrai & Németh, 2018).

The World Bank developed multiple indices from transparency, accountability, and control of corruption to governance indicators. The first World Bank index is also known as the Control of Corruption Index (CCI.) (Seri et al., 2014). They identified 13 red flags across four types of red flags: unobservable, uncollectible, potentially irrelevant, observable, collectable and relevant (Tátrai & Németh, 2018). Again, the CCI has been criticised for the divide between perception and reality because the indicators have been identified from no direct experiences and are insensitive to change (Fazekas et al., 2016).

The European Commission isn't the only council setting up corruption indices. Contrasting to OLAF and the World Bank, the OECD has created a simplified tool focusing on identifying good procurement methods by dividing red flags into the different phases of the procurement process compliant with its own rules and procurement processes (Tátrai & Németh, 2018). They have highlighted six supportive principles to prevent corruption in public procurement: integrity, transparency, stakeholder participation, accessibility, e-procurement, and oversight and control (OECD, 2016).

The non-governmental organisation Transparency International developed an index named the Corruption Perception Index (CPI) and International Country Risk Guide (ICRG) based on grand and petty corruption to signal corrupt risks (Spyromitros & Panagiotidis, 2022). The CPI has measured corruption through in-depth interviews, focus groups, and national quotas, while the ICRG is a statistical model for the comparisons between countries based on 22 elements in three risk categories (Spyromitros & Panagiotidis, 2022).

Not only do public entities set up anti-fraud offices, but scholars are also analysing how to detect red flags and whether the current indices are successful.

Fazekas et al. (2016) developed a CRI from Hungarian public procurement data. The CRI measures institutionalised grand corruption based on the corrupt rent extraction process, red flags, and objective data. It is used to compare institutionalised corruption from individual organisations to countries. First, they split the public procurement process into three phases: submission, assessment, and delivery. The following red flags were incorporated into their CRI: single bidder contract, call for tender (not published in official journal), procedure type, relative length of eligibility criteria, short submission period, the relative price of documentation, call for tenders modification, exclusion of all but one bid, weight of non-price evaluation criteria, annulled procedure re-launched subsequently, length of decision period, contract modification, contract lengthening, and contract value increase (Fazekas et al., 2016). The red flags are weighted between 0 and 1 to scale to the likeliness of corruption.

Abdou et al. (2022) further develop the CRI by questioning whether it explains higher prices paid for procured goods and/ or services based on single bidder contracts, non-open procedures, lack of publication of a call for tenders, a period for submitting bids, lack of publication of a call for tenders, a period for submitting bids, a period for selecting the winning bid, spending concentration, share of suppliers registered in jurisdictions offering limited company and banking transparency. The red flags were tested on the average score of corruption risk and budget implications with a dataset of public procurement contracts from Georgia, Indonesia, Paraguay, Romania, and Uganda through the CRI and regression model. According to (Abdou et al., 2022), price increases for procured goods and services result from red flags.

Decarolis and Giorgiantonio (2022) expands the list of red flags by incorporating elements of the call for tenders with the absence of tender call, call for tenders, Italian Anticorruption Authority (ANAC) information available, legality protocols, local regulations, design-build, open tender days, document verification, worksite verification, prohibition of pooling agreements, multiple contact points, and external contact points.

Ferwerda et al. (2017) identified from 192 public procurements cases by eight EU member states across five different sectors, eight indicators which truly related to predicting corruption, e.g.: contact office not subordinated to tender provider, the shortened period for the bidding process, tender substantial, complaints from non-winning bidders, significant changes in project scope/ costs after award, connections between bidders undermines competition, awarding authority did not fill in all fields, amount of missing information.

2.6 The Random Forest algorithm is the best-performing machine learning method for detecting collusion

There are currently many sources which produce different types of data. Nevertheless, the data itself doesn't give significant insights. Therefore, machine learning methods can extract information from data to have more useful knowledge in different fields based on the performance of the learning algorithm (Pugliese et al., 2021). Machine learning models are used for many issues such as prediction, clustering, estimating, and testing.

Fraudsters have been improving their methods through the use of technological advancements and therefore more intelligent and new approaches or technologies are being used to predict and prevent fraud, such as machine learning (Priya & Saradha, 2021). The use of machine learning models has advantages compared to historical approaches of algorithms written by fraud experts. Machine learning models make minimal assumptions about the data, they can learn by themselves, they are more efficient, fast, and accurate, and it is a continuous process (Bzdok et al., 2018; Priya & Saradha, 2021).

A machine learning model can be assessed and evaluated with the well-known parameters: accuracy, precision, recall, and F-measure (see Table 1) (Sokolova et al., 2006).

Measures	Definitions
Accuracy (A)	Accuracy determines the accuracy of the algorithm in predicting instances
Precision (P)	The classifier’s correctness/ accuracy is measured by Precision
Recall (R)	To measure the classifier's completeness or sensitivity, Recall is used
F-measure	F-measure is the weighted average of precision and recall

Table 1 - Machine learning measures (Sisodia & Sisodia, 2018)

The accuracy formula is (true positive + true negative)/ the number of samples. The value is a percentage between 0-100 per cent and it should be above 75% (James et al., 2021). The precision formula is true positive/ (true positive + false positive). The value is between 0-1 and should be around 1 (James et al., 2021). The recall formula is true positive/ (true positive + false negative). The value is a percentage between 0-100 per cent and it should be between 70 – 80 per cent (James et al., 2021). The F-measure formula is 2*(positive* recall)/ (positive + recall). The value is between 0-1 and should be around 0.7 (James et al., 2021).

The most used machine learning techniques for predictive modelling techniques are logistic regression models, Naive Bayes, k-nearest neighbours, decision trees, neural networks, vector- machine and Random Forest (Mackenzie, 2015). However, the Random Forest algorithm approach is one of the best-performing machine learning methods to detect collusion accurately. For instance, (García Rodríguez et al., 2022) assessed 11 different machine learning methods to determine that trees, Random Forests, and gradient boosting are the best methods to detect collusion accurately. Furthermore, (Lyra et al., 2022) found that the most effective and used approach is the Random Forest algorithm through a systematic literature review.

There are multiple cases where the Random Forest algorithm is used to predict fraud in numerous sectors, for risks, financial time series, traffic speed prediction etc. The following examples will cover fraud detection cases.

Firstly, the use of credit cards has been exponentially growing and, therefore, fraudsters have found another aspect to take advantage of by either forging false information (application fraud) or stealing an account and password from a genuine cardholder (behaviour fraud) (Xuan et al., 2018). Using the Random Forest algorithm, they found that fraud detection can be best detected through practical testing. The overall accuracy of the Random Forest algorithm was 99%.

Second, due to globalisation and technological advancements, the size of the firms has been growing and the fraudulent behaviour within those firms has also been increasing. Liu et al. (2015) found through the Random Forest algorithm that the debt-equity ratio is the most important variable, and the type II error probability is lower than the type I error. It can detect financial fraud significantly and improve efficiency. Not only did they use the Random Forest algorithm they also used four other models and still found that the Random Forest algorithm was the most accurate due to very high recognition efficiency, it ignores data normality assumptions. It measures the importance of each variable. The overall accuracy of the Random Forest algorithm was 88%, while the other four models' accuracy was between 42%-80%.

Lastly, fraudulent claims have been a challenge for car insurance companies. Using big data and data science, insurance companies can now predict claims. (Hanafy & Ming, 2021) analysed eight models and found that the Random Forest model is best suited to predict claim issues in insurance companies because it has better performance measures with an accuracy of 87%. While the other models had an accuracy between 60% - 80%. Moreover, Random Forest is the best in classifying and distinguishing classes with a sensitivity score of 97% and a specificity score of 71%.

3) Methodology: How to approach the research

3.1 CRISP-DM framework is most used for data mining

The main research question is: "Can past-defined red flags predict the level of possible corruption in Dutch healthcare tenders through machine learning?"

The independent variables are red flags. Red flags are indicators of corruption, but it is not a guarantee. It may point to the presence of corrupt activities and is based on characteristics of previous corrupt cases. The dependent variable is the risk of possible corruption. In Section 1.2, corruption was defined as "the misappropriation of authority, resources, trust or power for private or institutional gain" (Lyra et al., 2022). While this is a general definition of corruption, the definition of corruption related to healthcare covers specific practices and outcomes of corruption, such as bribery, improper marketing relations, and fraud and embezzlement of medicines and medical devices.

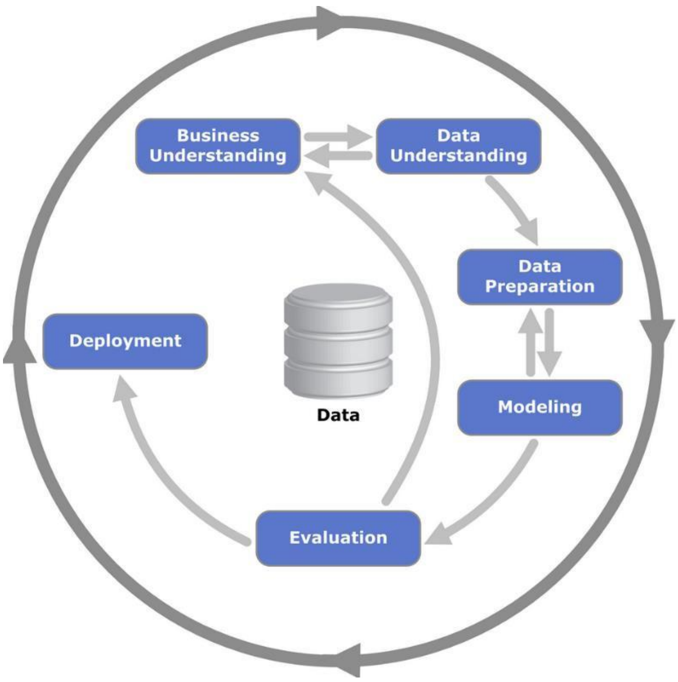

Figure 6 - CRISP-DM phases (Schröer et al., 2021)

The problem of predicting corruption by red flags through a Random Forest algorithm is categorised within Data Mining. The Cross-Industry Standard Process for Data Mining (CRISP-DM) framework is the most used methodology for data mining. The CRISP-DM model consists of six phases, see Figure 6. Data understanding encompasses data collection, exploration, and describing it (Schröer et al., 2021).

Data preparation covers cleaning the data by detecting and removing outliers, missing values, and transforming values. The modelling phase consists of selecting the modelling technique and building the test case and the model (Schröer et al., 2021). Model evaluation contains splitting the dataset into training and testing data, cross-validation, and bootstrap. Deployment is the final report or model.

3.2 Data sample description

The original data were extracted from the information files published by TenderNed. TenderNed contains national and international tenders from the Dutch government. The data on the second-level Classification Of the Functions Of Government (COFOG) from Centraal Bureau voor de Centraal Bureau voor de Statistiek (2023) shows that the main activity from TenderNed has the same value as COFOG. The dataset includes the COVID-19 pandemic, which is not a common occurrence and therefore rare compared to the population. Nevertheless, it represents the actions taken during the pandemic perhaps resulting in different tenders. It could lead to more insight into how red flags affect the healthcare sector when a country must react immediately to uncertainty.

The data from TenderNed contains 269,096 tenders between 2016 – medio 2024, of which 9,348 are related to the "Gezondheid" (Eng.: Health) at "Hoofdactiviteit" (Eng.: Main Activity). The 9,348 tenders are comprised as follows, see Table 2.

Split of Main Activity
5,006 tenders are "*Diensten"* (Eng.: Services)
4,158 tenders are "*Leveringen"* (Eng.: Supplies)
184 tenders are "*Werken"* (Eng.: Works)

Table 2 - Information columns

The Works shall be excluded from the data sample since this is such a small part of the tenders and only five tenders are suitable concerning the threshold of €5,538,000. Of the 9,164 tenders related to Diensten and Leveringen, 8,549 tenders were published before the 1st of January 2024. Lastly, the threshold for supplies and service contracts is €443,000, meaning tenders above the threshold must comply with European rules instead of national rules (European Union, n.d.). The tenders exceeding the threshold result in 1,248 tenders. So, 1,248 tenders will be used as the final sample. In the next paragraph, the red flags will be selected.

3.3 Twelve red flags are being used for the model

The dependent variable is the risk of possible corruption. The risk will be measured in levels because the approach accounts for the varying degrees of corruption likelihood based on the number of red flags present in a tender. A single red flag doesn't necessarily indicate corruption, so a scaled approach could provide better insights. The independent variable is red flags. Appendix I summarises all the red flags mentioned in Section 2.5. A limitation of selecting red flags from Appendix I is the lack of defining periods for certain red flags. For example, the red flag "Too short of an advertisement period" doesn't mention what is considered too short. Therefore, the Uniform Europees Aanbestedingswet (UAE) (Eng.: Uniform European Procurement Law) will be used to define those periods. In the UAE, the official legal proceedings are declared for the contracting authority, tender documents and contractors. In Appendix II, all the 186 red flags from Appendix I have been examined to determine whether they are compatible with the dataset. Furthermore, they are now accompanied by a definition (if available) and a label (green = compatibility, orange = aforementioned, red = no compatibility) to symbolise which red flags will be included in the analysed red flags and therefore in the model. Compatibility between the dataset and the red flag is important because there should be input from the dataset and tender to determine whether a red flag is triggered. The following red flags will be included in the research:

N	Red Flag	TenderNed dataset information
1	Procedure type	Procedure
2	Too short of an advertisement period	Publicatiedatum (Eng.: publication date) Sluitingsdatum aanbesteding (Eng.: tender closing date)
3	Hard-to-quantify evaluation criteria	Hoofdgunningscriterium (Eng.: main award criterion)
4	Excessively short or lengthy time used to decide on the submitted bids	Sluitingsdatum aanbesteding Datum gunning (Eng.: date of award)
5	Single-bidder contracts	Aantal inschrijvingen (Eng.: number of registrations Aantal elektronische inschrijvingen (Eng.: number of online registrations)
6	Selecting a supplier who bid well above the expected cost	Oorspronkelijk geraamde waarde (Eng.: original estimated value) Waarde (Eng.: value)
7	Repeatedly awarding contracts to the same suppliers	Naam aanbestedende dienst (Eng.: contracting authority) Officiële benaming (Eng.: official name)
8	Favouring getting procurement done quickly over following the proper process	Sluitingsdatum aanmelding (Eng.: closing date registration) Sluitingsdatum aanbesteding
9	Legality protocols	Juridisch kader (Eng.: legal framework)
10	Tender exceptionally large	Waarde
11	Amount of missing information	Empty cells

12	Time between bid award and actual contract signing date	Datum gunning Aanvang opdracht (Eng.: start of assignment)

Table 3 - Implemented red flags TenderNed

A red-flagged procedure type uses less open transparent procedure types such as open and invitation tenders. These non-open tenders are accelerated, restricted, awarded without publication, negotiated, and tenders without competition.

An advertisement period is the number of days between publishing a tender and the submission deadline. According to Article (2.71 lid 1 UAE, 2012), the time limit for submitting tenders shall be at least 45 days from the date of publication of the notice for open procedures. For non-open procedures, competitive procedures with negotiation, competitive dialogue procedures, and innovation partnership procedures, the time limit is at least 30 days.

Hard-to-quantify evaluation criteria are evaluation criteria which are not related to price.

Excessively short or lengthy time is used to decide on the submitted bids. According to Article (2.134 UAE, 2012), the contracting authority that has awarded a public contract shall publish the notice of the awarded public contract within 30 days.

Single-bidder contracts are contracts awarded in which only one bid was submitted. The bids are shown in the column “aantal inschrijvingen” (Eng.: number of registrations) and “aantal elektronische inschrijvingen” (Eng.: number of electronic registrations).

Selecting a supplier who bid well above the expected cost. There is no exact measure to define above the expected cost.(2.163d UAE, 2012) Article (2.163d UAE, 2012) states a government contract may be amended without a new tendering procedure if the increase in price does not exceed 50% of the value of the original contract. So, the limit shall be set at 50% above the expected cost.

Repeatedly awarding contracts to the same suppliers. This can be examined by looking at the column “Officiële benaming” (Eng.: Official name), and whether the contracting party awards them multiple times.

Favouring getting procurement done quickly over following the proper process is known as an expedited process. It has to abide as an urgent situation with no benefit for the contracting authority, according to (2.74 UAE, 2012). The deadline for submitting tenders may be set for at least 15 days from the publication date in an expedited process.

All legal protocols are based on “Aanbestedingswet 2012” (Eng.: Procurement Act 2012).

Exceptionally large tenders are “tenders which have an average size value plus two times the standard deviation” (Wensink & de Vet, 2013). In the final sample of the TenderNed dataset, this value is €142,113,919.

The amount of missing information is determined by the number of empty cells.

The time between the bid award and the actual contract signing date depends on whether the timing is accelerated or delayed. It gives a sense of false urgency and restriction of competition. According to article (2.127 UAE, 2012), the time between the bid award and the actual contracting signing is called the Alcatel term and lasts at least 20 days.

3.4 Using the Random Forest algorithm to solve a classification problem

Machine learning techniques refer to the method of red flag detection in different phases of the procurement process and techniques to verify the usefulness of red flags. Possible red flag detection machine learning techniques are Naïve Bayes, logistic regression, support vector, Random Forest, decision tree, gradient boosting, K-Nearest Neighbour, and binary classification. Measures depend on the problem type (classification or regression), but possible measures to verify usefulness are accuracy, precision, recall, F1-score, mean squared error and R-squared score (James et al., 2021).

This study employs a quantitative research approach to answer the research question and sub-questions because open data sources of public procurement will be analysed for the accuracy of machine learning (ML) models using chosen red flags. The study applies correlation research because it investigates the relationship between two variables. The correlation research will be performed through a predictive model. Predictive models are defined as: “the process of developing a mathematical tool or model that generates accurate prediction” (Kuhn & Johnson, 2013). Machine learning is related to predictive modelling because an algorithm must be developed to make computers learn from data and make predictions. This is otherwise known as supervised learning with a (multiclass) classification problem because the model will predict a label (corruption or no corruption). Statistical learning provides a framework for analysing data. Figure 7 presents a graph of different statistical learning methods showing the trade-off between flexibility and interpretability. The thesis will rely on secondary data. A predictive model used solely for prediction and not for interpretability uses a less flexible method (James et al., 2021).

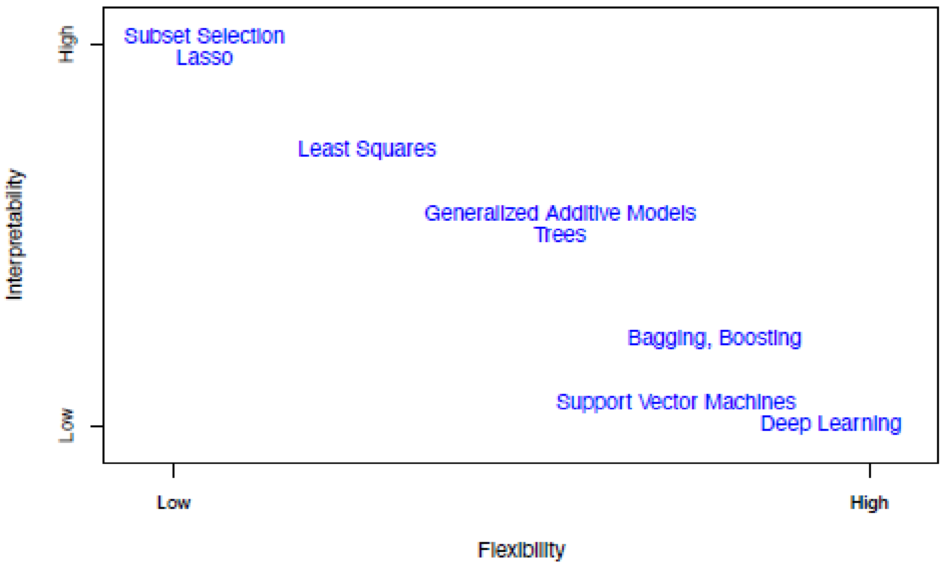

Figure 7 - Trade-off interpretability and flexibility on different statistical methods (James et al., 2021)

(Lima & Delen, 2020) found through his research that Random Forest was the most accurate classification technique in their prediction of Corruption Perception Indices (CPI). Random Forest is a tree technique. It is moderately flexible and interpretable, positioning it in the middle of the trade-off (see Figure 7). However, according to (James et al., 2021), trees have a lower level of predictive accuracy compared to other regression and classification approaches unless the decision trees are aggregated with methods such as Random Forest. In the literature review, the Random Forest algorithm is the most effective and most used approach to detect corruption (García Rodríguez et al., 2022; Hanafy & Ming, 2021; Lyra et al., 2022). A Random Forest algorithm will therefore be used to make predictions. The data and Random Forest will be analysed with the R package, using the programming language R[1].

3.5 Random Forest algorithm

Random Forest is an ensemble learning method comprised of decision trees because it comprises several decision trees. Decision trees are constructed as the tree's root, split into two nodes and then split again into two nodes etc. The default number of trees (M) in R is 500, but (Biau & Scornet, 2016) suggests letting it tend to infinity. The general steps of a Random Forest algorithm are presented in Figure 8.

[1] Appendix III – R-code Random Forest

Algorithm 1: Breiman's random forest predicted value at x.

Input: Training set $\mathcal{D}_n$, number of trees $M > 0$, $a_n \in \{1, \ldots, n\}$, $\texttt{mtry} \in \{1, \ldots, p\}$, $\texttt{nodesize} \in \{1, \ldots, a_n\}$, and $\mathbf{x} \in \mathcal{X}$.

Output: Prediction of the random forest at $\mathbf{x}$.

for $j = 1, \ldots, M$ **do**
- Select a_n points, with (or without) replacement, uniformly in $\mathcal{D}_n$. In the following steps, only these a_n observations are used.
- Set $\mathcal{P} = (\mathcal{X})$ the list containing the cell associated with the root of the tree.
- Set $\mathcal{P}_{\text{final}} = \emptyset$ an empty list.
- **while** $\mathcal{P} \neq \emptyset$ **do**
  - Let A be the first element of $\mathcal{P}$.
  - **if** *A contains less than* $\texttt{nodesize}$ *points or if all* $\mathbf{X}_i \in A$ *are equal* **then**
    - Remove the cell A from the list $\mathcal{P}$.
    - $\mathcal{P}_{\text{final}} \leftarrow Concatenate(\mathcal{P}_{\text{final}}, A)$.
  - **else**
    - Select uniformly, without replacement, a subset $\mathcal{M}_{\text{try}} \subset \{1, \ldots, p\}$ of cardinality $\texttt{mtry}$.
    - Select the best split in A by optimizing the CART-split criterion along the coordinates in $\mathcal{M}_{\text{try}}$ *(see text for details)*.
    - Cut the cell A according to the best split. Call A_L and A_R the two resulting cells.
    - Remove the cell A from the list $\mathcal{P}$.
    - $\mathcal{P} \leftarrow Concatenate(\mathcal{P}, A_L, A_R)$.
  - **end**
- **end**
- Compute the predicted value $m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)$ at $\mathbf{x}$ equal to the average of the Y_i falling in the cell of $\mathbf{x}$ in partition $\mathcal{P}_{\text{final}}$.

end

Compute the random forest estimate $m_{M,n}(\mathbf{x}; \Theta_1, \ldots, \Theta_M, \mathcal{D}_n)$ at the query point $\mathbf{x}$ according to (1).

Figure 8 - Random Forest algorithm steps(Biau & Scornet, 2016)

The tenders will be divided into training and testing sets through bootstrapping with a ratio of 80:20. Bootstrapping randomises the sample of 1,248 tenders and creates many bootstrap samples of the same size. So, the training set will contain tenders which can appear multiple times and some may never appear. The tenders which don't appear in the training data make up the test data. The bagging process is: "multiple decision trees are fitted on separate bootstrapped samples, and their predictions are averaged to reduce the overall variance" (Quantstart, n.d.).

Next, the nodes are also being randomised. The red flags (see Table 3) are predictors. The total number of red flags from the indices is 186 (see Appendix I). Usually, the number of predictors is the sum of the rounded square root of predictors, although the analyst may also choose the amount (Rigatti, 2017). So, 12 red flags (square root of 186) will be used to calculate the number of nodes. The algorithm tests possible thresholds and concludes the best split to ensure pure nodes, containing only cases or controls (Rigatti, 2017). This process is commonly repeated 100 to 1000 times.

The formula for the Random Forest model is:

$$g(x) = \mathrm{f0(x)} + \mathrm{f1(x)} + \mathrm{f2(x)} + \dots$$

Figure 9 - Formula Random Forest (Github, n.d.)

g(x) is the final prediction of either level 0 (low corruption risk) through level 5 (high corruption risk). fi is the simple base model and, therefore, a decision tree. The fi is the input feature labelled red flags, see Table 3. x are the tenders. So, the tested model is specified as:

$$g(x) = procedure(tenders) + advertisement\ period(tenders) + criteria(tenders) + time\ submitted\ bids(tenders) + single\ bidder\ contracts(tenders) + bidding(tenders) + awarding\ same\ supplier(tenders) + quick\ process(tenders) + legality(tenders) + large\ tender(tenders) + missing\ information(tenders) + time(tenders)$$

The recall, precision, accuracy, and F1-score measures will be calculated to assess and evaluate the model. Figure 10 shows the overall process of the predictive algorithm.

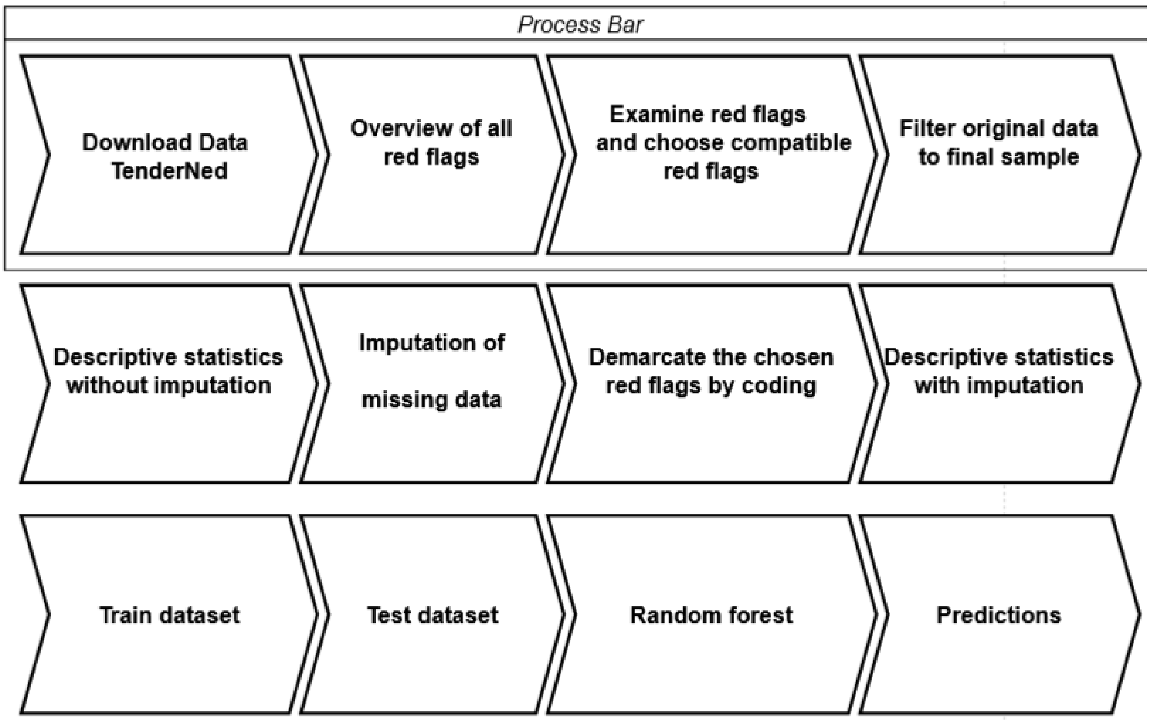

Figure 10 – Process

3.6 MICE is used to impute large amounts of missing data

A Random Forest algorithm doesn't operate with missing values. Therefore, the cases with missing values should be removed or imputed. The former is known as complete-case analysis. If all the missing values were removed, the dataset would only consist of 25 cases. However, too few cases mean a loss of precision because of less information (Roderick & Little, 2002). Imputation has two different methods: single imputation and multiple imputation.

Single imputation applies only one variable for each missing value and could lead to false precision due to the lack of sampling variability (Roderick & Little, 2002). Whereas multiple imputation applies multiple variables and combines the results. The two general multiple imputation methods are joint modelling and fully conditional specification. The former is based on the normal distribution, and the latter focuses on one variable at a time and influences (Huque et al., 2018). The fully conditional specification method: Multivariate Imputation by Chained Equations (MICE) was chosen to impute missing data because it is useful in handling large imputation procedures, it is flexible, and handles varying types of variables (Azur et al., 2011).

4) Results

The results help to answer the research question: “Can past-defined red flags predict the level of possible corruption in Dutch healthcare tenders through machine learning?” First, the handling of missing data is discussed. Secondly, descriptive statistics are discussed to summarise the data distribution and gain information about the patterns. Next, a comparison between the two models is presented. It looks at Out-Of-Bag (OOB) -error rate, classification error, specificity, Precision, Recall and F1-score to determine the best-suited model for a predictive machine learning model. Lastly, additional research was conducted because it relays information about the predictors’ influence, which could be useful for practitioners.

4.1 Missing values

During the data analysis phase, a significant number of missing data was discovered on the dataset filtered on health, date, threshold, and main activity (see Table 4). In total, 15% of the entire dataset is missing. There is no one clear reason why 15% of the data from TenderNed is missing, but a clarification from TED can be found. TED is short for Tenders Electronic Daily, and it contains public procurement tenders from the European Union and its members. The European Parliament requested a report on the limitations of TED data as a tool for data analysis. The report mentions multiple reasons for the large amounts of missing data which could also apply to TenderNed, such as the limited number of mandated (compulsory) fields; lack of compliance with and enforcement of (EU) legislation; and allowing much data to be marked as ‘not to be published’ (Ackermann et al., 2019).

Some predictors are missing most of their variables. As mentioned in Section 3.5, missing data was imputed with the MICE method, which has been credited with dealing with large amounts of missing data. However, the variable “Sluitingsdatum aanmelding” (Eng.: Closing date for registration) has been omitted due to 96% missing values during the imputation process (see Table 4). Therefore, RF8 (Favouring getting procurement done quickly over following the proper process) has been excluded from the final analysis.

Predictors	Number missing values
Sluitingsdatum aanmelding (Eng.: Closing date for registration)	1199
Aanvang opdracht (Eng.: Start of contract)	788
Hoofdgunningscriterium (Eng.: Main award criterion)	665
Sluitingsdatum aanbesteding (Eng.: Closing date for tender)	248
Waarde (Eng.: Value)	38
Procedure	6

Table 4 - Missing values per variable

4.2 Descriptive statistics

The section is divided into two parts. Both datasets have a count of 1,248 of the same tenders, but one dataset (original) isn't imputed while the other includes red flags and imputed values. If the section describing the statistics of the red flags shares the same variables as the original dataset, they'll be combined into one table in the second sub-paragraph.

4.2.1 Original

The names of the contracting authority and contractor parties are anonymised by giving them a number. Since there are 50 parties, Table 5 shows the top five contracting authorities who have awarded the most tenders. In this case, contracting authority 22 awards the most tenders. There are 376 contractors, in total, who are awarded tenders. Contractor 191 is awarded the most tenders, see Table 6.

Contracting authority	22	2	37	43	21
Count	309	209	171	76	70

Table 5 - Frequency awarded tenders by contracting authority

Contractor	191	6	50	55	349
Count	51	42	42	42	15

Table 6 - Frequency of awards to contractors

The number of registrations' median indicates that 50% of the measurements are below 4, and the other 50% is above 4. The highest recorded number of bidders is 36 for a tender. The same is true for the number of electronic registrations, but the median is 2, see Table 7.

The median of the original estimated value suggests that 50% of the data is below €2,890,000 and 50% is above. The difference between the 1st and 3rd quartile (Interquartile range) shows the variables are spread out and is also evident in the lowest and highest difference in cost. The same is true for the variable value, but the median is €1,800,000 see Table 7.

	Min	1st Quartile	Median	3rd Quartile	Max
Registrations	0	2	4	5	36
Electronic registrations	0	0	2	5	36
Original value	445,000	1,077,000	2,890,000	11,500,000	1,092,000,000
Value	0	800,000	1,800,000	11,500,000	668,577,000

Table 7 - Descriptive statistics original data

4.2.2 Imputation

The Procedures include five different types; see Table 8. Public procedures are the most frequently observed type and the competitive procedure with negotiation is the least observed type. The same is true for descriptive statistics with imputation, the missing values are all predicted as public procedures.

Procedure (RF1)	Competitive procedure with negotiation	Competitive dialogue	Non-public	Negotiation without publication	Public
Count original	3	4	61	71	1103
Count with imputation	3	4	61	71	1109

Table 8 - Frequency table RF1

The advertisement period is between publishing a tender and the submission deadline. It should be at least between 30 and 45 days depending on the procedure type. The lowest advertisement periods and 1st quartile are negative. However, it does have a median of 25 suggesting the typical advertisement date is around 25 days. The interquartile range indicates high variability in the advertisement periods, see Table 9. The negative advertisement period could indicate the contracting authority already has contacted a contractor to inform them about a tender. However, this seems unlikely because it would mean the tenders are submitted before the publication deadline and it would imply major issues within the procurement process. TenderNed defines publication date as the date an announcement was published. The online platform of TenderNed contains two publication dates: the announcement of an awarded contract and the announcement of an assignment. Naturally, the date of the announcement of an awarded contract is after the submission deadline. If the publication date in the dataset is defined as the date of the announcement of an award contract, the advertisement period will be negative.

The evaluation period is to decide on the submitted bids and should be decided within 30 days. The lowest evaluation period is negative and indicates a decision was already made before all the bids were submitted. The median lies around twice the number of days which is legally allowed. As observed in the advertisement period, the difference between the 1st and 3rd quartiles shows the variables are spread out, see Table 9. The negative evaluation period could indicate there was a bid early on which the contracting authority accepted. Another possibility is the "percelenregeling". It is a regulation where a contracting party sometimes needs to contract one specific party for a specific part of the contract (PIANOO, n.d.). Some scenarios to use the regulation are to contract smaller businesses or start-ups easily without an intensive registration process; to purchase innovative solutions easily; and a specific supplier can offer higher quality in a specific product group (PIANOO, n.d.).

The total bidders' median indicates that 50% of the measurements are below 4, and the other 50% is above 4. The highest recorded number of bidders is 72 for a tender, see Table 9.

Selecting a supplier who bids well above the expected cost is set at 50% of the value of the original contract. The median suggests that 50% of the data is below €1,445,024 and 50% is above. The difference between the 1st and 3rd quartile shows the variables are spread out and is also evident in the lowest and highest difference in cost, see Table 9.

Awarding the same supplier is calculated by determining whether the same contracting party awards a contract to the same supplier. The limit is set at 4 times because the boxplot's upper quartile is valued at 4, see Appendix III. The 3rd quartile suggests 75% of the tenders are awarded below 4 times from the same contracting party to the supplier. The remaining 25% of the tenders are awarded more than four times to the same supplier, see Table 9.

Exceptionally large tenders are defined by the average size plus two times the standard deviation. The median, 3rd quartile and highest difference in value indicate a small variability concerning the higher differences, see Table 9. However, the smallest value is negative and could suggest being an extreme outlier.

The contract execution period is the number of days between the bid award and the actual contract signing date and should be executed in at least 20 days. The negative values indicate the contract has already been signed before the announcement of the awarding of the bid. The 1st and 3rd quartiles indicate high variability in the contract execution periods and are supported by the huge difference between the lowest and highest differences, see Table 9.

	Min	1st Quartile	Median	3rd Quartile	Max
Advertisement period (RF2)	-2,312	-80	25	54.25	1,511
Evaluation period (RF4)	-1,699	39	66	135	2,307
Total bidders (RF5)	0	4	4	10	72
Difference cost (RF6)	222,500	538,578	1,445,024	5,750,000	546,087,500
Same supplier (RF7)	1	1	2	4	47
Difference value (RF10)	-526,773,992	130,303,008	140,003,008	141,033,610	141,803,008
Contract execution (RF12)	-2,194	-169	12	197.75	2,596

Table 9 - Descriptive statistics imputed data

The main award criteria are categorised into two criteria, see Table 10. The best price-quality ratio is the most prevalent criterion. Again, the missing values are all predicted to be the best-priced-quality ratio.

Main award criterion (RF3)	Lowest price	Best price-quality ratio
Count original	14	569
Count with imputation	14	1234

Table 10 - Frequency table RF3

4.3 Comparisons between Random Forest Models

In the previous section, two datasets were mentioned—one without imputed values and one with imputed values and red flags. The dataset with imputed values and demarcated red flags shall be used for the Random Forest models since it doesn't operate with missing values. The analysis looks at the confusion matrix, performance metrics and a SHAP beeswarm plot to determine the best-suited model for a predictive machine learning model and the predictor's influence, see Figure 11.

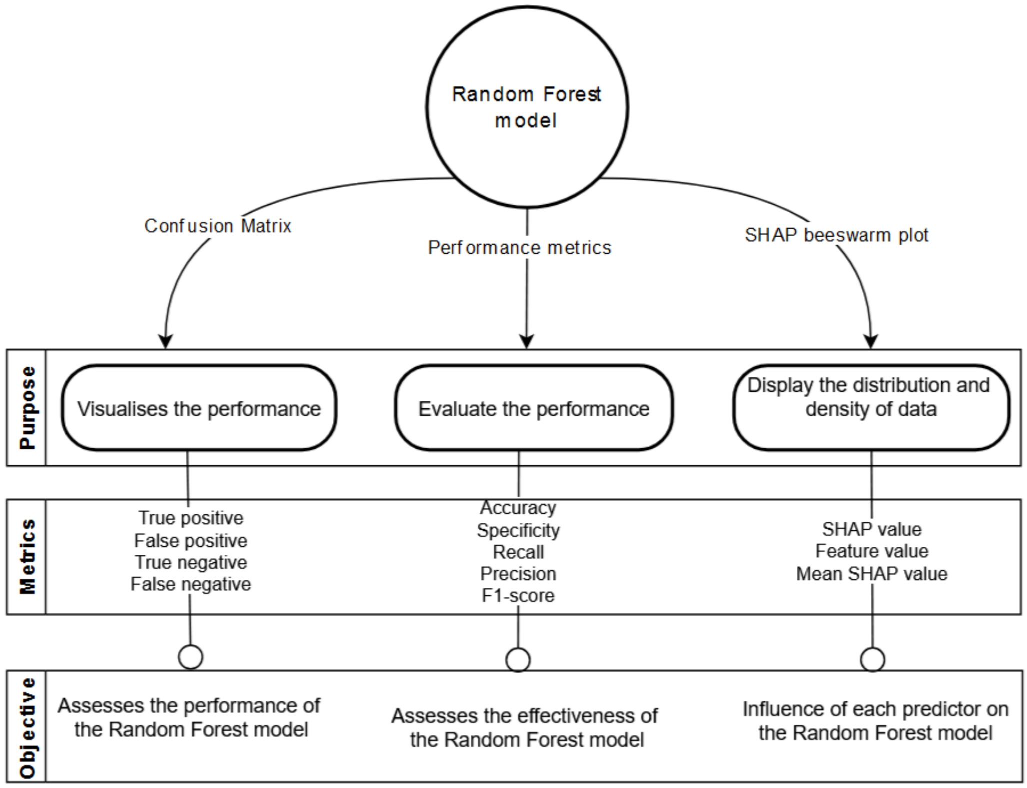

Figure 11 - Overview models

In this sub-section, a comparison between two Random Forest models is made. During the coding phase, four different models were created. Models 1 and 2 have different variables compared to Models 3 and 4. Model 1 contains dichotomous variables of RF1 t/m R12, where possible corruption is dichotomous (0-1). Model 2 contains dichotomous variables of RF1 t/m R12, where corruption is measured in levels 0 through 5. Model 3 contains numerical or categorical variables of the red flags, where corruption is dichotomous (0-1). Lastly, Model 4 contains numerical or categorical variables of RF1 t/m R12, where corruption is measured in levels 0 through 5. In Models 1 and 3, a tender with one red flag would be flagged as a high- risk of corruption. However, the presence of one red flag does not necessarily mean it would be a high-risk corrupt tender. Whereas Models 2 and 4 measure the risk of corruption on a scale from levels 0 through 5. Therefore, Models 2 and 4 were selected for comparison.

4.3.1 Confusion Matrix

The confusion matrix assesses the machine learning model by comparing the predicted values against actual values by determining the true positives, true negatives, false positives and false negatives (Murel & Kavlakoglu, 2024). True positives are the number of cases which are correctly predicted at the correct level. False positives are cases which are predicted as true but are false, and false negatives are cases which are predicted to be false when it is true. If a case is incorrectly predicted as level 2 while it is level 1, it will be a false negative for level 1 and a false positive for level 2. False negatives are the values of corresponding rows except for the TP value and false positives are the values of the corresponding column except for the TP value. True negatives are the remainder of the cases.

The OOB estimate of error rate includes samples outside of the bootstrapped samples and indicates the proportion of misclassified cases. The OOB estimate of the error rate in Model 2 is 1.99% higher than in Model 4, see Tables 11 and 12. Furthermore, the confusion matrix of Model 2 shows that levels 1 and 4 have a class.error rate of less than 10% and 20% respectively. To clarify, in level one 10% is wrongly predicted to level 2 and in level four 20% is wrongly predicted to level 3. Level 5 shows an error rate of 80%. In other words, 80% of the tenders are wrongly predicted as level 4 and 20% are correctly predicted as level 5. Model 4 shows a better class.error rate for level 5, but continuous to be a large value of 55%. Moreover, the accuracy of Model 4 is 1.6% higher than Model 2.

In other words, Model 4 is more accurate at predicting the correct level of possible corruption than Model 2. However, Model 2 outperforms Model 4 regarding levels 0 and 2. Nonetheless, Model 4 overall performance estimates the tenders at different levels of possible corruption better than Model 2. As mentioned previously, there are five levels of possible corruption the tenders can be categorised. Consequently, each level has its metrics such as sensitivity, specificity, precision etc. In Tables 13 and 14, both models are displayed with these metrics.

Actual/ Predicted	0	1	2	3	4	5	Class.error
0	34	0	0	0	0	0	0.0000
1	0	113	13	0	0	0	0.1032
2	0	4	351	1	0	0	0.0140
3	0	0	20	299	0	0	0.0627
4	0	0	0	27	116	0	0.1889
5	0	0	0	0	16	4	0.8000
OOB estimate of error rate: 8.12%				Accuracy: 0.94			

Table 11 - Random Forest output Model 2

Actual/ Predicted	0	1	2	3	4	5	Class.error
0	33	1	0	0	0	0	0.0294
1	1	118	7	0	0	0	0.0635
2	0	7	345	4	0	0	0.0309
3	0	0	15	301	3	0	0.0564
4	0	0	0	11	131	1	0.0839
5	0	0	0	0	11	9	0.5500
OOB estimate of error rate: 6.11%				Accuracy: 0.956			

Table 12 - Random Forest output Model 4

4.3.2 Performance metrics

Class-wise specificity focuses on the performance of each level compared to all other levels and looks at whether cases in one level are falsely identified for other levels (Kautz et al., 2017). A low specificity score means the classification of elements from a level is easily mistaken for another level, and a high specificity score is measured if the classification of elements in a level is different from all the other levels. In other words, a low specificity score means the levels aren’t well-defined while a high specificity score is. Both models show a high level of specificity around 0.95-1.0 on all levels, meaning the levels are well-defined and distinct from each other and tenders are easy to classify to the correct level, see Tables 13 and 14.

Precision or positive predictive value is the proportion of predicted positives being positives (Wang & Zheng, 2013). Model 2 shows perfect precision in level 0 and level 3. The tenders with an average possibility of corruption are correctly classified in level 3 and tenders with no possibility of corruption are correctly classified in level 0. Level 5 shows “Not available” as the precision metric. However, based on the confusion matrix (see Table 11), the precision is valued at 1.0000 (4/(4+0)). Consequently, it would indicate tenders which were predicted to be level 5 are assigned the correct level.

Recall or sensitivity measures how well a machine learning model can detect positive instances (True Positive) (Swift et al., 2020). Model 2 shows levels 0 through 4 having a high recall value (around 0.9-1.0), while level 5 shows a value of 0. A low recall value means the model fails to detect tenders belonging to level 5. Whereas Model 4 shows high values in levels 0 through 4 and a better value in level 5.

F1 assesses the predictive ability of a model by examining its performance in each class individually rather than considering overall performance like accuracy does (Goutte & Gaussier, 2005). Model 2 shows levels 0 through 4 with a high F1 score (around 0.9-1.0). Level 5 again shows a score of "Not available". However, based on the confusion matrix (see Table 11), the F1- score is valued at 0 (2*(1.000*0)/(1.000+0). A low F1 score means poor model performance of level 5. The poor performance can be observed in the low recall score. Whereas Model 4 shows better F1 scores and performance on levels 2 through 5.

	Level 0	Level 1	Level 2	Level 3	Level 4	Level 5
Specificity	1.0000	0.9953	0.9408	1.0000	0.9769	1.0000
Precision	1.0000	0.9722	0.9151	1.0000	0.8718	NA
Recall	1.0000	0.9211	0.9898	0.9118	1.0000	0.0000
F1	1.0000	0.9459	0.9510	0.9538	0.9315	NA

Table 13 - Model 2 metrics

	Level 0	Level 1	Level 2	Level 3	Level 4	Level 5
Specificity	0.9918	0.9953	0.9803	1.0000	0.9769	1.0000
Precision	0.7778	0.9714	0.9700	1.0000	0.8718	1.0000
Recall	1.0000	0.8947	0.9898	0.9559	1.0000	0.4000
F1	0.8750	0.9315	0.9798	0.9774	0.9315	0.5714

Table 14 - Model 4 metrics

4.4 Comparisons between SHAP beeswarm plots

In addition, the research was conducted about the influence of the predictors. A SHAP beeswarm plot displays the influence of the predictors (red flags) on the impact of a model (possibility of corruption) based on its feature value of the predictor. The plot shows the importance from most to least influential based on the mean SHAP value (see values next to the predictors). The focus is the comparison of the two plots by examining the range (wide or narrow) and density (high or low). The range implies variability in the SHAP value or impact on the level of possible corruption across various data points. The density implies the concentration of data points and can be highlighted with clusters.

4.4.1 Model 2

Model 2 predictors are dichotomously coded, and therefore the feature value is either low or high, see Figure 12. The lower feature values of the predictors negatively affect the corruption risk and therefore the level of possible corruption will be smaller. In comparison, the higher feature values of the predictors have a higher impact on the risk of corruption leading to a higher level.

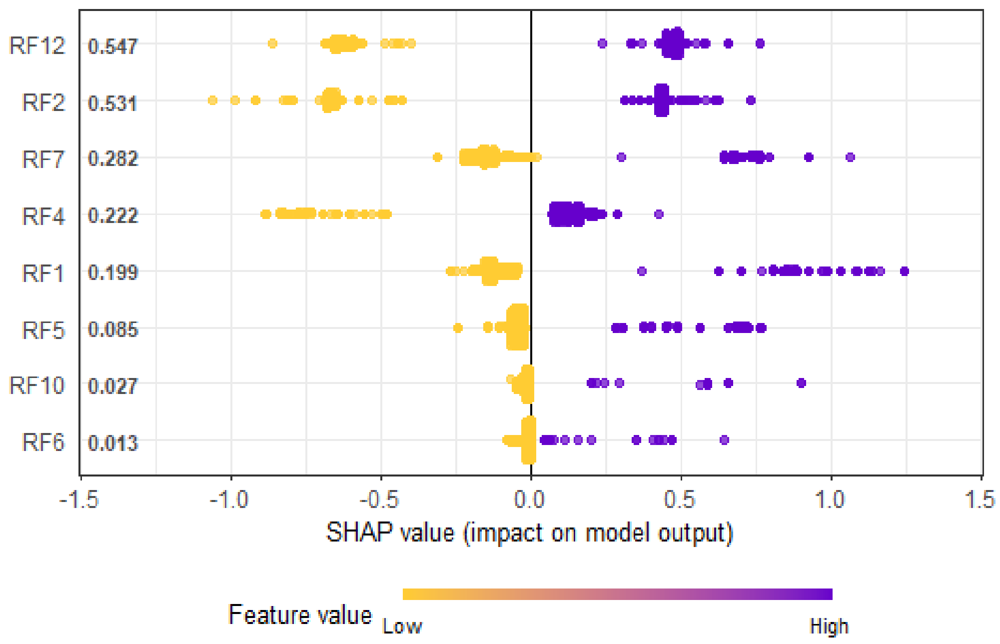

Figure 12 - SHAP beeswarm plot Model 2

Predictor RF12 shows tenders with a shorter Alcatel term (time between bid award and actual contract signing date) having a high feature value, which influences the risk of corruption towards a higher level. There are fewer tenders with a longer Alcatel term which influences the risk of corruption towards a lower level. The high feature values show a wide range with high density observed by a cluster of data points in the middle. The cluster indicates the data points have a similar value or are frequently observed. While the lower feature values show a smaller range with low density.

Predictor RF2 shows tenders with a short advertisement period (high feature value) influencing the risk of corruption towards a higher level of possible corruption. Whereas tenders with a longer advertisement period are influencing the risk of corruption to a lower level. The lower and higher feature values show a wide range of data points and low density but with a small cluster of data points. The observation indicates a varying distribution of values across the dataset for RF2.

Predictor RF7 shows that when tenders are awarded more than four times to the same supplier from the same contracting party, the feature value is high. When tenders are awarded to different suppliers, the feature value is low. The high feature values show a wide range of data points, which could suggest varying data in the dataset. The lower feature values show a cluster of data points with a low SHAP value or level of possible corruption.

RF4 shows when bids are submitted after 30 days the feature value is high, and before 30 days the feature value is low. The period of submitting bids before 30 days has a wide range of data points and has low density. It can be suggested the data, regarding RF4, varies greatly. Meanwhile, the period of submitting bids after 30 days has a high density, as observed by the concentration or cluster of data points. It suggests that certain values are more common or frequently observed, but they have a low SHAP value or positive impact on the level of possible corruption.

Predictor RF1 shows non-open procedures and procedures with negotiations having high feature values. The high feature value data points show a wide range and low density. The observation indicates considerable variation in the frequency of different types of non-open procedures. The open procedures (low feature value) show a narrow range with high density. However, they have a low impact on the level of possible corruption.

Predictor RF5 shows single bidder contracts as high feature values, while more bidders have a low feature value. The low feature value has a narrow range and high density. The cluster indicates there is minimal variation in the number of bidders.

Predictor RF10 shows most tenders aren’t exceptionally large. The data points with a low feature value are clustered with a low impact on the level of possible corruption. There are a few tenders which are exceptionally large with a high feature value and more impact.

Predictor RF6 shows like RF10 few suppliers who bid well above the expected costs. They do however have a high feature value with a higher impact on the level of corruption. The lower feature value shows a narrow range and high density with a big cluster around the 0 SHAP value. The observation indicates that tenders who have suppliers bidding below the excepted costs don’t have a positive or negative influence on the level of corruption.

Predictors RF1, RF5, RF10, and RF6 have similar distributions. The high feature values have a wide range, low density and higher impact on the level of corruption. While the low feature values have a narrow range, high density and lower impact on the level of corruption. Overall, these predictors rank among the lowest four because the amount of data points in the lower feature values contributes more toward the mean SHAP value.

4.4.2 Model 4

Model 4 predictors are numerically coded; therefore, the feature value shows more variety than Model 2 (see Figure 13).

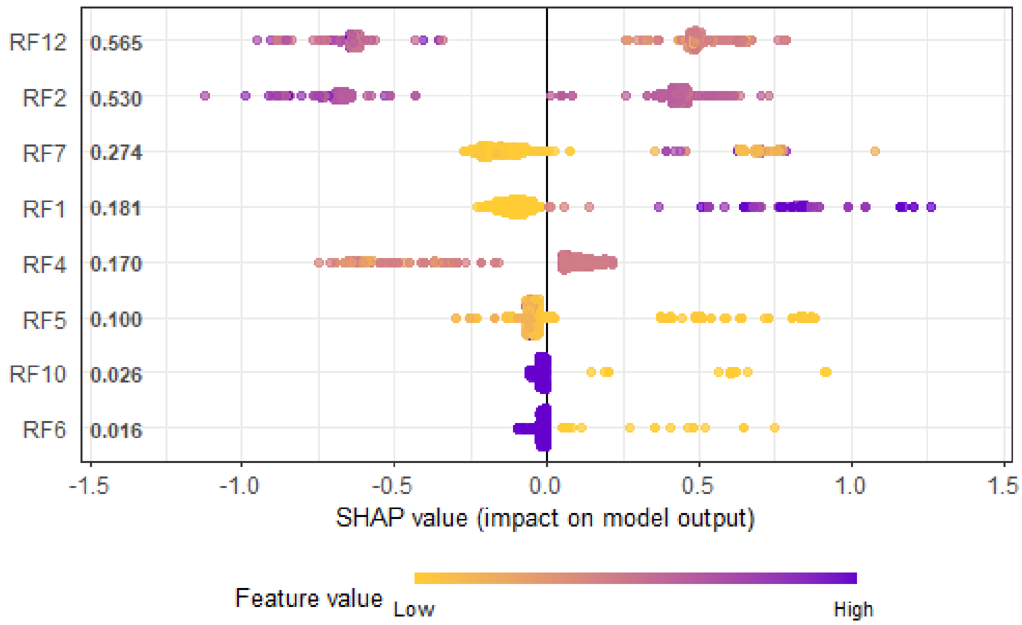

Figure 13 - SHAP beeswarm plot Model 4

Predictor RF12 shows its data points with a higher (SHAP) impact on the level of corruption are associated with lower to average Alcatel period. The data points with a lower (SHAP) impact on the level of corruption show an average to higher Alcatel period. Both have a wide range and a small cluster. A wide range indicates a varying distribution of SHAP values across the dataset, and the small cluster contains similar or frequently observed values. The predictor shows a lower Alcatel term leading to a bigger influence on the level of possible corruption.

Both RF2's sides exhibit an average to high feature value (advertisement period). The wide range from both sides indicates a varying distribution in impact, while a cluster at the higher SHAP value presents similar or frequently observed values. The predictor shows similar or frequently observed values, but the impact on the level of corruption differs depending on the advertisement periods.

Predictor RF7 shows when tenders have a low feature value their impact on the level of corruption is small. The range is narrow, and the density is high. The small number of averages to high feature values has a wide range and the density is low. The observation indicates when an average to a high number of times a tender is awarded to the same supplier is perceived, the level of possible corruption could vary in its impact. Whereas the lower number of times a tender is awarded to the same supplier, the level of corruption would hardly decrease.

Predictor RF1 shows when tenders have a low feature value their impact on the level of corruption is small. The range is narrow, and the density is high. The high feature values have a wide range and the density is low. The observation indicates when a tender follows a more non-open procedure the level of possible corruption varies in its impact (see Appendix III). Whereas the more open procedures hardly decrease the impact of the level of corruption.

Both RF4's sides exhibit an average feature value (evaluation period). The wide range from the low SHAP value side indicates a varying distribution in impact, while a cluster at the higher SHAP value presents similar or frequently observed values. The predictor shows the varying effects on the level of possible corruption from an average feature value.

Predictor RF5 mostly has low feature values. The higher impact data points are accompanied by a varying distribution. While the lower impact data points have a cluster of a SHAP value of 0. The observation indicates varying effects on the level of possible corruption from an average amount of bidders.

RF10 shows most tenders are exceptionally large. The data points with a low feature value are clustered with a high impact on the level of possible corruption. There are a few tenders which aren't exceptionally large with a high feature value and more impact. The observation indicates large tenders do not have a positive or negative influence on the level of corruption. Furthermore, it indicates smaller tenders vary in impact on the level of possible corruption.

Predictor RF6 shows like RF10 a lot of suppliers who bid well above the expected costs. They have a high feature value with a lower impact on the level of corruption. The higher feature value shows a narrow range and high density with a big cluster around the 0 SHAP value. While the lower feature values have a wide range. The observation indicates that tenders who have suppliers bidding above the excepted costs don't have a positive or negative influence on the level of corruption. Furthermore, it indicates lower bids vary in impact on the level of possible corruption.

4.4.3 Summary of findings from the SHAP beeswarm plots

Models 2 and 4 show the advertisement period (RF2) and contract execution period (RF12) as the most influential predictors of corruption. The predictors RF2 and RF12 have a major positive effect on the output with around 0.5 SHAP mean value units. So, if one of the predictors were present, the level of possible corruption would increase by 0.5 points. Another interesting finding is the difference between RF6 and RF10 between Model 2 and Model 4. It stands out because it presents a conflicting viewpoint.

Furthermore, both models are missing the same four red flags (RF3, RF8, RF9, and RF11) because they have a SHAP mean value of 0. Hard-to-quantify criteria (RF3), missing data (RF8), legality protocols (RF9), and the amount of missing data (RF11) all have very few tenders or data points that are considered to contribute to the level of possible corruption.

In addition, research was conducted on the influence of the predictors. Even though using a SHAP beeswarm plot doesn't directly answer the research question or is part of the sub-questions, it reveals all the past-defined red flags have a positive effect on the level of possible corruption. It suggests the chosen 12 past-defined red flags are indeed relevant and influential to the level of possible corruption. When looking at the mean SHAP value, none of the predictors hurt the output of the models. This means all the predictors increase the level of corruption positively. So, if RF7 is flagged within a tender the level of possible corruption will increase by about 0.3 and if RF12 is flagged within a tender the level of possible corruption will increase by about 0.6. In other words, when a red flag occurs within a tender the level of possible corruption increases.

5) Discussion

5.1 Key Findings

The purpose of this thesis was to verify the accuracy of analysed red flags using a machine learning predictive model. The corresponding research question of this thesis is: *"Can past-defined red flags predict the level of possible corruption in Dutch healthcare tenders through machine learning?"*

The first sub-question was: "How accurate are machine learning models while making the prediction?" Machine learning models are used for many issues, such as making predictions. Multiple machine learning models, such as logistic regression models, Naive Bayes, k-nearest neighbours, decision trees, neural networks, vector-machine, and Random Forest, have been introduced and assessed based on the existing literature to address the sub-question. Furthermore, well-known parameters were introduced to assess a machine-learning model such as accuracy, precision, recall and F-measure. In the literature overview, it was revealed the Random Forest algorithm is the best-performing and most effective machine learning method to detect collusion accurately based on previous cases. The Random Forest algorithm was shown to be the most accurate due to very high recognition efficiency. Furthermore, Random Forest is shown to be the better model for classifying and distinguishing classes. The literature overview showed the accuracy of the Random Forest algorithm was between 88% - 99%. Therefore, the answer to the first sub-question is that the machine learning models are quite accurate in making predictions, especially the Random Forest algorithm. Therefore, the decision was made to execute the research with the Random Forest algorithm.

The second sub-question that was studied was: "How should the machine learning predictive model operate?" The Random Forest algorithm operates with a dependent and independent variable. The dependent variable is the risk of possible corruption, and the independent variable is red flags. Corruption occurs in many forms but is within the healthcare sector "bribery; collusion; under-the-table payments for care; bid rigging; absenteeism; unethical drug promotion; inappropriate ordering of tests and procedures to increase financial gain; lack of accountability; biased application; nepotism; unnecessary referrals; political influence; and use of government resources for private practice as corrupt practices" (Vian, 2008). The literature overview revealed governmental entities and scholars created red flag indices to combat possible corruption. Red flags are an accumulation of traces that may point to the presence of corrupt activities. All 186 red flags were analysed whether they were compatible with the data and 12 red flags were left. To ensure the predictive model is coded correctly and the model is accurately predicting the outcomes, four different models were created. Two models were analysed in detail because the method of how the variables were coded was consistent with the research question. The models which were disregarded were coded with the assumption corruption is dichotomous, either a tender is corrupt or is not corrupt. The first model was coded with dichotomous predictors (Model 2) and the second model was coded with numerical or categorical predictors (Model 4). To extend the scope of the sub-question, the predictors were examined on their influence regarding the level of possible corruption. Using a SHAP beeswarm plot, two red flags were discovered to be most influential towards determining the level of possible corruption in both models.

5.2 Overfitting

Overfitting occurs when the test data follows the errors of the training data too closely. It could result in the algorithm not being able to make accurate predictions and being unable to generalise new data (IBM, n.d.). However, according to the trademark owner of Random Forest (Breiman, 2001), there is no possibility of overfitting when using the Random Forest model because it follows the Strong Law of Large Numbers. The Strong Law of Large Numbers means as the number of trees increases, the average of the output will become the expected value. Furthermore, the model has a bigger chance of getting more of the right and best possible predictions due to the high amount of trees (Breiman, 2001). Therefore, the model becomes less sensitive to fitting the training data too closely by the number of trees. According to (Hengl et al., 2018; Sarica et al., 2017), Random Forest algorithms have built-in protection against overfitting, are resistant to noise, and are stable in the presence of outliers compared to other machine learning algorithms.

5.3 Theoretical Implications

This thesis has several theoretical contributions. Firstly, it theoretically contributes to the literature because it is the first study to use Dutch healthcare data to determine whether previously defined red flags can predict the level of possible corruption. To my knowledge, it is the first thesis to combine the governmental red flag indices and corruption risk indices created by scholars and apply them to the Dutch healthcare sector. Previous research measured corruption in different contexts such as the European public procurement (Fazekas & Kocsis, 2020), bank projects (Kenny & Musatova, 2010), and an Italian private company (Decarolis & Giorgiantonio, 2022).

In addition, this thesis used four models and closely analysed two Random Forest models based on alternative coding for the variables. The two models were chosen because they measured possible corruption on a scale through levels. Moreover, the thesis shows the scale variable of possible corruption performs better than the models based on a dichotomous dependent variable. It gives insight into how variables should be defined.

Lastly, this thesis showed the SHAP beeswarm plot to be a helpful tool for determining which red flag is the most influential on potential corruption. It gives insight into how the red flags distinguish between lower and higher levels of potential corruption risk in new tenders, and how it affects the level of possible corruption.

5.4 Practical Implications

Section 2.5 presented an overview of the indices and red flags created by governmental entities and scholars to control corruption. The indices and red flags are treated with equal consideration. A contribution of this thesis in this field was to question the equal consideration of each red flag. Bribery; collusion; under-the-table payments for care; bid rigging; absenteeism; unethical drug promotion; inappropriate ordering of tests and procedures to increase financial gain; lack of accountability; biased application; nepotism; unnecessary referrals; political influence; and use of government resources for private practice are seen as corrupt practices in the healthcare sector (Vian, 2008). The SHAP beeswarm plot shows certain red flags to be more influential for possible corruption than other red flags. The red flags advertisement period and contract execution period are the most reliable when evaluating a healthcare-related tender because they are the most influential. Institutions that are investigating are now able to narrow down healthcare-related corrupt practices based on the red flags. Tenders which contain one or both of those red flags can now be investigated for bribery, collusion, bid rigging, and biased application. These healthcare-related corrupt practices stand out because they could be the cause of triggering the red flag advertisement period and/ or contract execution period. Furthermore, certain actors within the healthcare system are responsible for certain corrupt practices, such as governmental regulators taking bribes from providers (Vian, 2008). Consequently, they can be detected in a certain step of the procurement process such as the tendering and decision-making phase or the pre- and post-award phase.

5.5 Conclusion

This thesis has demonstrated the potential of accurately predicting possible corruption through levels using red flags from Dutch healthcare data. Furthermore, it demonstrated through the machine learning algorithm Random Forest that predictions about the level of potential corruption are possible.

The red flags were extracted from past-defined corruption indices which are based on past corrupt cases. The red flags were coded dichotomously for Model 2 and as a numeric variable for Model 4. According to the confusion matrix, Model 4 overall performance estimates the tenders at different levels of possible corruption more accurately than Model 2. The prediction of tenders which are assigned at lower levels of possible corruption is observed to be especially accurate. F1 assesses the predictive ability of a model by examining its performance in each class individually rather than considering overall performance. Model 4 shows better F1 scores and performance than Model 2.

Moreover, the levels of corruption are well-defined and distinct from each other and tenders are easy to classify to the correct level. It is observable in both models that the specificity scores were very high. These findings align with the framework of (Lima & Delen, 2020), who argue using multiple class options also leads to high predictive accuracy just like binary class options.

In summary, it is possible to predict the level of potential corruption of tenders in Dutch healthcare through machine learning with well-defined levels. The predictions are more accurate when predictors are coded numerically rather than dichotomously.

The levels of possible corruption are derived from red flags and are observed in the model. It determined that the red flags advertisement period and contract execution period were most influential on the risk of corruption. Using a SHAP beeswarm plot, it was discovered that both red flags from both models attribute an increase of 0.5 points in the level of possible corruption.

Overall, this thesis adds to the ongoing discussion about predicting corruption and red flags in a singular sector. Predicting possible corruption in a Dutch healthcare context fits into the framework of (Fazekas, 2020, 155-164) which posits looking into regions, sectors, organisations or individuals' behaviour, which has long been thought to be necessary for advancing the field.

5.6 Limitations and Future Research

Throughout the thesis, several limitations have been identified. A limitation of this thesis is the specific healthcare corruption risks. Section 2.3 encloses that corruption takes on different types of forms in the healthcare sector: misuse of level positions, embezzlement of medicines and medical devices, improper marketing relations different types of corrupt practices. However, they are not included as red flags in the model, because it was not possible to align those types of forms with the available data. Therefore, the models contain a more general view of what would be considered a possible corruption risk case.

Even though the great advantage of the approach is the large amount of data readily available, another limitation lies in the amount of missing data within the dataset. Around 15% of the values are missing in the entire dataset. However, some columns are missing most of their values such as “Sluitingsdatum aanmelding” (Eng.: Closing date for registration) for the input for RF8. Even though RF8 was relevant, the decision was made to remove the red flag. Furthermore, the remaining missing red flag values were imputed. However, RF11 addresses the missing data. Even though in the original dataset missing data does occur, no tenders are assigned with RF11 due to imputation. Such actions may result in a less authentic representation of assigning the correct level of possible corruption to the tenders. However, “Sluitingsdatum aanmelding” (Eng.: Closing date for registration) accounts for 6% of the missing data and is the highest proportion from the dataset. Moreover, the remaining missing red flag values were imputed based on a larger amount of original data through the MICE method which is equipped to handle large amounts of missing data. In summary, it currently assigns the level of possible corruption based on 11 red flags, rather than twelve with the assumption all data is available for use.

Another limitation lies in the SHAP beeswarm plot. The difference between RF6 (selecting a supplier who bid well above the expected cost) and RF10 (tender exceptionally large) from Model 2 and Model 4 cannot be explained in this plot, because the interaction between red flags isn’t visible.

There is room for future research by applying a different sample, from another country/ government, to capture the influence of red flags on healthcare-specific corruption risks. This would capture the difference in predictability between general corruption risks and healthcare-specific corruption risks. Moreover, there is also room for researching the interactions between predictors to present how the level of influence is explained. For example, a tender might display both a short advertisement period combined with repeatedly awarding contracts to the same supplier. Both red flags can influence each other and might affect the level of possible corruption differently than if only one was present. A restricted procedure type combined with single-bidder contracts might affect the influence on the level of possible corruption differently compared to a negotiated procedure with single-bidder contracts.

References

2.71 lid 1 UAE. (2012). *Article 2.71 lid 1 UAE: Termijnen bij openbare procedure*. Retrieved from https://wetten.overheid.nl/BWBR0032203/2022-03-02/0#Deel2_Hoofdstuk2.3_Afdeling2.3.2_Paragraaf2.3.2.3_Artikel2.71

2.74 UAE. (2012). *Article 2.74 UAE: Termijnen*. Retrieved from https://wetten.overheid.nl/BWBR0032203/2022-03-02/0#Deel2_Hoofdstuk2.3_Afdeling2.3.2_Paragraaf2.3.2.3_Artikel2.74

2.127 UAE. (2012). *Article 2.127 UAE: Gunningsbeslissing*. Retrieved from https://wetten.overheid.nl/BWBR0032203/2022-03-02/0#Deel2_Hoofdstuk2.3_Afdeling2.3.8_Paragraaf2.3.8.8

2.134 UAE. (2012). *Article 2.134 UAE: Verslaglegging en bekendmaking*. Retrieved from https://wetten.overheid.nl/BWBR0032203/2022-03-02/0#Deel2_Hoofdstuk2.3_Afdeling2.3.8_Paragraaf2.3.8.9_Artikel2.134

2.163d UAE. (2012). *Article 2.163d UAE: Wijziging van overheidsopdrachten*. Retrieved from https://wetten.overheid.nl/BWBR0032203/2022-03-02/0#Deel2_Hoofdstuk2.5

Abdou, A., Basdevant, O., David-Barret, E., & Fazekas, M. (2022). Assessing Vulnerabilities to Corruption in Public Procurement and Their Price Impact. *IMF: Working Papers*. https://www.imf.org/en/Publications/WP/Issues/2022/05/20/Assessing-Vulnerabilities-to-Corruption-in-Public-Procurement-and-Their-Price-Impact-518197

Ackermann, R., Sanz, M., & Sanz, A. (2019). *Gaps and Errors in the TED database*. https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/CONT/DV/2019/03-11/CONT_GapsandErrorsintheTEDdatabase_EN.pdf

Aman, A., & Kasimin, H. (2011). E-procurement implementation: a case of Malaysia government. *Transforming Government: People, Process and Policy*, *5*(4), 330-344. https://doi.org/10.1108/17506161111173586

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, *20*(1), 40-49. https://doi.org/10.1002/mpr.329

Badillo, S., Banfai, B., Birzele, F., Davydov, I. I., Hutchinson, L., Kam-Thong, T., Siebourg-Polster, J., Steiert, B., & Zhang, J. D. (2020). An Introduction to Machine Learning. *Clinical Pharmacology & Therapeutics*, *107*(4), 871-885. https://doi.org/https://doi.org/10.1002/cpt.1796

Bartlett Quintanilla, P., Cummins-Tripodi, P., Marié, A., & Mussio, E. (2018). *The cost of corruption across the EU*. https://www.greens-efa.eu/files/assets/docs/the_costs_of_corruption_across_the_eu.pdf

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197-227. https://doi.org/10.1007/s11749-016-0481-7

Bîzoi, A.-C., & Bîzoi, C.-G. (2023). Primum Non Nocere: How to Fight the “Pandemic” of Healthcare Corruption. In R. W. McGee & S. Benk (Eds.), *The Ethics of Bribery: Theoretical and Empirical Studies* (pp. 345-365). Springer International Publishing. https://doi.org/10.1007/978-3-031-17707-1_20

Bobowski, S., & Gola, J. (2018). E-procurement in the European Union. *The Asia-Pacific Journal of European Union Studies, materiał powielony złożony do druku*, *17*(1-2018), 23-35. https://www.researchgate.net/publication/329322569_E-Procurement_in_the_European_Union

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, *15*(4), 233-234. https://doi.org/10.1038/nmeth.4642

Camboni, R., & Valbonesi, P. (2021). Incumbent and Entrant Bidding in Scoring Rule Auctions: A Study on Italian Canteen Services. *The B.E. Journal of Economic Analysis & Policy*, *21*(1), 1-43. https://doi.org/10.1515/bejeap-2019-0146

Chadwick, L. (2024). What is behind the Malta hospital scandal that has led to charges against former PM Joseph Muscat? . *Euronews*. https://www.euronews.com/health/2024/05/09/what-is-behind-the-malta-hospital-scandal-that-has-led-to-charges-against-former-pm-joseph

Decarolis, F., & Giorgiantonio, C. (2022). Corruption red flags in public procurement: new evidence from Italian calls for tenders. *EPJ Data Science*, *11*(1), 16. https://doi.org/10.1140/epjds/s13688-022-00325-x

Dotoli, M., Epicoco, N., & Falagario, M. (2020). Multi-Criteria Decision Making techniques for the management of public procurement tenders: A case study. *Applied Soft Computing*, *88*, 106064. https://doi.org/10.1016/j.asoc.2020.106064

European Commission. (n.d.). *Public procurement*. Retrieved from https://single-market-economy.ec.europa.eu/single-market/public-procurement_en

European Public Prosecutor's Office. (2024a). *Czechia: Ten arrested in probe into corruption ring involving medical supplies to hospitals* https://www.eppo.europa.eu/en/media/news/czechia-ten-arrested-probe-corruption-ring-involving-medical-supplies-to-hospitals

European Public Prosecutor's Office. (2024b). *Romania: EPPO conducts searches in probe into €8 million fraud involving Covid-19 medical equipment* https://www.eppo.europa.eu/en/media/news/romania-eppo-conducts-searches-probe-eu8-million-fraud-involving-covid-19-medical

European Union. (n.d.). *Threshold*. Retrieved from https://single-market-economy.ec.europa.eu/single-market/public-procurement/legal-rules-and-implementation/thresholds_en

Fazekas, M., & Kocsis, G. (2020). Uncovering High-Level Corruption: Cross-National Objective Corruption Risk Indicators Using Public Procurement Data. *British Journal of Political Science*, *50*. https://doi.org/10.1017/S0007123417000461

Fazekas, M., Tóth, I. J., & King, L. P. (2016). An Objective Corruption Risk Index Using Public Procurement Data. *European Journal on Criminal Policy and Research*, *22*(3), 369-397. https://doi.org/10.1007/s10610-016-9308-z

Ferwerda, J., Deleanu, I., & Unger, B. (2017). Corruption in Public Procurement: Finding the Right Indicators. *European Journal on Criminal Policy and Research*, *23*(2), 245-267. https://doi.org/10.1007/s10610-016-9312-3

García Rodríguez, M. J., Rodríguez-Montequín, V., Ballesteros-Pérez, P., Love, P. E. D., & Signor, R. (2022). Collusion detection in public procurement auctions with machine learning algorithms. *Automation in Construction*, *133*, 104047. https://doi.org/10.1016/j.autcon.2021.104047

Github. (n.d.). *Random Forest Regression*. https://apple.github.io/turicreate/docs/userguide/supervised-learning/random_forest_regression.html

Goutte, C., & Gaussier, E. (2005). A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. Advances in Information Retrieval, Berlin, Heidelberg.

Hanafy, M., & Ming, R. (2021). Machine Learning Approaches for Auto Insurance Big Data. *Risks*, *9*(2), 42. https://www.mdpi.com/2227-9091/9/2/42

Heggstad, K. K., & Froystad, M. (2011). The basics of integrity in procurement. *Chr. Michelsen Institute*. https://open.cmi.no/cmi-xmlui/handle/11250/2475082

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B. (2018). Random

forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*. https://doi.org/10.7717/peerj.5518
Hume, C., & Ferda, G. (2024). NHS checked 'fake' nurse qualifications, jury told. *BBC*. https://www.bbc.com/news/articles/c0xxl2n8vwgo
Huque, M. H., Carlin, J. B., Simpson, J. A., & Lee, K. J. (2018). A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC Medical Research Methodology*, *18*(1), 168. https://doi.org/10.1186/s12874-018-0615-6
Ibem, E. O., & Laryea, S. (2015). e-Procurement use in the South African construction industry. *Journal of Information Technology in Construction (ITcon)*, *20*(23), 364-384. https://www.itcon.org/2015/23
IBM. (n.d.). *What is overfitting?* https://www.ibm.com/topics/overfitting
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to Statistical Learning*. https://doi.org/10.1007/978-1-0716-1418-1
Kautz, T., Eskofier, B. M., & Pasluosta, C. F. (2017). Generic performance measure for multiclass-classifiers. *Pattern Recognition*, *68*, 111-125. https://doi.org/10.1016/j.patcog.2017.03.008
Kenny, C., & Musatova, M. (2010). Red Flags of Corruption in World Bank Projects: An Analysis of Infrastructure Contracts (English). *Policy Research working paper*. http://documents.worldbank.org/curated/en/790591468321562564/Red-flags-of-corruption-in-world-bank-projects-an-analysis-of-infrastructure-contracts
Klaassen, N. R., Cyril. (2022). Onderzoek Sywert-deal: Van Lienden mocht winst maken, maar hij misleidde de overheid. *AD*. https://www.ad.nl/binnenland/onderzoek-sywert-deal-van-lienden-mocht-winst-maken-maar-hij-misleidde-de-overheid~ad18823e/
Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human Decisions and Machine Predictions*. *The Quarterly Journal of Economics*, *133*(1), 237-293. https://doi.org/10.1093/qje/qjx032
Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. https://doi.org/10.1007/978-1-4614-6849-3
Lima, M. S. M., & Delen, D. (2020). Predicting and explaining corruption across countries: A machine learning approach. *Government Information Quarterly*, *37*(1), 101407. https://doi.org/10.1016/j.giq.2019.101407
Lima, W., Lira, R., Paiva, A., Silva, J., & Silva, V. (2023). Methodology for automatic extraction of red flags in public procurement. 2023 International Joint Conference on Neural Networks (IJCNN),
Liu, C., Chan, Y.-C., Hasnain Alam, S., & Fu, H. (2015). Financial Fraud Detection Model: Based on Random Forest. *Econometrics: Econometric Model Construction*. https://www.semanticscholar.org/paper/Financial-Fraud-Detection-Model%3A-Based-on-Random-Liu-Chan/4734bb0c53378c9c59f2967ad071f466b7413162
Lyra, M. S., Damásio, B., Pinheiro, F. L., & Bacao, F. (2022). Fraud, corruption, and collusion in public procurement activities, a systematic literature review on data-driven methods. *Applied Network Science*, *7*(1), 83. https://doi.org/10.1007/s41109-022-00523-6
Mackenzie, A. (2015). The production of prediction: What does machine learning want? *European Journal of Cultural Studies*, *18*(4-5), 429-445. https://doi.org/10.1177/1367549415577384
Mackey, T. K., & Cuomo, R. E. (2020). An interdisciplinary review of digital technologies to facilitate anti-corruption, transparency and accountability in medicines procurement. *Global Health Action*, *13*(sup1), 1695241. https://doi.org/10.1080/16549716.2019.1695241
Mélon, L., & Spruk, R. (2020). The impact of e-procurement on institutional quality. *Journal*

of Public Procurement, *20*(4). https://doi.org/10.1108/JOPP-07-2019-0050
Mohemad, R., Hamdan, A. R., Othman, Z. A., & Noor, N. M. M. (2011). Modelling ontology for supporting construction tender evaluation process. 2011 International Conference on Semantic Technology and Information Retrieval,
Murel, J., & Kavlakoglu, E. (2024). *What is a confusion matrix?* https://www.ibm.com/think/topics/confusion-matrix
NOS. (2016). Onderzoek naar fraude met pgb-geld en witwassen in Emmen. *NOS*. https://nos.nl/artikel/2137498-onderzoek-naar-fraude-met-pgb-geld-en-witwassen-in-emmen
NOS. (2022). Mogelijke corruptie in Zwolle Isala-ziekenhuis groter dan gedacht. *NOS*. https://nos.nl/artikel/2434665-mogelijke-corruptie-in-zwols-isala-ziekenhuis-groter-dan-gedacht
OECD. (2016). *Preventing corruption in public procurement*. https://www.oecd-ilibrary.org/docserver/9789264061651-31-en.pdf?expires=1731960610&id=id&accname=id11881&checksum=D6AD2834E7ECFCF8607674796E5555B6
OECD. (2023). *Government at a glance 2023*. Retrieved from https://www.oecd.org/en/publications/government-at-a-glance-2023_3d5c5d31-en.html
OLAF. (2022). *The OLAF report 2021*. Retrieved from https://anti-fraud.ec.europa.eu/system/files/2022-09/olaf-report-2021_en.pdf
PIANOO. (n.d.). *De ruimte van de percelenregeling*. https://www.pianoo.nl/nl/document/16910/de-ruimte-van-de-percelenregeling
Priya, G. J., & Saradha, S. (2021). Fraud Detection and Prevention Using Machine Learning Algorithms: A Review. 2021 7th International Conference on Electrical Energy Systems (ICEES),
Pugliese, R., Regondi, S., & Marini, R. (2021). Machine learning-based approach: global trends, research directions, and regulatory standpoints. *Data Science and Management*, *4*, 19-29. https://doi.org/10.1016/j.dsm.2021.12.002
Puterman, E., Weiss, J., Hives, B. A., Gemmill, A., Karasek, D., Mendes, W. B., & Rehkopf, D. H. (2020). Predicting mortality from 57 economic, behavioral, social, and psychological factors. *Proceedings of the National Academy of Sciences*, *117*(28), 16273-16282. https://doi.org/10.1073/pnas.1918455117
Quantstart. (n.d.). *Bootstrap Aggregation, Random Forest and Boosted Trees*. https://www.quantstart.com/articles/bootstrap-aggregation-random-forests-and-boosted-trees/
Rigatti, S. J. (2017). Random Forest. *J Insur Med*, *47*(1), 31-39. https://doi.org/10.17849/insm-47-01-31-39.1
Roderick, J. A., & Little, D. B. R. (2002). *Statistical Analysis with Missing Data*. https://doi.org/10.1002/9781119013563
Ross, T. W., & Yan, J. (2015). Comparing Public–Private Partnerships and Traditional Public Procurement: Efficiency vs. Flexibility. *Journal of Comparative Policy Analysis: Research and Practice*, *17*(5), 448-466. https://doi.org/10.1080/13876988.2015.1029333
Sarica, A., Cerasa, A., & Quattrone, A. (2017). Random Forest algorithm for the classification of neuroimaging Data in Alzheimer's Disease: A Systematic Review. *Front. Aging Neurosci*. https://doi.org/10.3389/fnagi.2017.00329
Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526-534. https://doi.org/10.1016/j.procs.2021.01.199

Schulz, M., & Kourkoulas, D. (2023). *Directive 2014/24/EU of the European Parliament and of the Council of 26 February 2014 on public procurement and repealing Directive 2004/18/EC Text with EEA relevance*. Official Journal of the European Union Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32014L0024
Schwarz-Goerlich, A., Huebner, A., & Revill, J. (2023). Siemens cooperating with Austrian authorities in corruption probe. https://www.reuters.com/legal/siemens-cooperating-with-austrian-authorities-corruption-probe-2023-08-08/
Seri, P., Bianchi, A., & Matteucci, N. (2014). Diffusion and usage of public e-services in Europe: An assessment of country level indicators and drivers. *Telecommunications Policy*, *38*(5), 496-513. https://doi.org/10.1016/j.telpol.2014.03.004
Sisodia, D., & Sisodia, D. S. (2018). Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science*, *132*, 1578-1585. https://doi.org/10.1016/j.procs.2018.05.122
Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. AI 2006: Advances in Artificial Intelligence, Berlin, Heidelberg.
Spyromitros, E., & Panagiotidis, M. (2022). The impact of corruption on economic growth in developing countries and a comparative analysis of corruption measurement indicators. *Cogent Economics & Finance*, *10*(1), 2129368. https://doi.org/10.1080/23322039.2022.2129368
Swift, A., Heale, R., & Twycross, A. (2020). What are sensitivity and specificity? *Evid Based Nurs*, *23*(1), 2-4. https://doi.org/10.1136/ebnurs-2019-103225
Tátrai, T., & Németh, A. (2018). Improving red flag instruments for public procurement. *ERA Forum*, *19*(2), 267-285. https://doi.org/10.1007/s12027-018-0513-8
Thomann, E., Marconi, F., & Zhelyazkova, A. (2024). Did pandemic responses trigger corruption in public procurement? Comparing Italy and Germany. *Journal of European Public Policy*, *31*(9), 2907-2936. https://doi.org/10.1080/13501763.2023.2241879
Titl, V., Mazrekaj, D., & Schiltz, F. (2024). Identifying Politically Connected Firms: A Machine Learning Approach. *Oxford Bulletin of Economics and Statistics*, *86*(1), 137-155. https://doi.org/10.1111/obes.12586
Transparency International. (2023a). *Corruption Perceptions Index* (https://www.transparency.org/en/cpi/2023
Transparency International. (2023b). *CPI 2022 for Western Europe & EU: Undue influence and fragmented anti-corruption measures hurt progress*. https://www.transparency.org/en/news/cpi-2022-western-europe-eu-corruption-undue-influence-hurt-progress
Vian, T. (2008). Review of corruption in the health sector: theory, methods and interventions. *Health Policy and Planning*, *23*(2), 83-94. https://doi.org/10.1093/heapol/czm048
Wang, H., & Zheng, H. (2013). Positive Predictive Value. In W. Dubitzky, O. Wolkenhauer, K.-H. Cho, & H. Yokota (Eds.), *Encyclopedia of Systems Biology* (pp. 1723-1724). Springer New York. https://doi.org/10.1007/978-1-4419-9863-7_256
Wensink, W., & de Vet, J. M. (2013). *Identifying and Reducing Corruption in Public Procurement in the EU*. https://www.pianoo.nl/sites/default/files/documents/documents/identifyingreducingcorruptioninpublicprocurementen.pdf
Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S., & Jiang, C. (2018). Random forest for credit card fraud detection. 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC),

6) Appendices

6.1 Appendix I – Red flags

This table shows all the red flags from different studies and indices mentioned in this section.

Phase	Red flag	Study/ Index
Submission	Call for tender not published in official journal	(Fazekas & Kocsis, 2020)
	Procedure Type	
	Length of submission period	
Assessment	Weight of non-price evaluation criteria	
	Length of decision period	
Outcome	Single bidder contract	
Bidding process	Writing tender specifications in a way that favours a particular supplier	Australian government
	Frequent use exemptions to circumvent competitive procurement	
	Frequently extending contracts	
	Accepting late or suspicious bids	
	Not declaring connections with the bidder, as well as actual or perceived conflicts of interest	
	Accepting offers of gifts, benefits or hospitality	
	Releasing sensitive information to a particular bidder	
	Submitting suspicious bids to give the appearance of competition	
	Not declaring connections with another bidder	
	Submitting bids that vary significantly from others	
	Offering gifts, benefits or hospitality	
Selecting preferred supplier	Selecting a supplier who bid well above the expected cost	
	Repeatedly awarding contracts to the same suppliers	
	Not declaring a conflict of interest in the tender/ quote evaluation process	
	Favouring getting procurement done quickly over following the proper process	
	Not maintaining appropriate paperwork	
	Not submitting the appropriate paperwork to support the approval of the supplier	
	Undertaking all duties by themselves	
	Hiring a losing bidder as a sub-contractors	
	Complaining about the selection process of lack of competition	
Payment	Raising purchase orders after invoices have been received	
	Submitting false, inflated or duplicate invoices, or sloppy invoices with insufficient detail or obvious mistakes	

Delivery	Not receiving goods and services Receiving poor quality goods and services Receiving community/ staff complaints about the quality of goods and services	
Contract management	Splitting contracts to avoid the need for a certain number of quotes or tender process, or to keep purchases within a particular financial delegation Varying contracts	
	Single bidder contracts Non-open procedures Lack of publication of call for tenders Period for submitting bids Period for selecting the winning bid Spending concentration Share of suppliers registered in jurisdictions offering limited company and banking transparency	(Abdou et al., 2022)
	Absence of tender call Call for tenders: page and word number ANAC info available Negotiated procedures Legality protocols Local regulations Design-build Scoring rule (MEAT) Price Only No possibility of single source award Preferred firm indications Open tender days Document verification Worksite verification Ad hoc rules for subcontracting Prohibition of pooling agreements Multiple contact points External contact points	(Decarolis & Giorgiantonio, 2022)
	Contract office not subordinated to tender provider The shortened time for the bidding process Tender exceptionally large Complaints from non-winning bidders Substantial changes in project scope/costs after award Connections between bidders undermine competition	(Ferwerda et al., 2017)

	Awarding authority did not fill in all fields in TED/CAN Amount of missing information	
	Strong inertia in the composition of the evaluation team Conflict of interest members of the evaluation team Multiple contact points Contact office not subordinated to tender provider Contact person not employed by tender provider Preferred supplier indications The shortened time for the bidding process Accelerated tender Tender exceptionally large Time-to-bid does not conform to the law Bids after the deadline accepted Number of offers Artificial bids Complaints from non-winning bidders The award contract has new bid specifications Substantial changes in project scope/costs after award Connections between bidders undermine competition All bids higher than projected overall costs Not all bidders were informed of the award and its reasons Award contracts and selection documents are not all public Inconsistencies in reported turnover/number of staff Winning company not listed in the Chamber of Commerce % of EU funding % of public funding from MS Awarding authority did not fill in all fields in TED/CAN Audit certificates by auditor without credentials Negative media coverage	OLAF
Advertising/Bid opening	The time between advertising of the contract and bid opening (weeks) The time between bid opening and bid evaluation Number of submitted bids The ratio of submitted bids to the number of companies that bought bidding documents (%)	World Bank
Bid evaluations/Contract award	Time between bid award and actual contract signing date The ratio of non-responsive bidders to all bidders Was the lowest bidder considered non-responsive?	

	For ICB contracts: did international companies bid in the auction?	
	If the winner is the lowest bidder, what is the per cent gap between 1st and 2nd bid quotes?	
	Were any two bids submitted within 15 (Rigid Threshold) or 2,5% (Soft Threshold) of each other?	
	Difference between contract estimate and winning bid	
	Difference between contract award and final contract amount	
	Thresholds for procurement methods and prior review	
Pre-tendering phase	Needs assessment and market analysis	OECD
	Planning and budgeting	
	Development of specifications/ requirements	
	Choice of the procurement procedure	
Tender phase	Request for proposal/ bid	
	Bid submission	
	Bid evaluation	
	Contract award	
Post-award phase	Contract management/ performance	
	Order and payment	
Contract notice	The contracting authority has been convicted in a final judgment or has a bad reputation	Transparency International
	Framework agreement with one tenderer	
	Framework agreement with several tenderers (fewer than three tenderers participating)	
	Term of the framework agreement (long)	
	Total estimated value of framework agreement (high)	
	The object of public procurement (cartel risk)	
	High estimated value (contract of outstanding value)	
	Amounts overly uncertain (great difference allowed)	
	The contract can be renewed (several times, or for a longer time, or without any information)	
	Term of the contract (long or indefinite)	
	Omission of the definition of compulsory grounds of exclusion	
	Economic and financial ability—no minimum requirements	
	Economic and financial ability—conditions for capital (levels)	
	Economic and financial ability—required sales revenues > estimated value	

	Economic and financial ability—statement of sales revenues (period)
	Technical capacity—no minimum requirement defined
	Technical capacity—reference value > estimated value
	Technical capacity—period of reference requirement
	Technical capacity—a requirement of a reference performance under one contract
	Technical capacity—requirement for references co-financed by the EU
	Technical capacity—setting geographical requirements
	Technical capacity—experience of experts involved (number of years)
	Accelerated procedure (use, and/or without statement of reasons)
	Competitive (negotiated) procedure as legal grounds
	The actual or predefined number of candidates is low
	No criteria specified for the limitation of participant numbers
	Award criterion—definition is incomplete (no constituent factor or at least two constituent factors; no method defined)
	Award criterion—payment deadline
	The time limit for tendering/participation is short
	Opening date of tenders (differs from the time limit for tendering/participation)
	Tender guarantee (amount)
Contract award notice	Procedures without prior publication
	Number of tenders received (low)
	Winning economic actor(s)—related information
	The ratio of the total final value and the estimated value
	Unsuccessful procedure for risky reasons
	Unsuccessful procedure without a statement of reason
	Successful procedure without contracting
	Duration of evaluation (long)

6.2 Appendix II – Flags used in model

Green = information on TenderNed available for this red flag

Orange = previously mentioned as available information on TenderNed

Red = no information available for this red flag on TenderNed

Red Flag	Definition	TenderNed Characterstic
Call for tender not published in official journal	Published or not published	Publicatie ID
Non-open procedure types	Using less open transparent procedure types such as open and invitation tenders. Non-open tenders accelerated, restricted, awarded without publication, negotiated, tender without competition	Procedure or publicatie soort or publicatie type
Too-short of an advertisement period	Number of days between publishing a tender and the submission deadline	Publicatiedatum and sluitingsdatum aanmelding or sluitingsdatum aanbesteding
Hard-to-quantify evaluation criteria	Sum of weights for evaluation criteria which are not related to price	Hoofdgunningscriterium
Excessively short or lengthy time used to decide on the submitted bids	Number of days between the submission deadline and announcing the contract	Sluitingsdatum aanmelding or sluitingsdatum aanbesteding and datum gunning
Single-bidder contracts	Contracts awarded in procurement tenders in which only one bid was submitted	Aantal inschrijvingen and aantal elektronische inschrijvingen
Writing tender specifications in a way that favours a particular supplier		
Frequent use exemptions to circumvent competitive procurement		
Frequently extending contracts		
Accepting late or suspicious bids		
Not declaring connections with the bidder, as well as actual or perceived conflicts of interest		
Accepting offers of gifts, benefits or hospitality		
Releasing sensitive information to a particular bidder		

Submitting suspicious bids to give the appearance of competition		
Not declaring connections with another bidder		
Submitting bids that vary significantly from others		
Offering gifts, benefits or hospitality		
Selecting a supplier who bid well above the expected cost		Oorspronkelijk geraamde waarde and waarde
Repeatedly awarding contracts to the same suppliers		Naam aanbestedende dienst, and officiele benaming
Not declaring a conflict of interest in the tender/ quote evaluation process		
Favouring getting procurement done quickly over following the proper process		Sluitingsdatum aanmelding and sluitingsdatum aanbesteding
Not maintaining appropriate paperwork		
Not submitting the appropriate paperwork to support the approval of the supplier		
Undertaking all duties by themselves		
Hiring a losing bidder as a sub-contractor		
Complaining about the selection process of lack of competition		
Raising purchase orders after invoices have been received		
Submitting false, inflated or duplicate invoices, or sloppy invoices with insufficient detail or obvious mistakes		
Not receiving goods and services		
Receiving poor quality goods and services		
Receiving community/ staff complaints about the quality of goods and services		
Splitting contracts to avoid the need for a certain number of quotes or tender process, or to keep purchases within a particular financial delegation		
Varying contracts		

Single bidder contracts		see page 1
Non-open procedures		see page 1
Lack of publication of call for tenders		see page 1
Period for submitting bids		see page 1
Period for selecting the winning bid		see page 1
Spending concentration		
Share of suppliers registered in jurisdictions offering limited company and banking transparency		
Absence of tender call		
Call for tenders (page and word count)	Number of pages and words of the call for tender main document	
ANAC info available	Anticorruption Authority	
Negotiated procedures	Indicator for whether the procedure is negotiated or not	
Legality protocols	Indicator for whether the call requests bidders to adhere to any legality protocol or not	Juridisch kader
Local regulations	Indicator for whether the call requests bidders to adhere to any local	
regulation		
Design-build	Indicator for whether the contract involves both design and build or only build	
Scoring rule (MEAT)	Indicator for whether the award criterion entails multiple parameters or not	
Price Only	Indicator for price-only criterion & automatic exclusion of abnormally low bids or not	
Preferred firm indications	Indicator for preferences for firms enrolled in the buyer’s preferred	
Suppliers list or not		
Open tender days	Number of days between when the call is published and when it closes	see page 1
Document verification	Indicator for compulsory verification of inspection	

	of the project documents or not	
Worksite verification	Indicator for compulsory verification of inspection of the project worksite or not	
Ad hoc rules for subcontracting	Indicator for whether the call contains ad hoc rules for subcontracting or not	
Prohibition of pooling agreements		
Multiple and external contact points	Indicator for contact point personnel outside the employees of the public buyer or not	
Contract office not subordinated to tender provider		
The shortened time for the bidding process		see page 1
Tender exceptionally large		Waarde
Complaints from non-winning bidders		
Substantial changes in project scope/costs after award		
Connections between bidders undermine competition		
Awarding authority did not fill in all fields in TED/CAN		
Amount of missing information		Empty cells
Strong inertia in the composition of the evaluation team		
Conflict of interest members of the evaluation team		
Multiple contact points		
Contact office not subordinated to tender provider		
Contact person not employed by tender provider		
Preferred supplier indications		
The shortened time for the bidding process		see page 1
Accelerated tender		see page 1
Tender exceptionally large		see page 6
Time-to-bid does not conform to the law		
Bids after the deadline accepted		
Number of offers		see page 1

Artificial bids		
Complaints from non-winning bidders		
The award contract has new bid specifications		
Substantial changes in project scope/costs after award		
Connections between bidders undermine competition		
All bids higher than projected overall costs		see page 4
Not all bidders were informed of the award and its reasons		
Award contracts and selection documents are not all public		see page 1
Inconsistencies in reported turnover/number of staff		
Winning company not listed in the Chamber of Commerce		
% of EU funding		
% of public funding from MS		
Awarding authority did not fill in all fields in TED/CAN		
Audit certificates by auditor without credentials		
Negative media coverage		
The time between advertising of the contract and bid opening (weeks)		
The time between bid opening and bid evaluation		publicatiedatum and datum gunning
Number of submitted bids		see page 1
The ratio of submitted bids to the number of companies that bought bidding documents (%)		
Time between bid award and actual contract signing date		datum gunning and aanvang opdracht
The ratio of non-responsive bidders to all bidders		
Was the lowest bidder considered non-responsive?		
For ICB contracts	did international companies bid in the auction?	
If the winner is the lowest bidder, what is the per cent gap between 1st and 2nd bid quotes?		

Were any two bids submitted within 15 (Rigid Threshold) or 2,5% (Soft Threshold) of each other?		
Difference between contract estimate and winning bid		see page 2
Difference between contract award and final contract amount		
Thresholds for procurement methods and prior review		
Needs assessment and market analysis		
Planning and budgeting		
Development of specifications/ requirements		
Choice of the procurement procedure		Procedure
Request for proposal/ bid		
Bid submission		see page 1
Bid evaluation		
Contract award		see page 6
Contract management/ performance		
Order and payment		
The contracting authority has been convicted in a final judgment or has a bad reputation		
Framework agreement with one tenderer		
Framework agreement with several tenderers (fewer than three tenderers participating)		
Term of the framework agreement (long)		
Total estimated value of framework agreement (high)		
The object of public procurement (cartel risk)		Publicatie soort
High estimated value (contract of outstanding value)		see page 6
Amounts overly uncertain (great difference allowed)		
The contract can be renewed (several times, or for a longer time, or without any information)		

Term of the contract (long or indefinite)		
Omission of the definition of compulsory grounds of exclusion		
Economic and financial ability—no minimum requirements		
Economic and financial ability—conditions for capital (levels)		
Economic and financial ability—required sales revenues > estimated value		
Economic and financial ability—statement of sales revenues (period)		
Technical capacity—no minimum requirement defined		
Technical capacity—reference value > estimated value		
Technical capacity—period of reference requirement		
Technical capacity—a requirement of a reference performance under one contract		
Technical capacity—requirement for references co-financed by the EU		
Technical capacity—setting geographical requirements		
Technical capacity—experience of experts involved (number of years)		
Accelerated procedure (use, and/or without statement of reasons)		see page 1
Competitive (negotiated) procedure as legal grounds		
The actual or predefined number of candidates is low		page 1
No criteria specified for the limitation of participant numbers		
Award criterion—definition is incomplete (no constituent factor or at least two constituent factors		no method defined)
Award criterion—payment deadline		

The time limit for tendering/participation is short		see page 1
Opening date of tenders (differs from the time limit for tendering/participation)		
Tender guarantee (amount)		
Procedures without prior publication		
Number of tenders received (low)		see page 1
Winning economic actor(s)—related information		
The ratio of the total final value and the estimated value		see page 2
Unsuccessful procedure for risky reasons		
Unsuccessful procedure without a statement of reason		
Successful procedure without contracting		
Duration of evaluation (long)		see page 1
The final value of the contract is too high		see page 6

6.3 Appendix III – R-code Random Forest

```
install.packages("readxl")
install.packages("magrittr")
install.packages("dplyr")
install.packages("tidyr")
install.packages("ggplot2")
install.packages("tidyverse")
install.packages("readr")
install.packages("lubridate")
install.packages("class")
install.packages("caret")
install.packages("lattice")
install.packages("MASS")
install.packages("VIM")
install.packages("data.table")
install.packages("caTools")
install.packages("randomForest")
install.packages("datasets")
install.packages("boot")
install.packages("rfPermute")
install.packages("permute")
install.packages("kableExtra")
install.packages("psych")
install.packages("mice")

library(readxl)
library(magrittr)
library(dplyr)
library(tidyr)
library(ggplot2)
library(tidyverse)
library(readr)
library(lubridate)
library(class)
library(caret)
library(lattice)
library(MASS)
library(VIM)
library(data.table)
library(caTools)
library(randomForest)
library(datasets)
library(boot)
library(rfPermute)
library(permute)
library(kableExtra)
library(psych)
library(mice)

#original data
data_org<-read_xlsx("Dataset.xlsx")

#Removing not applicable columns
```

```r
data_mod<-data_org %>%
  dplyr::select(Publicatiedatum, `Naam Aanbestedende dienst`, `Sluitingsdatum aanmelding`, `Sluitingsdatum aanbesteding`, `Aanvang opdracht`, `Juridisch kader`,`Type opdracht`, Procedure, Hoofdactiviteit, Hoofdgunningscriterium, `Datum gunning`, `Aantal inschrijvingen`, `Aantal elektronische inschrijvingen`, `Officiele benaming`, `Oorspronkelijk geraamde waarde - bedrag`, `Waarde - bedrag`)

#original data filtered on health, date, opdracht, and threshold
data_mod<-data_mod %>%
  filter(Hoofdactiviteit == "Gezondheid") %>%
  filter(Publicatiedatum<'2024-01-01') %>%
  filter(`Type opdracht` == "Leveringen" | `Type opdracht` == "Diensten")
%>%
  filter(`Oorspronkelijk geraamde waarde - bedrag` >= 443000)


#Convert chr into date
data_mod<-data_mod%>%
  mutate(Publicatiedatum = as.Date(as.character(as.POSIXct(Publicatiedatum
))))
data_mod<- data_mod %>%
  mutate(`Sluitingsdatum aanbesteding` = as.Date(`Sluitingsdatum aanbesteding`, format = "%d-%m-%Y"))
data_mod <- data_mod %>%
  mutate(`Sluitingsdatum aanmelding` = as.Date(`Sluitingsdatum aanmelding`
, format = "%d-%m-%Y"))
data_mod$`Datum gunning`<-as.Date(data_mod$`Datum gunning`, format = "%d-%m-%Y")
data_mod$`Aanvang opdracht`<-as.Date(data_mod$`Aanvang opdracht`, format =
"%d-%m-%Y")
data_mod$`Waarde - bedrag`<-as.numeric(data_mod$`Waarde - bedrag`)

#NA in dataframe
na_counts<-colSums(is.na(data_mod))
print(na_counts)
sum(is.na(data_mod))/prod(dim(data_mod))
sum(is.na(data_mod$`Sluitingsdatum aanmelding`))/prod(dim(data_mod))
sum(is.na(data_mod$`Aanvang opdracht`))/prod(dim(data_mod))

#removing necessary columns
data_mod=subset(data_mod,select = -c(3,7,9))
####################################################################

#Descriptive statistics original data summary(data_mod)
data_mod<-data_mod %>%
  mutate("Naam Aanbestedende dienst" = as.numeric(factor(`Naam Aanbestedende dienst`)))
freq_table<-table(data_mod$`Naam Aanbestedende dienst`)
view(freq_table)

data_mod<-data_mod %>%
  mutate("Officiele benaming" = as.numeric(factor(`Officiele benaming`)))
freq_table<-table(data_mod$`Officiele benaming`)
```

```
view(freq_table)

table(data_mod$Procedure)
table(data_mod$Hoofdgunningscriterium)
#####################################################################
#MICE for predicting missing data, because Random Forest doesn't work with
missing data
md.pattern(data_mod)

colnames(data_mod)<-c("Column1","Column2","Column3","Column4","Column5", "
Column6","Column7","Column8",
                      "Column9","Column10", "Column11","Column12","Column1
3")
data_mod$Column2 <- as.factor(data_mod$Column2)
data_mod$Column5 <- as.factor(data_mod$Column5)
data_mod$Column6 <- as.factor(data_mod$Column6)
data_mod$Column7 <- as.factor(data_mod$Column7)
data_mod$Column11 <- as.factor(data_mod$Column11)

data_mod <- data_mod %>%
  mutate(Column1 = as.numeric(Column1))
data_mod <- data_mod %>%
  mutate(Column3 = as.numeric(Column3))
data_mod <- data_mod %>%
  mutate(Column4 = as.numeric(Column4))
data_mod <- data_mod %>%
  mutate(Column8 = as.numeric(Column8))

imp_data<-mice(data_mod, m=5, method = "rf")

data_final<-complete(imp_data)
data_final <- data_final %>%
  mutate(Column1 = as.Date(Column1, origin = "1970-01-01"),
         Column3 = as.Date(Column3, origin = "1970-01-01"),
         Column4 = as.Date(Column4, origin = "1970-01-01"),
         Column8 = as.Date(Column8, origin = "1970-01-01"))

colnames(data_final)<-c("Publicatiedatum","Naam Aanbestedende dienst","Slu
itingsdatum aanbesteding",
                        "Aanvang opdracht","Juridisch kader","Procedure","
Hoofdgunningscriterium",
                        "Datum gunning ","Aantal inschrijvingen","Aantal e
lektronische inschrijvingen",
                        "Officiele benaming","Oorspronkelijk geraamde waar
de - bedrag", "Waarde - bedrag")

na_counts<-colSums(is.na(data_final))
print(na_counts)

#####################################################################
#Red Flag 1
data_final <- data_final %>%
  mutate(RF1 = case_when(
  grepl("Mededingingsprocedure met onderhandeling|Niet-openbaar|Onderhande
```

```
ling zonder bekendmaking",
        Procedure, ignore.case = TRUE) ~ 1,
  Procedure %in% c("Concurrentiegerichte dialoog", "Innovatiepartnerschap
", "Openbaar") ~ 0,))

#Red Flag 2
data_final<-data_final%>%
  mutate("Advertisement period" = `Sluitingsdatum aanbesteding` - Publicat
iedatum)

data_final<-data_final%>%
  mutate(RF2 = case_when(
    Procedure == "Concurrentiegerichte dialoog" & `Advertisement period`<3
0~1,
    Procedure == "Innovatiepartnerschap" & `Advertisement period`<30~1,
    Procedure == "Mededingingsprocedure met onderhandeling" & `Advertiseme
nt period`<30~1,
    Procedure == "Niet-openbaar" & `Advertisement period`<30~1,
    Procedure == "Onderhandeling zonder bekendmaking" & `Advertisement per
iod`<30~1,
    Procedure == "Openbaar" & `Advertisement period`<45~1,
    TRUE ~ 0))

#Red Flag 3
data_final <- data_final %>%
  mutate(RF3 = case_when(
    grepl("Beste prijs-kwaliteit verhouding|Laagste prijs", Hoofdgunningsc
riterium, ignore.case = TRUE) ~ 0,
    Hoofdgunningscriterium %in% c(NA) ~ 1,))

#Red Flag 4
data_final<-data_final%>%
  mutate("Evaluation period" = `Datum gunning `- `Sluitingsdatum aanbested
ing`)
data_final$RF4<-ifelse(data_final$`Evaluation period`>30,1,0)

#Red Flag 5
data_final<-data_final %>%
  mutate("Total bidders" = `Aantal inschrijvingen` + `Aantal elektronische
inschrijvingen`)
data_final$RF5<-ifelse(data_final$`Total bidders`<2,1,0)

#Red Flag 6
data_final$`Waarde - bedrag`<-as.numeric(data_final$`Waarde - bedrag`)
data_final<-data_final %>%
  mutate("Cost limit" = `Oorspronkelijk geraamde waarde - bedrag` * 1.5)
data_final$RF6<-ifelse(data_final$`Cost limit`<data_final$`Waarde - bedrag
`,1,0)

#Red Flag 7
data_final<-data_final %>%
  mutate("Naam Aanbestedende dienst" = as.numeric(factor(`Naam Aanbesteden
de dienst`)))
```

```r
data_final<-data_final %>%
  mutate("Officiele benaming" = as.numeric(factor(`Officiele benaming`)))

data_final<-data_final%>%
  arrange(`Naam Aanbestedende dienst`, `Officiele benaming`) %>%
  group_by(`Naam Aanbestedende dienst`, `Officiele benaming`) %>%
  mutate(count=row_number()) %>%
  ungroup()

boxplot(data_final$count)
summary(data_final$count)

#Threshold is set at 4, because upper quarterly of the box plot is 4
data_final<-data_final %>%
  mutate(RF7 = ifelse(count > 4, 1, 0))

#Red Flag 8
#data_final = subset(data_final, select = c(-3))

#Red Flag 9
data_final<-data_final%>%
  mutate(RF9 = case_when(
    `Juridisch kader` == "Aanbestedingswet 2012" ~0,
    TRUE ~1))

#Red Flag 10
mean_value <- mean(data_final$`Waarde - bedrag`, na.rm = TRUE)
sd_value <- sd(data_final$`Waarde - bedrag`, na.rm = TRUE)
result <- mean_value + 2 * sd_value
print(result)

data_final<-data_final %>%
  mutate(RF10 = ifelse(`Waarde - bedrag`>result,1,0))

#Red Flag 11
data_final$RF11 <- apply(data_final, 1, function(row)
  ifelse(any(is.na(row)), 1, 0))

#Red Flag 12
data_final<-data_final%>%
  mutate("Contract execution period" = `Aanvang opdracht` - `Datum gunning
`)

data_final<-data_final %>%
  mutate(RF12 = ifelse(data_final$`Contract execution period`<20,1,0))

#Corruption
data_final<-data_final %>%
  mutate(Corruption = RF1+RF2+RF3+RF4+RF5+RF6+RF7+RF9+RF10+RF11+RF12)

data_final<- data_final%>%
  mutate(Corruption = ifelse(is.na(Corruption), 0, Corruption))
#####################################################################
#Descriptive statistics with imputation
```

```
data_final$`Advertisement period` <- as.numeric(data_final$`Advertisement
period`)
data_final$`Evaluation period` <- as.numeric(data_final$`Evaluation period
`)
data_final$`Contract execution period`<- as.numeric(data_final$`Contract e
xecution period`)
summary(data_final)

#Frequency table procedure RF1
table(data_final$Procedure)

#Advertisement period RF2
median(data_final$`Advertisement period`)

#Frequency table Main Award criterion RF3
table(data_final$Hoofdgunningscriterium)

#Evaluation period RF4
median(data_final$`Evaluation period`)

#Total bidders RF5
median(data_final$`Total bidders`)

#Original estimated value
median(data_final$`Oorspronkelijk geraamde waarde - bedrag`)

#Awarding same supplier RF7
median(data_final$`count`)

#Legal framework RF9
table(data_final$`Juridisch kader`)

#Contract execution period RF12
median(data_final$`Contract execution period`)
###############################################################
#Random Forest
#Version 1 contains dichotomous variables of RF1 t/m R12, and where corrup
tion are dichotomous
version1<-data_final
version1=subset(version1,select = c(14,16,17,19,21,23,25,26,27,28,30,31))

version1<-version1%>%
  mutate(Corruption = case_when(
    Corruption == 0~0,
    Corruption == 1~1,
    Corruption == 2~1,
    Corruption == 3~1,
    Corruption == 4~1,
    Corruption == 5~1,
    Corruption == 6~1))

version1$Corruption<-as.factor(version1$Corruption)

set.seed(123)
```

```r
train_indices1 <- sample(1:nrow(version1), 0.8 * nrow(version1))
train_data1 <- version1[train_indices1, ]
test_data1 <- version1[-train_indices1, ]

rf_model1<-randomForest(Corruption~RF1+RF2+RF3+RF4+RF5+RF6+RF7+
                          RF9+RF10+RF11+RF12, data = train_data1,
                        ntree = 500)
print(rf_model1)

prediction1<-predict(rf_model1, newdata = test_data1, type = 'response')
result1 <- caret::confusionMatrix(prediction1,test_data1$Corruption)
result1

cm1<-result1$table
cm1

metrics1<-as.data.frame(result1$byClass)
colnames(metrics1)<-"metrics"
kable(round(metrics1,4), caption = "F1-score, Precision and Recall ") %>%
  kable_styling(font_size = 16)

#Version 2 contains dichotomous variables of RF1 t/m R12, and where corrup
tion are levels 1 through 6
version2<-data_final
version2=subset(version2,select = c(14,16,17,19,21,23,25,26,27,28,30,31))
version2$Corruption<-as.factor(version2$Corruption)

set.seed(123)
train_indices2 <- sample(1:nrow(version2), 0.8 * nrow(version2))
train_data2 <- version2[train_indices2, ]
test_data2 <- version2[-train_indices2, ]
rf_model2<-randomForest(Corruption~RF1+RF2+RF3+RF4+RF5+RF6+RF7+
                          RF9+RF10+RF11+RF12, data = train_data2, ntree =
500)
print(rf_model2)

prediction2<-predict(rf_model2, newdata = test_data2, type = 'class')
result2 <- caret::confusionMatrix(prediction2,test_data2$Corruption)
result2
cm2<-result2$table
cm2

metrics2 <- as.data.frame(result2$byClass)
if (is.null(colnames(metrics2)) || any(is.na(colnames(metrics2))))
  {
  colnames(metrics2) <- c("F1-score", "Precision", "Recall")
}
kable(round(metrics2, 4), caption = "F1-score, Precision and Recall") %>%
  kable_styling(font_size = 16)

#version 3 contains numerical or categorical variables of red flags, and w
here corruption are dichotomous
version3<-data_final
version3=subset(version3,select = c(5,6,7,12,13,15,18,20,22,24,28,29,31))
```

```r
version3<-version3 %>%
  mutate("Juridisch kader" = case_when(
    `Juridisch kader` == "Aanbestedingswet 2012" ~1,
    `Juridisch kader` == "ARW 2016 - Aanbestedingsreglement Werken 2016"~2
))

version3<-version3 %>%
  mutate(Procedure = case_when(
    Procedure == "Concurrentiegerichte dialoog"~1,
    Procedure == "Innovatiepartnerschap"~2,
    Procedure == "Mededingingsprocedure met onderhandeling"~3,
    Procedure == "Niet-openbaar"~4,
    Procedure == "Openbaar"~5,
    Procedure == "Onderhandeling zonder bekendmaking"~6))

version3<-version3 %>%
  mutate(Hoofdgunningscriterium = case_when(
    Hoofdgunningscriterium == "Beste prijs-kwaliteit verhouding"~1,
    Hoofdgunningscriterium == "Laagste prijs"~2))

version3$`Waarde - bedrag`<-as.numeric(version3$`Waarde - bedrag`)
version3<-version3 %>%
  mutate("Waarde - bedrag" = `Oorspronkelijk geraamde waarde - bedrag` * 1
.5)
version3$`Waarde - bedrag`<-ifelse(version3$`Cost limit`<version3$`Waarde
- bedrag`,1,2)

version3$`Advertisement period`<-as.numeric(version3$`Advertisement period
`)
version3$`Evaluation period`<-as.numeric(version3$`Evaluation period`)
version3$`Contract execution period`<-as.numeric(version3$`Contract execut
ion period`)

version3<-version3%>%
  mutate(Corruption = case_when(
    Corruption == 0~0,
    Corruption == 1~1,
    Corruption == 2~1,
    Corruption == 3~1,
    Corruption == 4~1,
    Corruption == 5~1,
    Corruption == 6~1))

version3=subset(version3,select = -c(4))

version3=subset(version3,select = c(Procedure, `Advertisement period`, Hoo
fdgunningscriterium, `Evaluation period`, `Total bidders`,
                `Cost limit`, count, `Juridisch kader`, `Waarde - bedrag`,
RF11, `Contract execution period`,
                Corruption))

version3$Corruption<-as.factor(version3$Corruption)

set.seed(123)
```

```
train_indices3 <- sample(1:nrow(version3), 0.8 * nrow(version3))
train_data3 <- version3[train_indices3, ]
test_data3 <- version3[-train_indices3, ]

colnames(version3)<-c("RF1","RF2", "RF3", "RF4", "RF5", "RF6", "RF7", "RF9
", "RF10", "RF11","RF12", "Corruption")
colnames(train_data3)<-c("RF1","RF2", "RF3", "RF4", "RF5", "RF6", "RF7", "
RF9", "RF10", "RF11","RF12", "Corruption")
colnames(test_data3)<-c("RF1","RF2", "RF3", "RF4", "RF5", "RF6", "RF7", "R
F9", "RF10", "RF11","RF12", "Corruption")


rf_model3<-randomForest(Corruption~RF1+RF2+RF3+RF4+RF5+RF6+RF7+
                          RF9+RF10+RF11+RF12, data = train_data3,
                        ntree = 500)

print(rf_model3)

prediction3<-predict(rf_model3, newdata = test_data3, type = 'response')
result3 <- caret::confusionMatrix(prediction3,test_data3$Corruption)
result3

cm3<-result3$table
cm3

metrics3<-as.data.frame(result3$byClass)
colnames(metrics3)<-"metrics"
kable(round(metrics3,4), caption = "F1-score, Precision and Recall ") %>%
  kable_styling(font_size = 16)

#version 4 contains numerical or categorical variables of red flags, and w
here corruption are levels 1 through 6
version4<-data_final
version4=subset(version4,select = c(5,6,7,12,13,15,18,20,22,24,28,29,31))
version4<-version4 %>% mutate("Juridisch kader" = case_when(
  `Juridisch kader` == "Aanbestedingswet 2012" ~1,
  `Juridisch kader` == "ARW 2016 - Aanbestedingsreglement Werken 2016"~2))

version4<-version4 %>% mutate(Procedure = case_when(
  Procedure == "Openbaar"~1,
  Procedure == "Innovatiepartnerschap"~2,
  Procedure == "Concurrentiegerichte dialoog"~3,
  Procedure == "Mededingingsprocedure met onderhandeling"~4,
  Procedure == "Onderhandeling zonder bekendmaking"~5,
  Procedure == "Niet-openbaar"~6))

version4<-version4 %>%
  mutate(Hoofdgunningscriterium = case_when(
  Hoofdgunningscriterium == "Beste prijs-kwaliteit verhouding"~1,
  Hoofdgunningscriterium == "Laagste prijs"~2))

version4$`Waarde - bedrag`<-as.numeric(version4$`Waarde - bedrag`)
version4<-version4 %>%
  mutate(Overbidding = `Oorspronkelijk geraamde waarde - bedrag` * 1.5)
```

```
version4$Overbidding<-ifelse(version4$`Cost limit`<version4$`Waarde - bedr
ag`,1,2)

mean_value <- mean(version4$`Waarde - bedrag`, na.rm = TRUE)
sd_value <- sd(version4$`Waarde - bedrag`, na.rm = TRUE)
result <- mean_value + 2 * sd_value
print(result)

version4<-version4 %>%
  mutate("Large tender" = ifelse(`Waarde - bedrag`>result,1,2))

version4$`Advertisement period`<-as.numeric(version4$`Advertisement period
`)
version4$`Evaluation period`<-as.numeric(version4$`Evaluation period`)
version4$`Contract execution period`<-as.numeric(version4$`Contract execut
ion period`)
version4=subset(version4,select = -c(4,5,9))

version4$Corruption<-as.factor(version4$Corruption)
version4<-version4[,c(2,4,3,5,6,11,7,1,12,8,9,10)]

set.seed(123)
train_indices4 <- sample(1:nrow(version4), 0.8 * nrow(version4))
train_data4 <- version4[train_indices4, ]
test_data4 <- version4[-train_indices4, ]

colnames(version4)<-c("RF1","RF2", "RF3", "RF4", "RF5", "RF6", "RF7", "RF9
", "RF10","RF11","RF12", "Corruption")
colnames(train_data4)<-c("RF1","RF2", "RF3", "RF4", "RF5", "RF6", "RF7", "
RF9", "RF10","RF11","RF12", "Corruption")
colnames(test_data4)<-c("RF1","RF2", "RF3", "RF4", "RF5", "RF6", "RF7", "R
F9", "RF10","RF11","RF12", "Corruption")

rf_model4<-randomForest(Corruption~RF1+RF2+RF3+RF4+RF5+RF6+RF7+RF9+RF10++R
F11+RF12,
                        data = train_data4, ntree = 500)
print(rf_model4)

prediction4<-predict(rf_model4, newdata = test_data4, type = 'class')
result4 <- caret::confusionMatrix(prediction4,test_data4$Corruption)
result4
cm4<-result4$table
cm4

metrics4 <- as.data.frame(result4$byClass)
if (is.null(colnames(metrics4)) || any(is.na(colnames(metrics4)))) {
  colnames(metrics4) <- c("F1-score", "Precision", "Recall")
}
kable(round(metrics4, 4), caption = "F1-score, Precision and Recall") %>%
  kable_styling(font_size = 16)

######################################################################
```

```r
#SHAP beeswarm plot
install.packages("shapr")
install.packages("iml")
install.packages("SHAPforxgboost")
library(shapr)
library(iml)
library('SHAPforxgboost')
#Model 2
X2 = as.matrix(version2[,-12])
mod2 = xgboost::xgboost(
  data = X2, label = version2$Corruption, gamma = 0, eta = 1, lambda = 0,
nrounds = 1, verbose = FALSE, nthread = 1)
shap_values2 <- shap.values(xgb_model = mod2, X_train = X2)
shap_values2$mean_shap_score
shap_values_v2 <- shap_values2$shap_score

shap_long_v2 <- shap.prep(xgb_model = mod2, X_train = X2)
shap_long_v2 <- shap.prep(shap_contrib = shap_values_v2, X_train = X2)

shap.plot.summary(shap_long_v2, scientific = TRUE)
shap.plot.summary(shap_long_v2, x_bound = 1.5, dilute = 10)

shap.plot.summary.wrap1(mod2, X = as.matrix(version2[,-12]), top_n = 8)

#Model 4
X4 = as.matrix(version4[,-12])
mod4 = xgboost::xgboost(
  data = X4, label = version4$Corruption, gamma = 0, eta = 1, lambda = 0,
nrounds = 1, verbose = FALSE, nthread = 1)
shap_values4 <- shap.values(xgb_model = mod4, X_train = X4)
shap_values4$mean_shap_score
shap_values_v4 <- shap_values4$shap_score

shap_long_v4 <- shap.prep(xgb_model = mod4, X_train = X4)
shap_long_v4 <- shap.prep(shap_contrib = shap_values_v4, X_train = X4)

shap.plot.summary(shap_long_v4, scientific = TRUE)
shap.plot.summary(shap_long_v4, x_bound = 1.5, dilute = 10)

shap.plot.summary.wrap1(mod4, X = as.matrix(version4[,-12]), top_n = 8)
```