Predicting cost overruns on utility projects with data-driven models

Master Thesis

Name:	G.D. (Rick) Potkamp
Student number:	3045617
Institute:	University of Twente
Faculty:	Engineering Technology
E-mail:	rickpotkamp@outlook.com
Supervisors UT:	dr. ir. Léon olde Scholtenhuis
	dr. ir. Ramon ter Huurne
Supervisor Bam:	ing. Folkert Stijnenbosch
Date:	17-05-2025

UNIVERSITY OF TWENTE.

bam

Table of content

I Summary

The utility industry faces major challenges in the near future. These challenges find their origin in, among others, climate change and urbanisation. Moreover, projects in the underground utility infrastructure sector still face cost overruns. Cost overruns significantly impede the success of a project because they may lead to delays, disagreements and financial losses. At a time like this where there is a lot of work in this sector, it is desirable to create realistic economic expectations. Moreover, the problem is also likely to grow due to that increase in work.

Predicting cost overruns is hard because utility projects are prone to uncertainties. Nowadays, many cost overrun predictions are conducted manually by expert work planners and quantity surveyors. They use their memory and experience on past projects and intuitively compare these with their current projects. These conventional methods are subject to the available time and attention but also the experience of the professional. While this is a suitable method, lack of time or experience can be both time-consuming and prone to errors. The literature argues that cost overrun prediction is essential for proper construction project management. Therefore, studies have proposed approaches like, among others, three-stage least-squares models or risk-based estimating. Other researchers tend to find key performance indicators of projects to get insight into the cause of cost performance.

Since the access to more data and computational power in recent years, previous construction management studies show how predictive data-driven models help project teams in predicting cost overruns. Such a solution could reduce support for human assessment and increase the efficiency of prediction processes.

While data-driven approaches to predicting cost overruns have been studied for the construction sector, such models have not been developed for multi-utility construction projects. The goal of this study therefore is to develop and test a data-driven decision model that supports cost overrun prediction. It specifically focuses on comparing the performance of suitable binary classification machine learning models for this task.

This study is based on a case of Bam, a large utility contractor in the Netherlands. The organisation strived to improve its cost overrun assessment and owns sufficient data to test the various data-driven approaches. Throughout the study, I adopted the CRISP data mining cycle, comprising five phases. First, the influential causes (so-called features) of cost overruns were explored based on a literature review and semi-structured interviews with 10 specialists including three team leaders, one project leader, one estimator, and five site managers. Next, a subset of features was selected based on 6 criteria, namely availability, quality, categorisability/quantifiability, uniqueness, relevance, and scalability. This led to the removal of features. Finally, categorial data was one-hot encoded to prepare the data for the training stage.

Subsequently, I analysed the following categorical prediction models: random forest, K-nearest neighbours, decision tree, gradient boosting, light gradient boosting, extreme gradient boosting, extra trees, and artificial neural networks. While comparing the models, I also performed recursive feature elimination. This process helps eliminate features that do not

contribute significantly to the predictive performance of the various models. Next, I selected three models with the highest accuracy for further analysis, namely random forest, gradient boosting and extra trees. Accuracy provided sufficient insight into the overall performance because it indicates the extent to which the model can predict the correct classes. Again, I applied recursive feature elimination and hyperparameter optimization to configure the models, and I evaluated the trained model with 2 professional work planners.

During the design of the predictive model, a random forest binary classification model was developed based on 46 features and 888 records that were utility-based. These features included data about the type of methods (directional drilling, rocket drilling) and the site conditions (e.g., the use of dewatering systems). By using recursive feature elimination, it became clear that the three selected models performed optimally on the first 46 of the 71 most influential features. The following performance on the test set was achieved: an accuracy of 0.6367, a recall of 0.6385, a precision of 0.6241, an F1 score of 0.6367 and an area under the receiver operating characteristic curve of 0.7174. Based on the rule of thumb that AUC and ROC need to be more than 0.5 to learn patterns minimally and exceed the quality of random guessing (Narkhede, 2018), the model exceeds minimally necessary learning performance. The recall performance of the trained model (0.6385) shows that it correctly identifies 64% of all overrunning projects from the dataset. The stakeholders argued that this is sufficient for the supportive model to serve as a tool to predict cost overruns. Results show thus that a binary classification random forest model could classify projects which are prone to a cost overrun.

To the literature, this study contributes the insight that a binary categorial predictive model on cost overruns of construction projects in the utility sector is capable of distinguishing projects with and without cost overruns. Although this study has a unique use case, it followed roughly the same methodological approach as, among others, Al mnaseer et al. (2023) and Aung et al. (2023). Al mnaseer (2023) developed a better-performing model than the one developed in this study. A comparison with the model of Aung et al (2023) is difficult because their regression model uses different evaluation parameters. However, it can be said that their ANN model performed better than the other models they tested. This is surprising since ANN was the least-performing model in this study. Moreover, their model predicted costs, while the model developed in this study classifies a project as either a project with a cost overrun or a project without a cost overrun. Both studies used fewer projects for their training data than used in this study. This shows that using more data (projects) to train the predictive model does not necessarily lead to better results.

Limitations of this study can be found in feature selection and the evaluation of model performances. The first limitation is the inclusion of too unique case-specific features, such as the names of stakeholders involved in projects, which hinders future generalisation of the model to other cases. Future research should avoid this and instead use generic features to improve generalisation. The second limitation of this study is that the performance of the model on the test set is ca. 10% lower than the average performance on the folds of the cross-validation, which is the same as accuracy on the train set. Future research could be conducted to determine the reason for this. Potential directions could include examining whether the model was overfitting or if there was data leakage.

II Samenvatting

De nutssector staat in de nabije toekomst voor grote uitdagingen. Deze uitdagingen vinden hun oorsprong onder andere in klimaatverandering en verstedelijking. Bovendien kampen projecten in de ondergrondse nutsinfrastructuur nog steeds met kostenoverschrijdingen. Kostenoverschrijdingen belemmeren het succes van een project aanzienlijk, omdat ze kunnen leiden tot vertragingen, meningsverschillen en financiële verliezen. In een tijd waarin er veel werk is in deze sector, is het wenselijk om realistische economische verwachtingen te scheppen. Daarnaast is de kans groot dat dit probleem toeneemt vanwege de groei in werkzaamheden.

Het voorspellen van kostenoverschrijdingen is moeilijk, omdat nutsprojecten gevoelig zijn voor onzekerheden. Tegenwoordig worden veel voorspellingen van kostenoverschrijdingen handmatig uitgevoerd door ervaren werkvoorbereiders en kostendeskundigen. Zij gebruiken hun geheugen en ervaring met eerdere projecten en vergelijken deze intuïtief met huidige projecten. Deze conventionele methoden zijn afhankelijk van de beschikbare tijd en aandacht, maar ook van de ervaring van de professional. Hoewel dit een geschikte methode is, kan een gebrek aan tijd of ervaring het proces tijdrovend en foutgevoelig maken. De literatuur stelt dat het voorspellen van kostenoverschrijdingen essentieel is voor goed bouwprojectmanagement. Daarom zijn in studies benaderingen voorgesteld zoals onder andere three-stage least-squares modellen of risico-gebaseerde schattingen. Andere onderzoekers zoeken naar key performance indicators (KPI's) van projecten om inzicht te krijgen in de oorzaken van kostenprestaties.

Met de toename van beschikbare data en rekenkracht in de afgelopen jaren, tonen eerdere studies in bouwmanagement aan hoe voorspellende, data-gedreven machine learning modellen projectteams kunnen helpen bij het voorspellen van kostenoverschrijdingen. Een dergelijke oplossing kan de menselijke beoordeling ondersteunen en de efficiëntie van het voorspellingsproces vergroten.

Hoewel data-gedreven benaderingen voor het voorspellen van kostenoverschrijdingen zijn bestudeerd voor de bouwsector, zijn dergelijke modellen nog niet ontwikkeld voor nutsconstructieprojecten. Het doel van deze studie is daarom het ontwikkelen en testen van een data-gedreven model dat ondersteuning biedt bij het voorspellen van kostenoverschrijdingen. Hierbij wordt specifiek gekeken naar het vergelijken van de prestaties van geschikte binaire classificatie-modellen voor deze taak.

Deze studie is gebaseerd op een casus van Bam, een grote aannemer in Nederland. De organisatie streefde ernaar haar beoordeling van kostenoverschrijdingen te verbeteren en beschikt over voldoende data om verschillende data-gedreven benaderingen te testen. Gedurende de studie is de CRISP-dataminingcyclus toegepast, bestaande uit vijf fasen. Eerst zijn de invloedrijke oorzaken (zogenoemde features) van kostenoverschrijdingen verkend op basis van een literatuurstudie en semi-gestructureerde interviews met 10 specialisten, waaronder drie teamleiders, één projectleider, één calculator en vijf uitvoerders. Vervolgens is een subset van features geselecteerd op basis van 6 criteria: beschikbaarheid, kwaliteit, categoriseerbaarheid/kwantificeerbaarheid, uniekheid, relevantie en schaalbaarheid.

Hierdoor zijn bepaalde features afgevallen tijdens het proces. Ten slotte is categorische data one-hot encoded om deze klaar te maken voor de trainingsfase.

Vervolgens zijn de volgende categorische machine learning modellen geanalyseerd: random forest, K-nearest neighbours, decision tree, gradient boosting, light gradient boosting, extreme gradient boosting, extra trees en artificiële neurale netwerken. Bij het vergelijken van de modellen is ook recursive feature elimination toegepast. Dit proces helpt bij het verwijderen van features die niet significant bijdragen aan de voorspellende prestaties van de modellen. Daarna zijn de drie modellen met de hoogste nauwkeurigheid geselecteerd voor verdere analyse: random forest, gradient boosting en extra trees. Nauwkeurigheid gaf voldoende inzicht in de algehele prestaties, omdat het aangeeft in hoeverre het model de correcte klassen kan voorspellen. Opnieuw is recursive feature elimination toegepast en zijn hyperparameters geoptimaliseerd. Het getrainde model is geëvalueerd met twee professionele calculators.

Tijdens het ontwerp van het voorspellende model is een random forest binair classificatiemodel ontwikkeld op basis van 46 features en 888 records die nutsgericht waren. Deze features bevatten gegevens over de gebruikte methoden (gestuurde boringen, raketboringen) en de terreinomstandigheden (bijvoorbeeld het gebruik van bemalingssystemen). Met behulp van recursive feature elimination werd duidelijk dat de drie geselecteerde modellen optimaal presteerden op 46 van de 71 meest invloedrijke features. De prestaties op de testset waren als volgt: een nauwkeurigheid van 0.6367, een recall van 0.6385, een precision van 0.6241, een F1-score van 0.6367 en een area under the receiver operating characteristic curve (AUC) van 0.7174. Gebaseerd op de vuistregel dat AUC-ROC meer dan 0.5 moet zijn om minimaal patronen te kunnen leren en beter te presteren dan willekeurig raden (Narkhede, 2018), voldoet het model aan de minimale leereisen. De recall-prestatie van het getrainde model (0.6385) toont aan dat het model 64% van alle projecten met kostenoverschrijding correct identificeert. De betrokken belanghebbenden gaven aan dat dit voldoende is voor het model om als ondersteunend hulpmiddel voor kostenoverschrijdingsvoorspellingen te dienen. De resultaten tonen dus aan dat een binair classificatiemodel gebaseerd op random forest in staat is om projecten te classificeren die vatbaar zijn voor kostenoverschrijding.

Deze studie draagt bij aan de literatuur door te laten zien dat een binair categorisch voorspellingsmodel in de nutssector in staat is onderscheid te maken tussen projecten met en zonder kostenoverschrijdingen. Hoewel deze studie een uniek praktijkvoorbeeld betreft, volgde het methodologisch gezien grotendeels dezelfde aanpak als onder andere Al Mnaseer et al. (2023) en Aung et al. (2023). Al Mnaseer (2023) ontwikkelde een model dat beter presteerde dan het model in deze studie. Een vergelijking met het model van Aung et al. (2023) is lastig omdat hun regressiemodel andere evaluatieparameters gebruikt. Wel kan worden gezegd dat hun ANN-model beter presteerde dan de andere modellen die zij testten. Dit is verrassend, aangezien ANN in deze studie het slechtst presteerde. Bovendien voorspelde hun model kosten, terwijl het in deze studie ontwikkelde model projecten classificeert als óf met óf zonder kostenoverschrijding. Beide studies gebruikten minder projecten in hun trainingsdata dan deze studie. Dit toont aan dat het gebruik van meer data (projecten) niet per se tot betere resultaten leidt.

Beperkingen van deze studie liggen bij de featureselectie en de evaluatie van modelprestaties. De eerste beperking is de opname van te unieke, casus-specifieke features, zoals de namen van betrokken partijen in projecten, wat de generaliseerbaarheid van het model naar andere casussen belemmert. Toekomstig onderzoek zou dit moeten vermijden en in plaats daarvan generieke features moeten gebruiken om de generaliseerbaarheid te verbeteren. De tweede beperking is dat de prestaties van het model op de testset ongeveer 10% lager zijn dan het gemiddelde van de cross-validatie, wat hetzelfde is als accuraatheid op de train set. Toekomstig onderzoek zou kunnen onderzoeken wat hier de oorzaak van is. Mogelijke richtingen zijn het nagaan of het model overfit was of dat er sprake was van data lekkage.

III List of figures

IV List of tables

1 Introduction

The utility industry faces major challenges in the near future. These challenges find their origin in, among others, climate change and urbanisation. Regarding climate change, the demand for utilities that make electric-driven mobility possible has been increasing since the number of electric vehicles increased (Klein, 2023). As the electrification of mobility continues to develop, the power grid cannot be left behind. As a result, the pressure on this power grid grows. More capacity is urgently needed (NOS, 2024). Solutions are complex, which will increase the costs of this type of utility project in the future (Enexis, 2024).

It is not only projects regarding electricity that are expected to become more common in the future. As urbanisation in The Netherlands increases (CBS, 2023), the demand for all types of utilities is also growing. Although electricity is probably the most crucial one (Antea Group, 2024), utilities such as data, sewer and water should also be built in the urbanizing area. Utility projects are therefore expected to be more frequent in the future.

Moreover, utility projects are prone to uncertainty. Some examples of uncertain, unpleasant occurrences are leaks, subsidence, and contaminated soil (nginfra, 2024). Moreover, projects are becoming more complex. This is because the underground is a big tangle of cables and pipes, see Figure 1. The complication is that realistic project planning is difficult because many uncertainties have to be taken into account.

These uncertainties and a higher frequency of utility projects may cause future utility projects to face more cost overruns. Hence, any measure that can give insight into the cost overruns of utility projects to predict these for future projects is desirable by Bam Energie & Water, the organisation where this research is conducted. This organisation works in the utility sector and has to deal with calculation works where an estimate of a project outcome is often needed. Nowadays, many cost overrun predictions are conducted manually by expert work planners and quantity surveyors in this organisation. They use their memory and experience on past projects and intuitively compare these with their current projects. These conventional methods are subject to the available time and attention but also the experience of the professional. While this is a suitable method, lack of time or experience can be both time-consuming and prone to errors (Khodabakhshian et al., 2024).

With the come of more computational power and the development of artificial intelligence (AI), machine learning models (which is a sub-field in AI) could be developed to support the project team of Bam Energie & Water in predicting cost overruns. While such machine learning applications in combination with cost overruns have been studied for the construction sector, a model capable of predicting cost overruns on multi-utility construction projects has not yet been developed. Hence, in this study, it is explored whether data-driven decision models could potentially support in predicting cost overruns on utility projects. In particular, eight binary classification machine learning models were compared to find the most suitable model for predicting cost overruns. Bam Energie & Water is a suitable client to conduct this research on because it collects project data on a large scale.

This comparison of the models and the development of the final model was done using the CRISP-DM (cross-industry standardized process for data mining) cycle, which includes the phases of business understanding, data understanding, data preparation, modelling, and evaluation. The deployment phase of this cycle was excluded in this research. In the business understanding phase, the influential features according to scientific literature and specialists were examined using a literature study and semi-structured interviews. Moreover, the features from the specialists were validated against the literature. After that, available data was reviewed and collected in the data understanding phase. These data were then prepared for the modelling phase by selecting appropriate features and then one-hot encoding the categorical features. This is a technique used to indicate binary which category within the feature applies to the record in the dataset. In the modelling phase, eight models, namely random forest (RF), K-nearest neighbour (KNN), decision tree (DT), gradient boosting (GB), light gradient boosting (LGB), extreme gradient boosting (XGB), extra trees (XT), and artificial neural network (ANN), were trained. After that, the three best models were selected and further optimised and evaluated. In the last phase, the final model was evaluated by the client.

After training the Random Forest (which is a binary classification model), it could predict cost overruns with an accuracy of 0.6367. This finding shows that the model can distinguish projects which are prone to a cost overrun and projects which does not and thus could serve as a supportive tool for cost overrun prediction.

The rapport is structured as follows: in Chapter 2, the theoretical background on cost overruns will be elaborated on. After that, the research method will be introduced and described in Chapter 3. Subsequently, the results are presented in Chapter 4. These results will be discussed in Chapter 5. Also in Chapter 5, the theoretical and practical contribution will be described and limitations and proposals for future research will be given. Lastly, a conclusion is drawn in Chapter 6.

Figure 1: Current situation in the underground (Potkamp, 2024)

2 Theoretical background

Cost overruns have been broadly examined by researchers through the past years. This is done in various ways. From technical studies to more descriptive studies. This Chapter describes based on the literature what cost overruns are and how different studies used machine learning (ML) models to predict cost overruns.

A cost overrun is defined as follows: *Cost overrun is the amount by which actual costs exceed estimated costs, where costs are measured in local currency, constant prices and at a consistent baseline* (Flyvbjerg et al., 2018). A cost overrun could be measured absolutely or relatively. Absolutely measure means that a certain amount of money is determined whereas a relative measure often displays a percentage or ratio. According to the definition given by Flyvbjerg et al. (2018), it is important to determine what the actual costs, estimated costs, local currency, and baseline are. Flyvbjerg et al. (2018) state that the choice of the baseline should reflect what should be measured. They state moreover that the data used for the research on cost overruns should be a sample of a population of projects.

Construction sectors are facing cost overruns due to tight schedules, complexity in projects and budget limits. Cost overruns significantly impede the success of a project because they may lead to delays, disagreements and financial losses. Cost overrun prediction is essential for proper project management. If cost overruns could be reasonably well predicted, risk could be mitigated better and stakeholders can make more motivated decisions (Aung et al., 2023).

There are studies which tend to find key performance factors of projects (Gunduz et al., 2024; Naji et al., 2023; Yamany et al., 2024) or take a traditional approach (absence of ML) to estimate cost overruns on future projects e.g. three-stage least-squares models, risk-based estimating, and parametric estimation (Abhishek et al., 2010; Liu & Napier, 2010; Melin Jr., 1994). However, these traditional approaches are inefficient and time-consuming and lead often to cost overruns (Khodabakhshian et al., 2024; Shah & Gopinath, 2024).

With the rapid development of AI in recent years, supportive, predictive machine-learning models could enhance the estimation of cost overruns (Cao et al., 2018; Coffie & Cudjoe, 2023a). Machine learning methods contribute to the cost estimation of construction projects because these models can learn from data and anticipate outcomes, which leads to a growing interest in applying machine learning models for cost estimation in the construction industry (Aung et al., 2023). Predicting cost overruns could be done using classification or regression models. When using regression, a value, such as a price, is predicted which is done by e.g. Aung et al. (2023). A classification predicts a class or category. Examples of classes are projects with a cost overrun and projects without a cost overrun. This approach, in addition to regression, is followed by Al mnaseer et al. (2023).

This growing interest in the past years has led to several studies which aimed to predict cost overruns of construction projects or construction services (Al mnaseer et al., 2023; Arabiat et al., 2023; Coffie & Cudjoe, 2023b; Khodabakhshian et al., 2024; Matel et al., 2022; Tajziyehchi et al., 2020). These studies followed roughly the same pattern to come to conclusions i.e. gathering data, selecting influential features, training a model, optimising the

hyperparameters, and evaluating the model. Data is used from general construction projects and, with some exceptions, consists of 100 to 200 samples.

The authors used multiple methods for selecting influential features e.g. recursive feature elimination (RFE), SHAP, and random forest feature selection. In terms of RFE, a feature is a characteristic of the dataset e.g. project duration. SHAP stands for Shapley additive explanations and could explain the output of each machine learning model (Awan, 2023). Moreover, the authors used a variety of, mostly regression, models i.e. K-nearest neighbour (KNN), artificial neural network (ANN), random forest (RF), support vector regressor (SVR), gradient boosting (GB), decision tree (DT), ridge regression (RR), and extra trees (XT). The motivation behind the choice of which model(s) is (are) developed could not always be found. In most situations, it is the starting point of a study that is not always well-motivated. All the models except SVR and RR mentioned above are capable of performing both regression and classification tasks. The models SVR and RR are only regressors (Scikit-learn, 2025). Moreover, only ANN is a deep-learning algorithm whereas the others are shallow-learning algorithms (Hassaan, 2024).

To evaluate the performance of a model, evaluation parameters are used. Examples of evaluation parameters used in these studies to evaluate regression models are mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), R-squared (R^2), and mean absolute percentage error (MAPE). To evaluate classification models, accuracy, precision, recall, and F1-score are used.

Although the growing interest in machine learning applications in the construction industry and the studies resulting from it, few studies have been done where the probability of cost overruns is predicted specifically for utility projects. Because the utility sector has different work from construction, the determined influential features in already performed studies cannot be generalised for utility projects since utility projects include very specific work related to the underground. Moreover, the studies use raw project data, which is subsequently transformed. Enriching data with open-source data is not found in the studies.

Therefore, this study aims to develop a supportive, predictive classification model that could predict cost overruns on utility projects trained on influential features determined based on utility project data, which is enriched with open-source data. Although the majority of the studies developed regression models, in consultation with the client it was decided to apply classification, as the client was convinced that this would lead to better implementation of such a model in the organisation. Therefore, some models earlier mentioned in the literature study are excluded if they only support regression. The following classification models will be used to accomplish this goal: RF, KNN, DT, GB, LGB, XGB, XT, and ANN. LGB and XGB are models that are added to the scope of the research because they are variants of GB, which is mentioned in this literature section. The influential features of cost overruns and associated data on which the models will be trained are sought with a literature review, interviews and a project database of the client.

3 Research method

In this research, the CRISP-DM approach is used. This method stands for Cross-Industry Standard Process for Data Mining and is used to develop data-driven solutions (Huang & Hsieh, 2020; Poh et al., 2018; Sanni-Anibire et al., 2021). A visualization of this method can be found in Figure 2. This method focuses on the process of data analysis. It is a collection of phases which guide through data analysis projects. The sequence of these phases is not rigid. The phase which should be conducted next is dependent on the findings of the previously conducted phase. However, there is some logical order in the phases, which is indicated by the arrows between them. The circle with the arrows inside of it surrounding the phases indicates the iterative character of the method. The (iterative) phases are business understanding, data understanding, data preparation, modelling, and evaluation. Business understanding is a phase where the understanding of the project is developed. Moreover, the requirements from a business perspective become clear. In the data understanding phase, the initial data will be collected and the researcher will become familiar with the data. In the data preparation phase, the final dataset will be created out of the initial dataset. This is done by modifying the initial data, e.g. feature selection and scaling. What covers the modelling phase is the establishment of a model and optimising this. Lastly, in the evaluation phase, the built model is evaluated if it meets the requirements and expectations of the client (Tummers et al., 2025). The deployment phase is out of scope in this study. In the following sub-chapters, the interpretation of the phases in the context of this study is described.

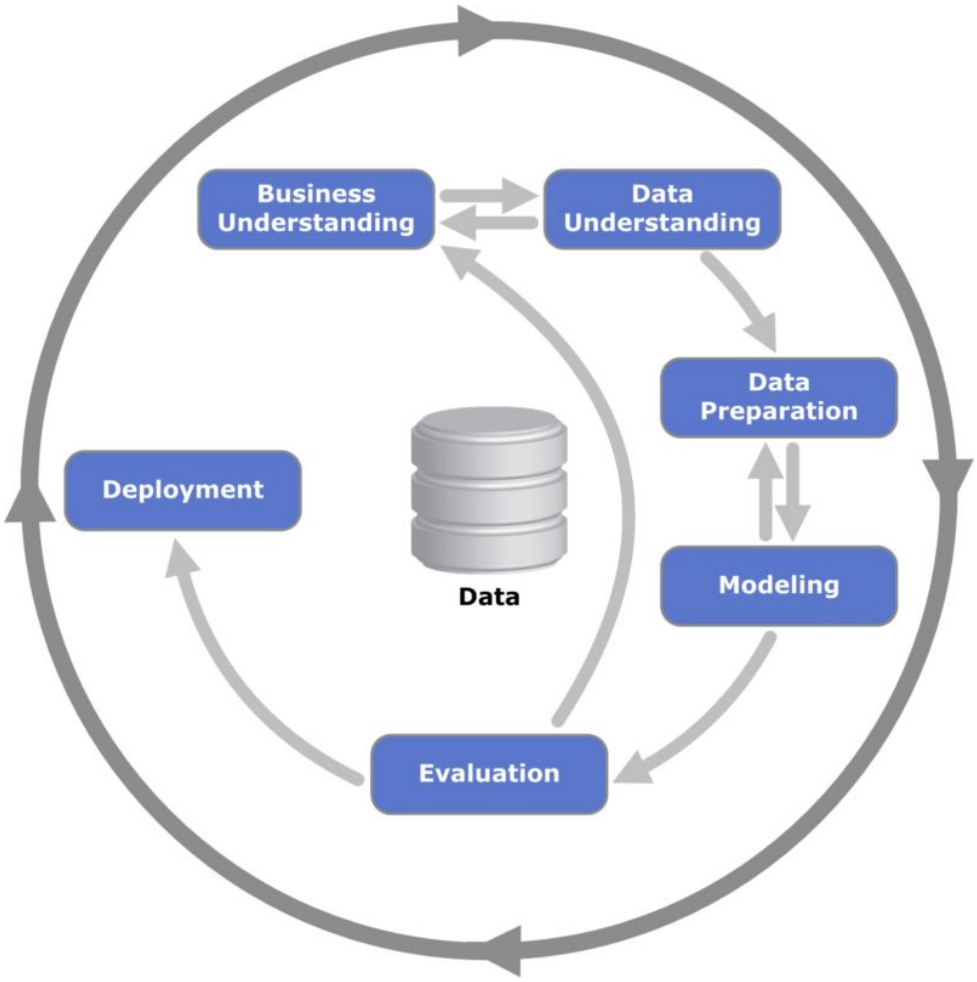

Figure 2: CRISP-DM method (Jensen, 2012)

3.1 Business understanding

In the context of business understanding, a literature review and interviews were conducted to get familiar with the influential features of cost overruns according to these sources. Moreover, this examination established a list of influential features that served as input for the initial features of the train data.

3.1.1 Literature review

To structure the search of the literature, a part of the PRISMA 2020 flow chart will be followed (Page et al., 2021). This method aims to systemize the process of literature reviews. The process of the selected part consists of five phases: identification, screening, retrieval, eligibility check, and inclusion.

To conduct the first phase, studies must be identified. This is done using Scopus. During the searching process, different queries are used, which are listed in Table 1. The article title, abstract, and keywords were searched. These are the standard settings of Scopus, which are considered sufficient for the identification phase.

	Used queries	**Records**
1	"cost overruns" AND "construction industry" AND "artificial intelligence" OR "machine learning"	26
2	"cost overruns" AND "construction industry" AND "prediction" OR "estimation"	100
3	"cost estimation" AND "construction projects" AND "artificial intelligence" OR "machine learning"	56
4	Web findings without specific query	5

Table 1: Used queries to find literature in Scopus

First, the 187 initial records are all combined into one list. Subsequently, 16 duplicates are removed. After that, a record with no known author or authors is excluded. This results in a list of 170 records from searching the queries on Scopus, which can be analysed further in the screening phase. In the screening phase, articles are screened for relevance based on title. Articles that developed cost-predictive machine learning models were primarily considered relevant. After reviewing the articles, 66 out of 170 articles were considered relevant. Then, the relevant articles (so far, based on their title) are checked for availability. In the end, 5 articles were not available, so these are excluded for further analysis. The last phase was about checking articles for eligibility. A list of 35 articles with corresponding features is established. This process is visualised in Figure 3.

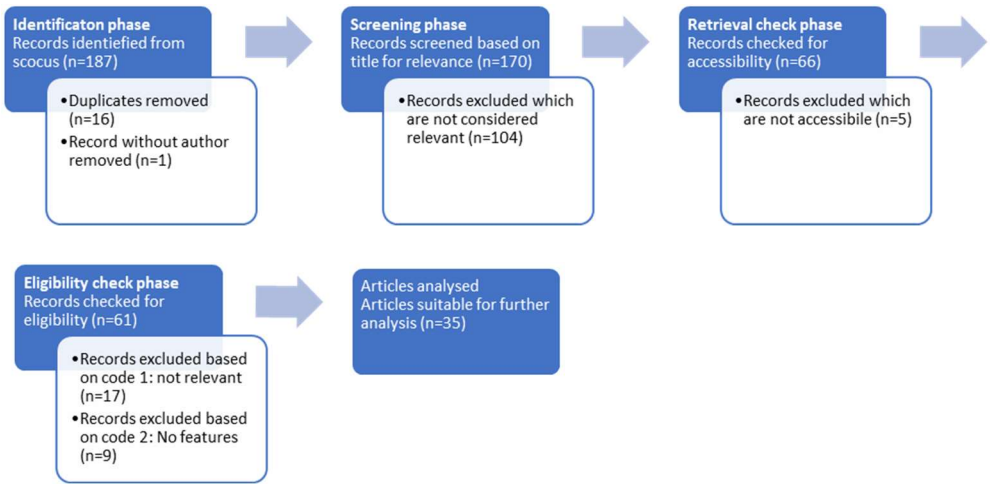

Figure 3: Literature retrieval and assessment process

The identified features were categorized to get insight into which feature category was cited most frequently by the articles. This is done to smoothen little discrepancies between the features of different articles. Features with the same meaning are sometimes called just a little differently among various studies. Moreover, some features were so niche that a minor generalisation was made. In some cases, more features out of the same article are coded with one code because of this generalisation.

The features are allocated to categories based on the knowledge of the researcher of the work field. This knowledge originated from both practice and scientific literature. The categories acted as collection bins in which features with the same meaning were grouped. In contrast, features with identical meanings were described in different ways by various authors of the studies.

3.1.2 Interviews

In addition to the literature review, specialists are interviewed to identify influential features on utility project cost overruns. The interview consists of a single question: “Which factors do you think have the most influence on cost overruns of utility projects, and why?” During the interview, more questions followed, leading to a conversation.

The public comprised various types of specialists. In this context, specialists are defined as individuals involved in the daily workflow of underground utility projects. A total of nine interviews were conducted with ten specialists (one duo). These ten specialists included three team leaders, one project leader, one estimator, and five site managers. Each interview lasted approximately 30 minutes. The audio was recorded to improve the processability.

After the interviews are done, the audio is transcribed using AI and subsequently, the data is analysed by coding in the form of categorising. This is needed to ensure the validity and reliability of the findings (Medelyan, 2024). Coding qualitative data is linking answers from respondents to labels. In this case, inductive labelling is used. By allocating qualitative data to the labels, the results of the interviews could be better analysed (Dingemanse, 2021). As with the analysis of the literature, the data is categorised and a generalisation was made.

Furthermore, some features (i.e. influential characteristics on cost overruns) were added afterwards by the researcher. This was done because some features were almost or vaguely

described but not exactly so named. Therefore, they were added, to distinguish the original features that came from the interviews and the added features.

3.1.3 Pattern matching

Pattern matching is used to compare the results of the literature review with the interviews to get insight into the similarities. While the results of the literature study are generic, the interviews are from a case-specific perspective. Pattern matching is a method which enables a comparison between these perspectives to explore the similarities and differences between the results (Sinkovics, 2018). In this case, pattern matching is performed to validate the features from the interviews with the features from the literature.

In this case, pattern matching is done to get insight into the similarities between the features from the literature, the features from the interviews and the added features. The comparison is made to determine similar features and combine them into one feature to ultimately come to a list of unique features.

The features originating from the literature are listed. These were used as a baseline for the pattern-matching process. For each feature from the literature, it was determined whether there was a feature from the interviews or the researcher with the same meaning. This resulted in a list with unique features since the duplicates were identified among the results of the literature study and the interviews.

3.2 Data understanding

In this phase, the initial database was established, and the researcher became familiar with the data. First, the data availability is examined. After that, the data collection process is described.

3.2.1 Data availability

The starting situation of checking the data availability was a list of features originating from literature, interviews and the added features by the researcher for the reason described in subchapter 3.1.2. After combining equal features from the literature and the interviews, 122 original features are left over as a baseline to check for the availability of the data.

To check data availability, three databases from Bam, the client, were used in combination with open-source data which adds up to a total of four data sources. Each feature from the identified list of features, drawn from a combination of literature, interviews and the researcher, has been assessed for its suitability to participate in the features on which the model will be trained. This was done using the following 6 criteria: availability, quality, categorisability/quantifiability, uniqueness, relevance, and scalability. Availability was a criterion which checked if data exists of a certain feature. The criterion quality assessed if the found data is of sufficient quality. Categorisability and/or quantifiability checked if the data (if it was not numeric) could be quantified or categorised within the scope, such as available time, of this research. Uniqueness was needed as a second check for pattern matching to make sure the feature and its data were unique and did not correspond with other features and/or datasets. The criterion relevance assessed if the feature or its data was relevant for the case of this study. Finally, scalability checked whether the data were widely available so that they could be included in the data without extraordinary effort to extract the values from their

source. Note that these criteria were not introduced as outcomes of this study but were the researcher's underlying rationale for assessing the features and associated data. The conclusions based on them can be found in Table 2.

The first row of the table means that the data is directly usable and is added to the project database. The other reason why features are still usable where data is not directly available, but could be calculated, is that the data is not available directly in the databases available to the researcher but could be calculated with the available data.

Further, there are five reasons defined why a feature is not usable. The first reason states that the data is not available in private or public databases. This does not have to mean that the feature is irrelevant. The second reason why a feature is not usable is when the feature cannot be represented in numbers or categories. An example of this is the skills of the workpeople or safety on the worksite. The next reason is that when a feature determined by literature, interviews or the researcher could be broken down into other features, which are also on the list. This means that this feature is subject to overlap with other features. The subsequent reason why a feature is not usable is that the feature is not relevant for utility projects in NL. It could be that the feature is not relevant for projects in The Netherlands or that the feature is not relevant for utility projects. An example of this is the number of manholes in a project. The article where this feature originated was dedicated to sewage pipelines (Abbas & Aswed, 2024). Therefore, this feature is too specific and is not relevant to the whole utility sector. The last reason is that data of a feature is available, but not on a large or standardised scale. This makes implementing the feature in a large dataset time-consuming or even impossible.

Usability	**Reason**
Yes	Data is directly usable
Yes	Data not directly available, but could be calculated
No	Data not available
No	Feature cannot be categorized or quantified
No	Not this feature, is broken down in more (detailed) feature(s)
No	Not relevant for utility projects in NL
No	The data is not available on a large standardized scale

Table 2: Feature availability reasons

Moreover, on top of the usability of the features, other information is stored such as the origin of the feature, an additional description (if necessary), the coding of the feature (e.g. int, boolean, or float) and the unit (e.g. percentage or euro).

Additional information is given for available features such as data category, owner, interface, format, granularity, granularity level, availability (private or open source), reliability, and a web link. Granularity level and reliability are given in a score of 1, 2, or 3, where 1 is the worst score and three the best.

3.2.2 Data collection

The data from the features which are considered usable in the data availability check were collected from four types of data sources. Three of these sources are databases from the client, which are private. The fourth type of data source is open-source databases. After that, data

about cost types of projects was also found in the private databases and added but was not mentioned in the feature list. This is done because the databases and their contents were examined after the list of features was established. Therefore, in order not to exclude data beforehand, this data was also included in the study. The database is project-based, which means that information on each feature is given per project. The number of records in the database is therefore equal to the number of projects included.

3.3 Data preparation and final feature selection

After the raw data from the features mentioned in the list is established based on literature and interviews, it is gathered and stored in the database, the data is modified to create a database which is suitable for training machine learning models. The first indicators that indicate the financial results of projects were determined based on other features and added for each project in the database. Then, clustering is applied to features which have items with a few occurrences ($n < 25$) such as the feature client. These features are clustered to generalise the database and thus avoid becoming too specific.

Thereafter, the prepared dataset is adjusted to get a balanced distribution into projects that did have cost overruns and those that did not. This is done because evaluation parameters are easier to interpret when the classes in the data are evenly distributed. Since the projects which encountered a cost overrun are in the minority (444 out of 1506), non-cost-overrun projects are deleted to get an equal number of samples per class. The projects with a positive financial result chosen to be retained in the database were selected at random.

Subsequently, on top of the first screening based on Table 2, the database is cleaned up from features that could not be used in the training process of the machine learning models. The following reasons were determined why a feature could not be used in the training process: a feature did not have an appropriate format, was related to the target variable, was not known in advance, was not relevant for prediction or was not used as a target variable. The last reason is added since there were binary and categorial target values (labels) determined. In consultation with the client, it was decided to proceed with binary classification and not investigate categorical classification further. Therefore, the target variable generated for categorial classification is dropped.

After that, the remaining features that are categorical, e.g. project leaders or season, are one hot encoded. This is a method which can be used to encode categorical data with more than two categories (Andishgar et al., 2025).

The train and test splits are made with a ratio of 70% train and 30% test. Although scaling the data is often considered an important step in data preparation (Khoong, 2023), it is chosen as an approach to not scale the data on the forehand but to compare the results of the models with and without scaling the data. To compare the need for scaling, the data is scaled using the min-max scaler, which scales the values between 1 and 0. In the modelling phase, performances with and without scaled data were compared. The best-performing option is worked further with.

3.4 Modelling

In this subsection, the process of modelling is described. First, the best models are selected using feature importance and recursive feature elimination. After that, the three best models will be further designed and optimised.

3.4.1 Feature importance and model selection

The first step in the modelling stage was to determine feature importance. This was done to enable recursive feature elimination (RFE) and to give insight into the contribution of features to cost overruns to the client. RFE is a method to search for the best subset of features in a dataset that contributes the most to the predictive capability (Brownlee, 2020). First, all eight models were trained without hyperparameter optimisation or RFE. These performances were measured in two variants of the accuracy metric, namely the mean accuracy of the folds of k-fold cross-validation on the train set and the accuracy measured based on a test set. Accuracy is used because it is an evaluation parameter which provides insight into the overall performance. Moreover, the standard deviation is given over the values of the cross-validation accuracies to get insight in the distribution of the values. Stratified cross-validation was used to get an indication of how the model would perform on unseen data. Further, it gave a more generalised indication of this than when the model was trained once (Brownlee, 2023).

Based on these results, it was determined whether scaling the data improves the model's performance and subsequently, whether scaling was implemented or not. Furthermore, the model was chosen, which served as the basis to calculate the SHAP values. SHAP values were used to get insight into the distribution of feature values compared to the contribution of the model's performance and as input for RFE. SHAP stands for Shapley additive explanations and could explain the output of each machine learning model (Awan, 2023).

Once the relevance of the features was understood, the training phase could get underway to find the three best-performing models, which were examined further. This was done by training the eight models again with RFE to prevent exclusion from models too early in the training process because it provided insight into the performances of the models on different subsets and not only on the total set of features.

The performance of the models is again, as in the process of arriving at the model for the basis of feature importance, mapped to average accuracy of cross-validation on the train set and accuracy based on the test set. Subsequently, the three best-performing models were chosen and examined further.

3.4.2 Model design, optimization, and evaluation

After analysing the results, the three best models were selected. RFE was conducted to gain insight into the optimal feature sub-set to maximise the predictive capability of the models. Hyperparameter optimization, a method to find the optimal parameters for a model, was performed at each run of RFE. This was done because the optimal hyperparameters were unknown upfront and hence could improve the performance in the learning phase. Grid search hyperparameter optimization was used to find the best combination of certain parameters. The grids were determined by an arbitrary approach. Further, stratified cross-validation with five folds was performed on each run within RFE.

The best models were evaluated on the optimal subset of features. This optimum is determined using mean cross-validation performances on the train set and test set performances which are calculated for each run in the RFE process. This conclusion is drawn visually based on the highest scores seen in the RFE accuracy plots.

The parameters used were accuracy, recall, precision, and F1 score. The formulas to calculate these parameters are given below (Seol et al., 2023). In the formulas, TP stands for true positive, TN for true negative, FP for false positive and FN for false negative. Accuracy indicates how capable the model is of finding the right classes of all instances of the dataset. Recall is giving insight into the capability of finding the true positives out of all positives. Precision indicates how often the model classifies a positive while it is a negative.

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

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

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

On top of the evaluation parameters, the receiver operating characteristic (ROC) curve was plotted, and the area under the curve (AUC) value was calculated. This curve indicates how good the model is in distinguishing categories, namely cost overrun projects and non-cost-overrun projects. The ROC curve used true positive rates (TPR) or recall values, which were plotted on the y-axis, and false positive rates (FPR), which were plotted on the x-axis. The more the line deflects to the upper left corner, the better the model is. A ROC curve with an AUC of 0,5 (diagonal line from left under to right upper corner) means that the model is not capable of distinguishing classes. An AUC of e.g. 0,7 indicates that the model has a 70% chance that it is capable of distinguishing between the positive and the negative classes. The formulas for TPR and FPR are given below (Narkhede, 2018).

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

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

Based on the performance parameters and the AUC of the ROC plots, the best model is chosen to be the most capable of predicting cost overruns based on utility project data. The model with the best mean scores was selected as the best model.

3.5 Evaluation of the model by the client

After the evaluation of the model was done and the best model was selected, the model was introduced to the client and tested. This was done using a discussion in which two people were

present (excluding the researcher). In this discussion, after the introduction to the model, a few imaginary cases were determined and used as input for the model. Subsequently, the change of a cost overrun was predicted for these cases. Some modifications were made in the parameters of the imaginary cases after running to get insight into the behaviour of the model.

To bring structure to the discussion, some questions were asked to the attendees to evaluate the model. The attendees were both senior work planners. The questions were:

1. What do you think about the usability of the model?
2. To what extent do you trust the model?
3. What do you think of the features (input values) used by the model?
4. Do you think you will use the model (possibly a further developed version) in your daily work?

In question one, usability is defined as the ease with which the model can be used. The definition as determined in this question states that the fewer barriers there are before output can be extracted from the model, the easier it is to use. These questions were discussed together. After discussing the question between the two attendees, one answer was given to the question. Thus, it can be said that there was a joint response from the attendees. No individual opinions were collected. From these answers, conclusions were drawn based on the main message of the answer.

4 Results

In this chapter, the results of the research are described. This is done using the structure of the methodology, namely the CRISP-DM cycle, presented in Figure 2.

4.1 Business understanding

In this chapter, the influential features according to literature and specialists are presented.

4.1.1 Literature review, interviews and pattern matching

The influential features of 35 papers are listed. Further, 495 features were coded, which resulted in 115 different codes. The different codes represent 115 different features determined by the literature. These 115 features are grouped into 8 groups. The underlying features of each code group are given in Appendix A. Further, in Appendix B could be seen which authors mentioned which single feature in their studies.

In the nine interviews, a total of 24 different features were proposed by the interviewees that are influential on cost overruns. The features, and the interviews in which they were mentioned, are shown in Table 14, which can be found in Appendix C

The results of the pattern-matching process can be found in Appendix D. After all features are allocated to each other, 122 unique features are left over. In this Appendix it is indicated whether it is mentioned in the literature, the interviews, or whether the researcher added the feature. The resulting table will be used in the next stage to search data for all the features.

4.2 Data understanding

In this chapter, the results from the data understanding phase are presented. First, data availability results are described, subsequently, results from data collection are given.

4.2.1 Data availability

The availability per feature is presented in Appendix E. Further, additional information on the available data is presented in Appendix F. In Table 3, presented below, the number of features connected to usability is given. The table could be seen as a summary of the two appendices.

Usability	**Reason**	**N**
Yes	Data is directly usable	22
Yes	Data not directly available, but could be calculated	6
No	Data not available	42
No	Feature cannot be categorized or quantified	39
No	Not this feature, is broken down in more (detailed) feature(s)	12
No	Not relevant for utility projects in NL	9
No	The data is not available on a large standardized scale	5

Table 3: Summarizing results of data availability

4.2.2 Data collection

The first features of the project database are presented in Table 4. The database consists of 83 features (due to early stage one-hot encoding) and 1506 projects. Note that after the data

availability check as done in the former subchapter, other data is added from the databases of the client that was not in the scope of this data availability check because the databases were not known to the researcher at that point. Table 4 includes the initial features after the binning of features with a few occurrences (n < 25) is done. The following actions regarding data preparation, as described in 3.3, are performed on this feature set.

#	Feature	Description
1	Projectcode	Unique project code
2	Project description	Description of the project
3	Year	Year of execution of the project
4	Turnover	The turnover made on the project
5	Profit	The profit made on the project
6	Cost	The cost made to realise the project
7	Client 1	Binary: 1 if client is included, 0 if not
8	Client 2	Binary: 1 if client is included, 0 if not
9	Client 3	Binary: 1 if client is included, 0 if not
10	Client 4	Binary: 1 if client is included, 0 if not
11	Client 5	Binary: 1 if client is included, 0 if not
12	Client 6	Binary: 1 if client is included, 0 if not
13	Other clients	Binary: 1 if client is included, 0 if not
14	Elektra LS	Binary: 1 if discipline is included, 0 if not
15	Elektra MS	Binary: 1 if discipline is included, 0 if not
16	Water	Binary: 1 if discipline is included, 0 if not
17	Gas LD	Binary: 1 if discipline is included, 0 if not
18	CAI	Binary: 1 if discipline is included, 0 if not
19	Gas HD	Binary: 1 if discipline is included, 0 if not
20	Media	Binary: 1 if discipline is included, 0 if not
21	Engineering	Binary: 1 if discipline is included, 0 if not
22	Other disciplines	Binary: 1 if discipline is included, 0 if not
23	Cost_overrun	Binary: 1 if there is cost overrun, 0 if not
24	Calculatie	The expected turnover on the project
25	Location	Town where project is executed
26	Street	Street where project is executed
27	Zipcode	Zipcode where project is executed
28	Province	Province where project is executed
29	Execution_length	Execution length of the project in days
30	Start_date	Start date of execution of the project
31	End_date	End date of execution of the project
32	Urbanisation	Urbanisation number. 1 is high, 5 is low
33	Season	Season in which the project in executed
34	Mean_temp	Mean temperature during execution
35	Mean_rain	Mean rainfall during execution
36	Length_trench	Sum of length of trench per discipline
37	Steered_drilling	Meters of steered drilling
38	Transfers	Number of house transfers
39	Rocket_drilling	Meters of rocket drilling
40	HxT Work preperation	Hours x tariff work preperation
41	HxT Execution	Hours x tariff execution
42	Procurement hiring general	Hiring staff

#	Feature	Description
43	Permanent materials	Cost of permanent materials
44	Cost equipment	Cost of the equipment
45	Subcontractor	Cost of subcontractor
46	Indirect and general cost	Indirect and general cost
47	Storage overhead cost	11 % add-up on other costs
48	Damage costs	Cost of damages
49	Overhead cost own work	Overhead cost own work
50	HxT Specialists	Cost of specialists
51	HxT Workshop	Cost of workshop
52	HxT migration account	Hours x tariff from migration account
53	Travel and lodging expenses	Cost of travel and lodging
54	HxT Staff	Hours x tariff staff
55	Charged indirect and overhead costs	Diverse costs
56	HxT Calculation	Hours x tariff calculation
57	HxT Designcost	Hours x tariff design
58	HxT Projectmanagement	Hours x tariff projectmanagement
59	HxT Planning	Hours x tariff planning
60	HxT Advice	Hours x tariff advice
61	HxT Mechanic E&W	Hours x tariff mechanic from Bam E&W
62	HxT Miscellaneous	Hours x tariff miscellaneous
63	HxT Groundworker E&W	Hours x tariff groundworker from Bam E&W
64	OA Miscellaneous	Subcontractors
65	Fees, precariotax, etc	Fees, precariotax, etc such as other taxes.
66	Projects Housing	Cost of realising work spaces on project
67	Damages	Cable damagees
68	HxT Miscellaneous	Transfer of incorrectly booked hours
69	Rent terrain/building	Cost of renting terrain/buildings
70	HxT Hiring	Hours x tariff hired staff
71	Productivity [m/day]	The productivity of making trench in m/day
72	Pandemic_active	Binary: 1 if the pandemic was active, 0 if
73	Executor	Name of the executor
74	Project leader	Name of the project leader
75	Drainage_used	Binary: 1 if the pandemic was active, 0 if
76	Length_LS	Length of low voltage cable used
77	Length_MS	Length of mid voltage cable used
78	Length_gasLD	Length of low pressure gas pipe used
79	Length_gasHD	Length of high pressure gas pipe used
80	Length_water	Length of water pipe used
81	Length_cable_diverse	Length of other cables used
82	Lengte_pipe_diverse	Length of other pipes used
83	Cost_overun_category	Catogorial: 0 when project is tie. 1 between 0 and 10%. 2 between 10 and 30% and 3 if result was 30% or higher. Same for negative results.

Table 4: Initial features of the project database

4.3 Data preparation and final feature selection

The features that are removed with their corresponding reasons, are given in the table below. The features province (nr. 28), urbanisation (nr. 32), season (nr. 33), executor (nr. 73), and project leader (nr. 74) are one hot encoded. The database dimensions after preparation were 888 projects and 71 features. The number of projects was reduced from 1506 to 888 to achieve a balance between those that did have cost overruns and those that did not. The final features can be seen in Table 6.

Feature	Feature numbers	Reason
Projectcode	1	Not relevant for predicting
Project description	2	Not numerical format
Location	25	Not numerical format
Street	26	Not numerical format
Zipcode	27	Not numerical format
Start_date	30	Not numerical format
End_date	31	Not numerical format
Profit	5	Relates to target variable
Cost	6	Relates to target variable
Turnover	4	Not known on forehand
Mean_temp	34	Not known on forehand
Mean_rain	35	Not known on forehand
HxT Work preparation till HxT Hiring	40-70	Not known on forehand
Year	3	Not relevant for predicting
Cost_overrun_category	83	Not used as target variable

Table 5: Removed features with corresponding reasons

#	Feature name	#	Feature name	#	Feature name	#	Feature name
1	Client 1	19	Length_trench	37	Spring	55	Executor 16
2	Client 2	20	Steered_drilling	38	Summer	56	Executor 17
3	Client 3	21	Transfers	39	Winter	57	Executor 18
4	Client 4	22	Rocket_drilling	40	Executor 1	58	Executor 19
5	Client 5	23	Productivity [m/day]	41	Executor 2	59	Executor 20
6	Client 6	24	Pandemic_active	42	Executor 3	60	Executor 21
7	Other clients	25	Drainage_used	43	Executor 4	61	Project leader 1
8	Elektra LS	26	Length_LS	44	Executor 5	62	Project leader 2
9	Elektra MS	27	Length_MS	45	Executor 6	63	Project leader 3
10	Water	28	Length_gasLD	46	Executor 7	64	Project leader 4
11	Gas LD	29	Length_gasHD	47	Executor 8	65	Project leader 5
12	CAI	30	Length_water	48	Executor 9	66	Project leader 6
13	Gas HD	31	Length_cable_diverse	49	Executor 10	67	Urbanisation 1
14	Media	32	Lengte_pipe_diverse	50	Executor 11	68	Urbanisation 2
15	Engineering	33	Drenthe	51	Executor 12	69	Urbanisation 3
16	Other disciplines	34	Groningen	52	Executor 13	70	Urbanisation 4
17	Calculatie	35	Overijssel	53	Executor 14	71	Urbanisation 5
18	Execution_length	36	Autumn	54	Executor 15		

Table 6: Final features after data preparation

4.4 Modelling

In this chapter, the results are given from the modelling process. First, feature importance and model selection results are shared. After that, the final model design, optimisation and evaluation are presented.

4.4.1 Feature importance and model selection

The performances of the initial training of the models with scaling are presented in Table 7. The performances without training are presented in Table 8. This training iteration is done to check if scaling the data has improved the performance of the models. Based on the two tables

presented below, it could be concluded that scaling did not improve the performances. Therefore, unscaled data is used for the rest of this study.

Model	CV (train set) accuracy	σ CV (train set) accuracy	Test set accuracy
RF	0.6940	0.0239	0.6180
KNN	0.5201	0.0239	0.5281
DT	0.6555	0.0245	0.6030
GB	0.6876	0.0519	0.6704
LGB	0.6811	0.0542	0.6142
XGB	0.7037	0.0363	0.6255
XT	0.6748	0.0255	0.6592
ANN	0.5039	0.0581	0.5094

Table 7: Model performances with non-scaled data

Model	CV (train set) accuracy	σ CV (train set) accuracy	Test set accuracy
RF	0.6843	0.0293	0.5843
KNN	0.6295	0.0498	0.5805
DT	0.6054	0.0346	0.6330
GB	0.6794	0.0689	0.6442
LGB	0.6650	0.0325	0.6067
XGB	0.6537	0.0239	0.6479
XT	0.6618	0.0253	0.6142
ANN	0.6378	0.0574	0.6030

Table 8: Model performances with scaled data

Subsequently, the SHAP values were calculated based on the Random Forest model. Although XGB performed a bit better, SHAP values based on this model are given in log odds and are not per class. Therefore, the second-best performing model (based on train set mean CV accuracy) is used to calculate the SHAP values. The 20 most influencing features of cost overruns can be seen in Figure 4. The feature importance of all features can be seen in Appendix G. How higher the features are in the table, the more influential a feature is. Moreover, red indicates a high value for a single instance and blue is low. Moreover, the more to the right the measurement points are, the more effect on cost overruns.

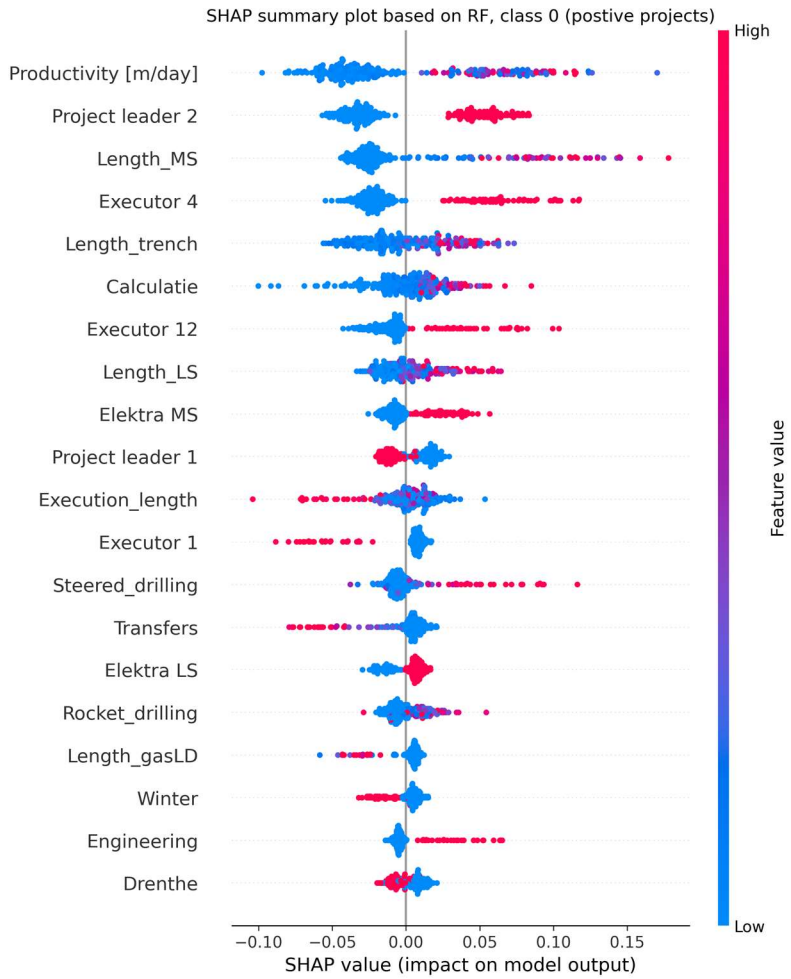

Figure 4: 20 most influencing features on cost overruns

The models are run with RFE based on the feature importances. The performances of the models are presented in Table 9. Graphs showing accuracies by feature during the RFE process can be found in Appendix H. In these graphs, max CV is the maximum mean cross-validation accuracy on the train set reached throughout the RFE process. The max AC represents the maximum accuracy on the test set reached throughout the RFE process.

Model	**Max CV (train set) accuracy**	**Max test set accuracy**	**Mean**
RF	0.7214	0.6667	0.6941
KNN	0.6118	0.5918	0.6018
DT	0.6764	0.6217	0.6491
GB	0.7053	0.6816	0.6935
LGB	0.6973	0.6554	0.6764
XGB	0.7166	0.6442	0.6804
XT	0.6957	0.6891	0.6924
ANN	0.6121	0.6142	0.6131

Table 9: Performances of models with RFE

Table 9, presented above, shows that RF, GB, and XT performed best on average. These models were investigated further.

4.4.2 Model design, optimisation and evaluation

The final performance of the models is presented in Table 10. The performance is based on the subset of the 46 most influential features (first 46 features in Table 6). This was done because this subset performs best according to the graphs in Appendix I. Further, the ROC curves for the three models are given in Appendix I. It could be seen that RF performed the best on average. The hyperparameters used for the final models are presented in Appendix J.

Evaluation parameter	**Score**	**Standard deviation**
RF		
CV (train set) Accuracy	0.7166	0.0453
CV (train set) Recall	0.7163	0.0630
CV (train set) Precision	0.7208	0.0411
CV (train) F1-score	0.7179	0.0493
Test set Accuracy	0.6367	n/a
Test set Recall	0.6385	n/a
Test set Precision	0.6241	n/a
Test set F1-score	0.6367	n/a
ROC AUC	0.7191	n/a
Average score RF	**0.6919**	
GB		
CV (train set) Accuracy	0.6989	0.0508
CV (train set) Recall	0.7037	0.0461
CV (train set) Precision	0.7053	0.0611
CV (train set) F1-score	0.7033	0.0454
Test set Accuracy	0.6404	n/a
Test set Recall	0.6385	n/a
Test set Precision	0.6241	n/a
Test set F1-score	0.6367	n/a
ROC AUC	0.7143	n/a
Average score GB	**0.6850**	
XT		
CV (train set) Accuracy	0.6940	0.0393
CV (train set) Recall	0.7164	0.0403
CV (train set) Precision	0.6916	0.0424
CV (train set) F1-score	0.7032	0.0362
Test set Accuracy	0.6255	n/a
Test set Recall	0.6385	n/a
Test set Precision	0.6103	n/a
Test set F1-score	0.6255	n/a
ROC AUC	0.7174	n/a
Average score XT	**0.6692**	

Table 10: Final performances of the three best models

4.5 Evaluation of the model by the client

A discussion is held with experts to evaluate the model. This is done using the questions listed in Chapter 3.5. The answers to the questions are given below. The evaluation of the experts is

processed in the final model. The experts agreed with each other on the answers, sometimes after a brief discussion. The discussions held were mostly about the details in the answers. One specialist was more optimistic than the other about the future of the model.

The experts had less discussion about the first question. The model is not practical enough to use it in the daily workflow. This is because the features should be easier to fill in before the model can make a prediction. In the current situation, every single feature should be filled in manually. Since some features, such as executor or project leader, are one-hot encoded, one executor or project leader can be selected and the others should be unselected. Moreover, the interface of the model is underdeveloped. On usability, improvement could be made.

In question two, it is asked to what extent the experts trust the model. The experts' trust in the model is limited. This is because the model is a relatively generic reflection of a so much more complex practice since not everything from practice is included in the model. Moreover, the model is a black box; it is not visible what processes take place behind the interface. This decreases the trust in the model. It can be concluded that specialists have confidence in the model to some extent. However, their own feelings on projects far outweigh it.

Based on the answer to question three, the features have been approved by the client, except for a few. For example, the feature year was taken out at the customer's request. This feature is irrelevant in a predictive algorithm. Furthermore, they had their doubts about the feature pandemic because this was an exceptional situation. The feature year has been removed, but the feature pandemic has been retained.

The last question tries to indicate whether the specialists will use the model in their daily work. The client wants to start using the model in the future. However, this will require further development. For instance, the existing features need to be linked to each other to make it more efficient to fill them in. An example is that if one selects a single season in which the project was carried out, the other seasons are deselected right away. A brief overview of the questions and answers can be found below, in Table 11.

#	**Question**	**Answer**
1	What do you think about the usability of the model?	Too unwieldy now, should be easier to fill in. Appearance could be even better. It is functional with this look though.
2	To what extent do you trust the model?	The model is based on generalized reality. Outside, it's all more complex. As a sparring partner it is nice, as a decisive advisor it is not.
3	What do you think of the features (input values) used by the model?	Years should be taken out because we are predicting something in the future. Pandemic was exceptional.
4	Do you think you will use the model (possibly a developed version) in your daily work?	Yes, if it is easy and the features are paired with each other. And then mainly to get gut feelings confirmed.

Table 11: Results of the discussion with experts

5 Discussion

This study developed a supportive data-driven, binary classification model that is capable of predicting cost overruns of utility projects. The study's main finding is that a random forest binary classification model trained 71 features, such as client, utility type, execution methods, season, location, and 888 records, which are projects, performs optimal on the 46 most influential features and could predict cost overruns on utility projects with an accuracy of 0.6367, a recall of 0.6385, a precision of 0.6241, an F1 score of 0.6367 and an area under the receiver operating characteristic curve of 0.7174.

This main finding shows that a binary classification random forest model as described above is capable of distinguishing projects which are prone to a cost overrun and projects which does not. This can be seen from the fact that the accuracy and AUC of ROC are above 0.5 because this is the threshold where the model does more than if it were guessing randomly and had not learnt patterns in the data. Important in this case is that the model detects as many cost overruns as possible. It is rather demanded that the model predicts a false positive than a false negative. This is because in that case, a project (which, according to the model, will not face a cost overrun) be checked. The ability of the model to detect the projects which will face a cost overrun is best indicated by the evaluation parameter recall. Since the recall is 0.6385, it could be concluded that the model is capable of finding ca. 64% of the projects which faced a cost overrun. According to the client, this is sufficient for a supportive model.

However, besides the fact that the best-performing model was a random forest, other models are also capable of distinguishing cost-overrun projects from non-cost-overrun projects. From the 8 analysed models, RF, GB, and XT performed best. The difference in the performance is minimal. All evaluation parameters are in a bandwidth of 5% in relation to the same parameters measured based on the other models. This means that even though RF was the best model in this case, the other 2 models performed almost equally to RF.

To the literature, this study contributes the insight that a binary categorial predictive model on cost overruns of construction projects in the utility sector is capable of distinguishing projects with and without cost overruns. Although this study has a unique use case, it followed roughly the same methodological approach as, among others, Al mnaseer et al. (2023) and Aung et al. (2023). However, the majority of studies conducted on predicting cost overruns in the construction industry used regression models and thus predicted the costs of a project instead of whether a project was going to have a cost overrun or not. Such models use other evaluation parameters. Therefore, a comparison with the (values of) evaluation parameters of the existing studies and this study is not relevant. What could be concluded is that ANN performed best of the models Aung et al (2023) analysed. In this study, ANN was the least-performing model. Although Al mnaseer (2023) developed a regression ANN model, they also performed classification on top of that which led to comparable evaluation parameters. Their results are an accuracy of 0.9220, a precision of 0.6816, a recall of 0.8068 and an F1 score of 0.7104. The ANN model in their study performed reasonably better than in this study. A high recall indicates that their model is better capable of finding the projects which face a cost overrun than the random forest model developed in this study.

The second contribution to the literature is the insight that using more samples (projects) to train the model on lead not always to better results. Aung et al. (2023) used data from 250 construction projects for their study which is remarkably less than the amount of projects used for this study. Although comparison is hard between regression and classification, they stated that the model performed better than traditional results. Al mnaseer et al. (2023) used 291 construction projects for the training of the model. As indicated earlier, they have achieved better results on classification than those achieved in this study with less training data.

The practical contribution is that cost estimators can be informed of the possibility of a cost overrun for a certain utility project. Such a model converts certain experiences of these people into a program which inexperienced cost estimators can use. However, although the model developed in this study has predictive capabilities, the outcomes of the predictions of the model were in line with the expectations of the specialists. On one hand, this means that the model is validated by them. On the other hand, this emphasises that the model only serves as a supporting tool and not as a decisive. Another practical contribution is that the features the experts expected to be influential did indeed affect cost overruns according to the data-driven feature importance analysis. This proves that they have a good idea of the influencing factors on cost overruns, and can thus help them assess project risks in the future.

Limitations of this study can be found in feature selection and the evaluation of model performances. The first limitation is the inclusion of case-specific features in the model, such as executor or project leader, which hinders future generalisation of the model to other cases. This is because other companies do not have the same executors or project leaders. Future research can be done to identify all these features and to break them down into properties of those features so that the generalisability towards other organisations can be done more easily since it becomes possible to add data to the model from another organisation.

The second limitation of this study is that the performance of the model on the test set is ca. 10% lower than the average performance on the folds of the cross-validation on the train set. Future research could be done to find a reason for this. Potential directions could be to examine if the model was overfitting or if there was data leakage.

6 Conclusion

The utility industry faces major challenges in the near future. These challenges find their origin in, among others, climate change and urbanisation. Moreover, nowadays, many cost overrun predictions are conducted manually. This is inefficient and time-consuming. (Khodabakhshian et al., 2024). In addition to that, the construction sector, and therefore the utility infrastructure sector, is still suffering from cost overruns (Eizakshiri et al., 2011). Since the rise of computational power in recent years (Cao et al., 2018; Coffie & Cudjoe, 2023a), the demand for a supportive predictive data-driven model to help work planners predict cost overruns has increased to reduce, inefficiency, and time consumption. However, such a model is not developed yet for the utility sector. Hence, in this study, such a model is developed based on project data from the utility sector.

This was done using the CRISP-DM cycle. The influential features of cost overruns were explored due to the conduction of a literature review and semi-structured interviews. When data was collected based on these findings, features were selected based on certain criteria and data was prepared. Subsequently, eight models were trained and analysed to come to the three best-performing models, which are further examined. The initial analysed models were RF, KNN, DT, GB, LGB, XGB, XT, and ANN. RFE was used in this phase to avoid missing a well-performing model on a particular subset of features. The evaluation parameter 'accuracy' is used to evaluate these models. From these models, RF, GB, and XT were the best-performing ones. These were further analysed with, again, RFE and hyperparameter optimization. The evaluation of the model with the client was done by a discussion.

During the design of the predictive model, a random forest binary classification model was ultimately developed based on features and 888 records that were utility-based and included utility project parameters such as directional drilling, rocket drilling and use of drainage. By using recursive feature elimination, it became clear that the models performed optimally on the 46 most influential features. The following performance on the test set was achieved: an accuracy of 0.6367, a recall of 0.6385, a precision of 0.6241, an F1 score of 0.6367 and an area under the receiver operating characteristic curve of 0.7174.

To the literature, this study contributes the insight that a binary categorial predictive model on cost overruns of construction projects in the utility sector is capable of distinguishing projects with and without cost overruns. Although this study has a unique use case, it followed roughly the same methodological approach as, among others, Al mnaseer et al. (2023) and Aung et al. (2023). Al mnaseer (2023) developed a better-performing model than the one developed in this study. While comparison with the model of Aung et al. (2023) is hard because regression models use other evaluation parameters, their ANN model also outperforms other models while ANN was the least-performing model in this study. This means that, although the model developed in this study has predictive capabilities, the results obtained in this study may not be groundbreaking compared to other works. Given this fact, the second contribution to the literature is the insight that using more samples (projects) to train the model on lead not always to better results. Both studies used fewer projects for their training data than used in this study.

7 Bibliography

Abbas, A., & Aswed, G. K. (2024). Enhancing Sewage Pipeline Project Cost Estimations in Iraq through Artificial Neural Network Models. *IOP Conference Series: Earth and Environmental Science*, *1374*(1). https://doi.org/10.1088/1755-1315/1374/1/012086

Abhishek, B., Ch., A. P., Samuel, L., C, S. K., & L, M. F. (2010). Three-Stage Least-Squares Analysis of Time and Cost Overruns in Construction Contracts. *Journal of Construction Engineering and Management*, *136*(11), 1207–1218. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000225

Al mnaseer, R., Al-Smadi, S., & Al-Bdour, H. (2023). Machine learning-aided time and cost overrun prediction in construction projects: application of artificial neural network. *Asian Journal of Civil Engineering*, *24*(7), 2583–2593. https://doi.org/10.1007/s42107-023-00665-7

Andishgar, A., Bazmi, S., Lankarani, K. B., Taghavi, S. A., Imanieh, M. H., Sivandzadeh, G., Saeian, S., Dadashpour, N., Shamsaeefar, A., Ravankhah, M., Deylami, H. N., Tabrizi, R., & Imanieh, M. H. (2025). Comparison of time-to-event machine learning models in predicting biliary complication and mortality rate in liver transplant patients. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-89570-4

Antea Group. (2024). *Verstedelijking & Energie* . https://www.rijksoverheid.nl/documenten/rapporten/2024/07/04/rapport-verstedelijking-energie

Arabiat, A., Al-Bdour, H., & Bisharah, M. (2023). Predicting the construction projects time and cost overruns using K-nearest neighbor and artificial neural network: a case study from Jordan. *Asian Journal of Civil Engineering*, *24*(7), 2405–2414. https://doi.org/10.1007/s42107-023-00649-7

Aung, T., Liana, S., Htet, A., & Bhaumik, A. (2023). Using Machine Learning to Predict Cost Overruns in Construction Projects. *Journal of Technology Innovations and Energy*, *2*, 1–7. https://doi.org/10.56556/jtie.v2i2.511

Awan, A. A. (2023, June). *An Introduction to SHAP Values and Machine Learning Interpretability*. https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability?dc_referrer=https%3A%2F%2Fwww.google.com%2F

Brownlee, J. (2020, August 28). *Recursive Feature Elimination (RFE) for Feature Selection in Python*. https://machinelearningmastery.com/rfe-feature-selection-in-python/

Brownlee, J. (2023, October 4). *A Gentle Introduction to k-fold Cross-Validation*. https://machinelearningmastery.com/k-fold-cross-validation/

Cao, Y., Ashuri, B., & Baek, M. (2018). Prediction of Unit Price Bids of Resurfacing Highway Projects through Ensemble Machine Learning. *Journal of Computing in Civil Engineering*, *32*. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000788

CBS. (2023, June 27). *Dossier Verstedelijking*. https://www.cbs.nl/nl-nl/dossier/dossier-verstedelijking

Coffie, G. H., & Cudjoe, S. K. F. (2023a). Toward predictive modelling of construction cost overruns using support vector machine techniques. *Cogent Engineering*, *10*(2). https://doi.org/10.1080/23311916.2023.2269656

Coffie, G. H., & Cudjoe, S. K. F. (2023b). Using extreme gradient boosting (XGBoost) machine learning to predict construction cost overruns. *International Journal of Construction Management*. https://doi.org/10.1080/15623599.2023.2289754

Dingemanse, K. (2021, October 26). *Voorbeeld kwalitatief gecodeerd interview*. https://www.scribbr.nl/onderzoeksmethoden/voorbeeld-kwalitatief-gecodeerd-interview/

Eizakshiri, F., Chan, P., & Emsley, M. (2011). Delays, what delays? A critical review of the literature on delays in construction. In *Association of Researchers in Construction Management, ARCOM 2011 - Proceedings of the 27th Annual Conference* (Vol. 2).

Enexis. (2024, May 23). *STRATEGIE & WAARDECREATIE*.

Flyvbjerg, B., Ansar, A., Budzier, A., Buhl, S., Cantarelli, C., Garbuio, M., Glenting, C., Holm, M. S., Lovallo, D., Lunn, D., Molin, E., Rønnest, A., Stewart, A., & van Wee, B. (2018). Five things you should know about cost overrun. *Transportation Research Part A: Policy and Practice*, *118*, 174–190. https://doi.org/10.1016/J.TRA.2018.07.013

Gunduz, M., Naji, K. K., & Al-Sharafi, S. (2024). Assessment of the Critical Dispute Factors in Public-Private Partnership Infrastructure Projects. *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, *16*(4). https://doi.org/10.1061/JLADAH.LADR-1138

Hassaan, I. (2024, October 27). *Shallow Learning vs. Deep Learning: Is Bigger Always Better?* https://medium.com/@hassaanidrees7/shallow-learning-vs-deep-learning-is-bigger-always-better-51c0bd21f059

Huang, C.-H., & Hsieh, S.-H. (2020). Predicting BIM labor cost with random forest and simple linear regression. *Automation in Construction*, *118*. https://doi.org/10.1016/j.autcon.2020.103280

Jensen, K. (2012). *CRISP-DM Process Diagram*.

Khodabakhshian, A., Malsagov, U., & Re Cecconi, F. (2024). Machine Learning Application in Construction Delay and Cost Overrun Risks Assessment. *Lecture Notes in Networks and Systems*, *921 LNNS*, 222–240. https://doi.org/10.1007/978-3-031-54053-0_17

Khoong, W. H. (2023, January 21). *Why Scaling Your Data Is Important*. https://medium.com/codex/why-scaling-your-data-is-important-1aff95ca97a2

Klein, C. (2023, March 4). Enorme toename elektrische voertuigen, maar wat is de klimaatwinst? *NOS*. https://nos.nl/artikel/2466089-enorme-toename-elektrische-voertuigen-maar-wat-is-de-klimaatwinst

Liu, L., & Napier, Z. (2010). The accuracy of risk-based cost estimation for water infrastructure projects: preliminary evidence from Australian projects. *Construction Management and Economics*, *28*(1), 89–100. https://doi.org/10.1080/01446190903431525

Matel, E., Vahdatikhaki, F., Hosseinyalamdary, S., Evers, T., & Voordijk, H. (2022). An artificial neural network approach for cost estimation of engineering services. *International Journal of Construction Management*, *22*(7), 1274–1287. https://doi.org/10.1080/15623599.2019.1692400

Medelyan, A. (2024, September 11). *Coding Qualitative Data: How To Guide*. https://getthematic.com/insights/coding-qualitative-data/

Melin Jr., J. B. (1994). Parametric estimation. *Cost Engineering (Morgantown, West Virginia)*, *36*(1), 19–24. https://www.scopus.com/inward/record.uri?eid=2-s2.0-0028125512&partnerID=40&md5=63d3d313224f0c16af5cdcab41b48188

Naji, K. K., Gunduz, M., & Adalbi, M. (2023). Analysis of Critical Project Success Factors—Sustainable Management of the Fast-Track Construction Industry. *Buildings*, *13*(11). https://doi.org/10.3390/buildings13112890

Narkhede, S. (2018, June 26). *Understanding AUC - ROC Curve*. https://medium.com/towards-data-science/understanding-auc-roc-curve-68b2303cc9c5

nginfra. (2024, May 23). *Verrassingen uit het ondergrondse*.

NOS. (2024). Toezichthouder ACM wil druk op stroomnet verminderen met pakket maatregelen. *2024*. https://nos.nl/artikel/2517302-toezichthouder-acm-wil-druk-op-stroomnet-verminderen-met-pakket-maatregelen

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Systematic Reviews*, *10*(1), 89. https://doi.org/10.1186/s13643-021-01626-4

Poh, C. Q. X., Ubeynarayana, C. U., & Goh, Y. M. (2018). Safety leading indicators for construction sites: A machine learning approach. *Automation in Construction*, *93*, 375–386. https://doi.org/10.1016/j.autcon.2018.03.022

Potkamp, H. (2024). *Photo of pipes and cables*.

Sanni-Anibire, M. O., Zin, R. M., & Olatunji, S. O. (2021). Machine learning - Based framework for construction delay mitigation. *Journal of Information Technology in Construction*, *26*, 303–318. https://doi.org/10.36680/j.itcon.2021.017

Scikit-learn. (2025, April 2). *Scikit-learn*.

Seol, D., Choi, J., Kim, C., & Hong, S. (2023). Alleviating Class-Imbalance Data of Semiconductor Equipment Anomaly Detection Study. *Electronics*, *12*, 585. https://doi.org/10.3390/electronics12030585

Shah, S., & Gopinath, S. (2024). Machine Learning-Based Dynamic Cost Estimation Model for Construction Projects. *Lecture Notes in Civil Engineering*, *388*, 625–633. https://doi.org/10.1007/978-981-99-6233-4_56

Sinkovics, N. (2018). *Pattern matching in qualitative analysis* (pp. 468–485).

Tajziyehchi, N., Moshirpour, M., Jergeas, G., & Sadeghpour, F. (2020). A Predictive Model of Cost Growth in Construction Projects Using Feature Selection. *Proceedings - 2020 IEEE 3rd International Conference on Artificial Intelligence and Knowledge Engineering, AIKE 2020*, 142–147. https://doi.org/10.1109/AIKE48582.2020.00029

Tummers, S. C. M. W., Hommersom, A., Bolman, C., Lechner, L., & Bemelmans, R. (2025). A new data science trajectory for analysing multiple studies: a case study in physical activity research. *MethodsX*, *14*, 103104. https://doi.org/10.1016/J.MEX.2024.103104

Yamany, M. S., Abdelhameed, A., Elbeltagi, E., & Mohamed, H. A. E. (2024). Critical success factors of infrastructure construction projects. *Innovative Infrastructure Solutions*, *9*(4). https://doi.org/10.1007/s41062-024-01394-9

Appendix A: Features of each code group

Project related	People related	External related	Cost related	Time related	Tender/contract related	Management related	Design related
amount of concrete	Client	Absence of data	Actual cost	Consecutive	bureaucracy in tendering	Cost management	design complexity
architectural properties	Client's experience	Competitors	Additional cost	Decision making delay	contract procedures	Management quality	Design/scope change
Computer technology used in design stage	Collaboration of designer and contractor	Contamination of ground	Awarded target cost	Delay between design and bid periods	Contract type	Relation management and labour	lack of flexibility in design
construction errors	Disputes on site	Currency exchange	Cost fluctuation	delay in approving drawings	Discrepancies in tender document	risk management	pre-contract design
Delivery strategy	Experience of (sub)contractor	Economic instability	Cost of insurance	delay in information	Lowest bid problems	Subcontractor management	
equipment	Experience of project manager	Existing utilities	cost of reworks	delay in supply	Procurement		
Excavation depth	Fraud	Force majeure	Cost/time overrun	delay in work	Tender strategy		
Execution method	Inadequate labour	government polices	Excavation cost	initial duration			
Inadequate material quantity	Inadequate site investigation	Impact of foreign companies	financial status	Payment delay			
Insufficient / changed material type	Lack of contractor performance	inflation and taxes	Initial cost	Planning quality			
Length of pipe	Level of knowledge/expe-rience consultant/desig ner	Interest rate	Installation cost	Project duration			

Location	Level of quality control	Market conditions	Labour cost				
name of project	level of understanding contract	number of parallel projects	Machinery cost				
Number of manholes	multidisciplinarity	pandemic	material cost				
Pipe diameter	productivity	Season	Profit				
progress percentage of construction	Project leadership	Social and cultural impacts	Project financing				
Project phases	Quality of communication	Stakeholders	Quality of cost estimation				
Project size	Quality of financial control	Type of soil	Testing cost				
project status	size project team	Underground water	Transport cost				
Safety	Supplier manipulation	weather	Type of finance				
Space to work	Unrealistic expectations						
Type of formwork	worked hours per week						
Type of material							
Type of work							
wastage							
Worksite							

Table 12: Features per feature group

Appendix B: Influential features from literature

#	Influential factors	Abbas & Aswed (2024)	Ahiaga-Dabuii & Smith (2014)	Akinradewo et al. (2019)	Al mnaseer et al. (2023)	Almahameed & Bisharah (2024)	Al-Nahhas et al. (2024)	Alsugair et al. (2024)	Ariyawansha & Francis (2023)	Arabiat et al. (2024)	Aung et al. (2022)	Cheng & Hoang (2023)	Cheng & Hoang (2014)	Coffie & Cudjoe (2024)	Coffie & Cudjoe (1) (2023)	Coffie & Cudjoe (2) (2023)	Das et al. (2023)	Johnson & Babu (2024)	Khodabakhshian et al. (2020)	Khodabakhshian et al. (2024)	Kuljaroenwirat & Seresangtakul (2016)	Lee & Yun (2024)	Leu et al. (1) (2023)	Leu et al. (2) (2023)	Matel et al. (2022)	Mhady et al. (2024)	Safaeian Hamzehkolaei & Alizamir (2022)	Obianyo et al. (2022)	Rasheed & Rezouki (2022)	Salleh et al. (2021)	Shah & Gopinath (2023)	Sharma et al. (2024)	Shoar et al. (2020)	Sohu et al. (2022)	Taiyzyehchi et al. (2020)	Ujong et al. (2020)	Yaqubi & Salhotra (2022)	Zahmak et al. (2019)	Zahmak et al. (2020)	Quantity
1	Project duration	●	●	●							●		●	●	●						●	●					●					●			●						12
2	Initial cost					●					●		●	●	●	●	●				●	●						●		●						●						12
3	Design/scope change			●				●						●		●			●						●	●	●	●	●							●					●	11
4	Planning quality			●				●	●		●								●								●	●		●						●	●				●	11
5	Management quality							●																	●	●			●			●	●	●	●		●	●			●	10
6	weather							●			●			●											●			●	●			●	●				●				●	10
7	Experience of (sub)contractor			●				●	●		●																				●		●				●	●		●	9
8	Actual cost				●				●					●	●		●					●							●					●						●	●	9
9	Location		●					●					●					●				●	●									●	●								●	9
10	architectural properties														●	●	●						●									●	●						●	●		8
11	Project size							●					●										●	●			●			●	●							●				8
12	Qualtity of cost estimation							●	●		●										●	●										●				●					●	8
13	Inadequate labour			●				●	●																				●				●	●		●				●	8
14	Knowledge/experience consultant/designer							●			●															●			●			●	●	●	●						8
15	initial duration									●					●	●						●	●							●								●			●	8
16	Worksite	●						●			●				●								●						●			●						●				7
17	Type of work		●										●										●				●					●		●				●				7
18	Discrepancies in tender document							●	●		●																		●			●				●				●	7
19	contract procedures			●				●	●						●			●							●				●												7
20	Decision making delay							●			●									●									●				●			●				●	7
21	Market conditions							●			●															●	●					●		●		●					7
22	Client		●																				●				●						●	●	●	●		●			7
23	Quality of financial control			●					●		●																		●			●	●							●	7
24	Payment delay							●			●												●						●				●				●				6
25	Labour cost			●							●	●															●											●			●	6
26	Procurement		●	●							●									●						●												●			6
27	material cost			●			●				●	●																										●			5
28	Subcontractor management															●										●	●		●									●			5
29	Tender strategy		●								●																		●			●								●	5
30	financial status							●			●									●									●								●				5
31	Inadequate material quantity			●							●																		●				●							●	5
32	productivity			●				●																		●			●											●	5
33	Type of soil	●	●																											●	●									●	5
34	Cost/time overrun				●				●														●			●															4
35	government polices			●				●																					●												●	4
36	Additional cost			●																																		●	●	●	●	4
37	delay in work							●		●										●										●												4
38	Disputes on site			●				●			●																													●	4
39	Experience of project manager							●															●				●									●						4
40	Quality of communication							●			●																									●	●					4
41	Safety							●			●																											●			●	4
42	Social and cultural impacts			●				●			●																													●	4
43	Type of material	●																								●					●				●						4
44	delay in supply							●			●																		●												3
45	Space to work							●																●															●			3

#	Influential factors	Abbas & Aswed (2024)	Ahiaga-Dabuii & Smith (2014)	Akinradewo et al. (2019)	Almahameed & Bisharah (2024)	Al mnaseer et al. (2023)	Al-Nahhas et al. (2024)	Alsugair et al. (2024)	Ariyawansha & Francis (2022)	Arrabat et al. (2023)	Aung et al. (2023)	Cheng & Hoang (2014)	Cheng & Khasani (2024)	Coffie & Cudjoe (1) (2023)	Coffie & Cudjoe (2) (2023)	Das et al. (2024)	Johnson & Babu (2020)	Khodabakhshian et al. (2024)	Kuljaroenwirat & Seresangtakul (2016)	Lee & Yun (2024)	Leu et al. (1) (2023)	Leu et al. (2) (2023)	Matel et al. (2022)	Mhady et al. (2024)	Obianyo et al. (2022)	Rasheed & Rezouki (2022)	Safaeian Hamzehkolaei & Alizamir (2021)	Salleh et al. (2023)	Shah & Gopinath (2024)	Sharma et al. (2020)	Shoar et al. (2022)	Sohu et al. (2020)	Tajziyehchi et al. (2020)	Ujong et al. (2022)	Yaqubi & Salhotra (2019)	Zahmak et al. (2020)	Quantity
46	Stakeholders								●		●															●											3
47	Unrealistic expectations								●		●															●											3
48	Absence of data				●						●																									●	3
49	Client's experience																		●					●									●			3
50	Contract type												●							●				●												3
51	Cost fluctuation													●												●									●	3
52	Economic instability				●																								●						●	3
53	equipment																					●								●	●					3
54	inflation and taxes								●		●																									●	3
55	Profit							●																			●								●		3
56	Type of finance	●			●																															●	3
57	Competitors								●																											●	2
58	Computer technology used in desing stage																																●		●		2
59	amount of concrete																	●											●								2
60	construction errors																									●				●							2
61	Cost management													●																					●		2
62	Currency exchange				●						●																										2
63	Delay between design and bid periods				●				●																												2
64	Delivery strategy								●																										●		2
65	Force majeure																						●									●					2
66	Inadequate site investigation				●																					●											2
67	Interest rate				●				●																												2
68	Level of quality control								●																								●				2
69	level of understanding contract																		●														●				2
70	Machinery cost				●																															●	2
71	number of paralel projects								●																											●	2
72	pre-contract design																							●								●					2
73	project status																				●		●														2
74	Relation management and labour				●																															●	2
75	risk management																		●							●										2
76	size project team																							●									●				2
77	Supplier manipulation				●																															●	2
78	Transport cost				●																															●	2
79	wastage																																		●	●	2
80	Project financing				●																																1
81	Awarded target cost		●																																		1
82	bureaucracy in tendering				●																																1
83	Collaboration of designer and contractor				●																																1
84	Consecutive				●																																1
85	Contamination of ground		●																																		1
86	Cost of insurance				●																																1
87	cost of reworks										●																										1
88	delay in approving drawings								●																												1
89	delay in information										●																										1
90	design complexity																												●								1

#	Influential factors	Abbas & Aswed (2024)	Ahiaga-Dabuii & Smith (2014)	Akinradewo et al. (2019)	Almahameed & Bisharah (2024)	Al mnaseer et al. (2023)	Al-Nahhas et al. (2024)	Alsugair et al. (2024)	Ariyawansha & Francis (2022)	Arabiat et al. (2023)	Aung et al. (2023)	Cheng & Hoang (2014)	Cheng & Khasani (2024)	Coffie & Cudjoe (1) (2023)	Coffie & Cudjoe (2) (2023)	Das et al. (2024)	Johnson & Babu (2020)	Khodabakhshian et al. (2024)	Kuljaroenwirat & Seresangtakul (2016)	Lee & Yun (2024)	Leu et al. (1) (2023)	Leu et al. (2) (2023)	Matel et al. (2022)	Mhady et al. (2024)	Obianyo et al. (2022)	Rasheed & Rezouki (2022)	Safaeian Hamzehkolaei & Alizamir (2021)	Salleh et al. (2023)	Shah & Gopinath (2024)	Sharma et al. (2020)	Shoar et al. (2022)	Sohu et al. (2020)	Tajziyehchi et al. (2020)	Ujong et al. (2022)	Yaqubi & Salhotra (2019)	Zahmak et al. (2020)	Quantity
91	Excavation cost							●																												1
92	Excavation depth	●																																		1
93	Execution method				●																															1
94	Existing utilities																																	●		1
95	Fraud				●																															1
96	Impact of foreign companies										●																									1
97	Installation cost							●																												1
98	Insufficient / changed material type								●																											1
99	Lack of contractor performance																																	●		1
100	lack of flexibility in design																			●																1
101	Length of pipe	●																																		1
102	Lowest bid problems										●																									1
103	multidisciplinarity																							●												1
104	name of project																			●																1
105	Number of manholes	●																																		1
106	pandemic										●																									1
107	Pipe diameter	●																																		1
108	progress percentage of construction																									●										1
109	Project leadership	●																																		1
110	Project phases																							●												1
111	Season																			●																1
112	Testing cost										●																									1
113	Type of formwork																	●																		1
114	Underground water	●																																		1
115	worked hours per week																							●												1

Table 13: Features mentioned by authors

Appendix C: Influential features from interviews

#	Influential factors	Interview 1	Interview 2	Interview 3	Interview 4	Interview 5	Interview 6	Interview 7	Interview 8	Interview 9	Quantity
1	Existing utilities	●	●	●	●		●	●	●	●	7
2	Kind of utility	●	●	●		●	●	●	●	●	7
3	preperation	●	●	●	●	●	●		●	●	7
4	Rain	●	●	●	●		●	●	●		7
5	Type of soil	●	●	●	●	●		●	●		7
6	Temperature	●	●	●	●			●	●		6
7	urbanity of an area	●		●		●	●	●	●	●	6
8	Watch project on forehand	●	●	●	●	●		●			6
9	Experience of workpeople	●			●		●	●	●	●	5
10	Size of project		●		●		●	●	●	●	5
11	Space to work	●		●		●	●		●		5
12	Type of contract		●		●	●		●		●	4
13	client				●	●	●				3
14	Diameter						●	●	●		3
15	Length of project					●	●		●		3
16	Motivation of workpeople		●	●				●			3
17	Subcontractors				●	●			●	●	3
18	Communication between parties						●		●		2
19	Connections to houses							●	●	●	2
20	Extern supervisor	●	●								2
21	Not skilled people	●							●	●	2
22	Number of drillings							●	●		2
23	type of subcontractor					●					1
24	type of work							●			1

Table 14: Influential features from interviews

Appendix D: Pattern matching

#	Literature	Interview	Added	Features from literature	Features from interviews	Added by researcher
1	●	●		Project duration	Length of project	
2	●			Initial cost		
3	●			Design/scope change		
4	●			Planning quality		
5	●	●		Management quality	preperation	
6	●	●		weather	Rain Temperature	
7	●			Experience of (sub)contractor		
8	●			Actual cost		
9	●			Location		
10	●			architectural properties		
11	●	●		Project size	Size of project	
12	●			Qualtity of cost estimation		
13	●	●	●	Inadequate labour	Motivation of workpeople Not skilled people Experience of workpeople	Group of workpeople that conducted work
14	●			Knowledge/experience consultant/designer		
15	●			initial duration		
16	●			Worksite		
17	●	●		Type of work	type of work	
18	●			Discrepancies in tender document		
19	●			contract procedures		
20	●			Decision making delay		
21	●			Market conditions		
22	●	●		Client	client	
23	●			Quality of financial control		
24	●			Payment delay		
25	●			Labour cost		

#	Literature	Interview	Added	Features from literature	Features from interviews	Added by researcher
26	●			Procurement		
27	●			material cost		
28	●	●		Subcontractor management	type of subcontractor Subcontractors	
29	●			Tender strategy		
30	●			financial status		
31	●			Inadequate material quantity		
32	●			productivity		
33	●	●		Type of soil	Type of soil	
34	●			Cost/time overrun		
35	●			government polices		
36	●			Additional cost		
37	●			delay in work		
38	●			Disputes on site		
39	●			Experience of project manager		
40	●	●		Quality of communication	Communication between parties	
41	●			Safety		
42	●			Social and cultural impacts		
43	●			Type of material		
44	●			delay in supply		
45	●	●		Space to work	Space to work	
46	●			Stakeholders		
47	●			Unrealistic expectations		
48	●			Absence of data		
49	●			Client's experience		
50	●	●		Contract type	Type of contract	
51	●			Cost fluctuation		
52	●			Economic instability		
53	●			equipment		
54	●			inflation and taxes		
55	●			Profit		

#	Literature	Interview	Added	Features from literature	Features from interviews	Added by researcher
56	●			Type of finance		
57	●			Competitors		
58	●			Computer technology used in desing stage		
59	●			amount of concrete		
60	●			construction errors		
61	●			Cost management		
62	●			Currency exchange		
63	●			Delay between design and bid periods		
64	●			Delivery strategy		
65	●			Force majeure		
66	●	●		Inadequate site investigation	Watch project on forehand	
67	●			Interest rate		
68	●			Level of quality control		
69	●			level of understanding contract		
70	●			Machinery cost		
71	●			number of paralel projects		
72	●			pre-contract design		
73	●			project status		
74	●			Relation management and labour		
75	●			risk management		
76	●			size project team		
77	●			Supplier manipulation		
78	●			Transport cost		
79	●			wastage		
80	●			Project financing		
81	●			Awarded target cost		
82	●			bureaucracy in tendering		
83	●			Collaboration of designer and contractor		
84	●			Consecutive		
85	●			Contamination of ground		

#	Literature	Interview	Added	Features from literature	Features from interviews	Added by researcher
86	●			Cost of insurance		
87	●			cost of reworks		
88	●			delay in approving drawings		
89	●			delay in information		
90	●			design complexity		
91	●			Excavation cost		
92	●			Excavation depth		
93	●			Execution method		
94	●	●		Existing utilities	Existing utilities urbanity of an area	
95	●			Fraud		
96	●			Impact of foreign companies		
97	●			Installation cost		
98	●			Insufficient / changed material type		
99	●			Lack of contractor performance		
100	●			lack of flexibility in design		
101	●			Length of pipe		
102	●			Lowest bid problems		
103	●			multidisciplinarity		
104	●			name of project		
105	●			Number of manholes		
106	●			pandemic		
107	●			Pipe diameter	Diameter	
108	●			progress percentage of construction		
109	●	●		Project leadership	Extern supervisor	
110	●			Project phases		
111	●			Season		
112	●			Testing cost		
113	●			Type of formwork		
114	●			Underground water		
115	●			worked hours per week		

#	Literature	Interview	Added	Features from literature	Features from interviews	Added by researcher
116		●			Kind of utility	
117		●			Connections to houses	
118		●			Number of drillings	
119			●			Initial scheduled work hours
120			●			Length cable per utility type
121			●			Actual project size
122			●			Money billed to client

Table 15: Pattern matching

Appendix E: Data availability per feature

#	Feature	Origin	Additional description	Coding	Unit	Feature usable	Comment
1	Project duration	Literature/Interviews	Actual duration of project	int	days	yes	Data not directly avaiable, but could be calculated
2	Initial cost	Literature		int	euros	no	The data is not available on a large standardized scale
3	Design/scope change	Literature	Number of scope changes	int	-	no	Data not available
4	Planning quality	Literature		string	-	no	Feature cannot be categorized or quantified
5	Management quality	Literature		string	-	no	Not this feature, is broken down in more (detailed) feature(s)
6	preperation	Interviews	Hours of preperation	float	euros	yes	Data is usable and available
7	weather	Literature		string	-	no	Not this feature, is broken down in more (detailed) feature(s)
8	Rain	Interviews		float	millimeter	yes	Data is usable and available
9	Temperature	Interviews	Avarage temperature over project	float	degrees Celcius	yes	Data is usable and available
10	Experience of (sub)contractor	Literature		int	years	no	Data not available
11	Actual cost	Literature		int	euros	yes	Data not directly avaiable, but could be calculated
12	Location	Literature	The geograhpical location of the project	float	coordinates/zipcode/adress	yes	Data is usable and available
13	architectural properties	Literature	Architectural properties of a building	string	-	no	Not relevant for utility projects in NL
14	Project size	Literature/Interviews	Initial length of trench	int	meters	no	Data not available
15	Qualtity of cost estimation	Literature		string	-	no	Feature cannot be categorized or quantified
16	Inadequate labour	Literature		String	-	no	Feature cannot be categorized or quantified
17	Motivation of workpeople	Interviews		string	-	no	Feature cannot be categorized or quantified
18	Not skilled people	Interviews		string	-	no	Feature cannot be categorized or quantified

#	Feature	Origin	Additional description	Coding	Unit	Feature usable	Comment
19	Experience of workpeople	Interviews		int	years	no	Data not available
20	Group of workpeople that conducted work	Researcher		int	-	yes	Data is usable and available
21	Knowledge/experience consultant/designer	Literature		int	years	no	Data not available
22	initial duration	Literature		int	days	no	The data is not available on a large standardized scale
23	Worksite	Literature	Properties of the worksite	string	-	no	Feature cannot be categorized or quantified
24	Type of work	Literature/Interviews	The type of work (e.g. new build, removal)	string	-	no	Data not available
25	Discrepancies in tender document	Literature		string	-	no	Feature cannot be categorized or quantified
26	contract procedures	Literature		string	-	no	Feature cannot be categorized or quantified
27	Decision making delay	Literature		int	days	no	Data not available
28	Market conditions	Literature		string	-	no	Feature cannot be categorized or quantified
29	Client	Literature		string	-	yes	Data is usable and available
30	Quality of financial control	Literature/Interviews		string	-	no	Feature cannot be categorized or quantified
31	Payment delay	Literature		int/boolean	day/-	no	Data not available
32	Labour cost	Literature	Cost of the workpeople who carry out the work	int	euros	yes	Data is usable and available
33	Procurement	Literature	Procurement process	string	-	no	Feature cannot be categorized or quantified
34	material cost	Literature		int	euros	yes	Data is usable and available
35	Subcontractor management	Literature		string	-	no	Not this feature, is broken down in more (detailed) feature(s)
36	type of subcontractor	Interviews	Type of contract with the subcontractor (per hour or contract)	string	-	no	Data not available
37	Subcontractors	Interviews	Amount of subcontractors on a project	float	euros	yes	Data is usable and available

#	Feature	Origin	Additional description	Coding	Unit	Feature usable	Comment
38	Tender strategy	Literature		string	-	no	Feature cannot be categorized or quantified
39	financial status	Literature	Financial status of the company	string	-	no	Feature cannot be categorized or quantified
40	Inadequate material quantity	Literature		boolean	-	no	Data not available
41	productivity	Literature	Productivity of the workpeople	float	meters per day	yes	Data not directly avaiable, but could be calculated
42	Type of soil	Literature/Interviews		string	-	no	Data not available
43	Cost/time overrun	Literature		boolean	-	yes	Data not directly avaiable, but could be calculated
44	government polices	Literature		string	-	no	Feature cannot be categorized or quantified
45	Additional cost	Literature	Difference in cost between initial cost and actual cost	int	euros	no	Data not available
46	delay in work	Literature		int	days	no	The data is not available on a large standardized scale
47	Disputes on site	Literature		string	-	no	Data not available
48	Experience of project manager	Literature		string	-	yes	Data is usable and available
49	Quality of communication	Literature		string	-	no	Feature cannot be categorized or quantified
50	Communication between parties	Interviews		string	-	no	Feature cannot be categorized or quantified
51	Safety	Literature	Safety on the worksite	string	-	no	Feature cannot be categorized or quantified
52	Social and cultural impacts	Literature		string	-	no	Feature cannot be categorized or quantified
53	Type of material	Literature		string	-	no	Data not available
54	delay in supply	Literature		int	days	no	Data not available
55	Space to work	Literature/Interviews		float	square meters	no	Data not available
56	Stakeholders	Literature	Amount of stakeholders	int	-	no	Data not available
57	Unrealistic expectations	Literature		boolean	-	no	Data not available
58	Absence of data	Literature		boolean	-	no	Feature is irrelevant
59	Client's experience	Literature		int	years	no	Data not available

#	Feature	Origin	Additional description	Coding	Unit	Feature usable	Comment
60	Contract type	Literature/Interviews	contract based on total work or per meter	string	-	no	Data not available
61	Cost fluctuation	Literature	standard deviation of cost fluctuation over project	float	-	no	Data not available
62	Economic instability	Literature		string	-	no	Feature cannot be categorized or quantified
63	equipment	Literature	Equipment used on worksite	string	-	no	Data not available
64	inflation and taxes	Literature		int	euros	no	Not relevant for utility projects in NL
65	Profit	Literature	Profit on a project	int	euros	yes	Data is usable and available
66	Type of finance	Literature	In how many installments have been paid	int	-	no	Not relevant for utility projects in NL
67	Competitors	Literature	How many competitors bid along	int	-	no	Data not available
68	Computer technology used in desing stage	Literature		string	-	no	Feature cannot be categorized or quantified
69	amount of concrete	Literature	Amount of concrete used in the project	float	cubic meters	no	Not relevant for utility projects in NL
70	construction errors	Literature		int	-	no	The data is not available on a large standardized scale
71	Cost management	Literature		string	-	no	Feature cannot be categorized or quantified
72	Currency exchange	Literature	If there was a currency exchange	boolean	-	no	Not relevant for utility projects in NL
73	Delay between design and bid periods	Literature		int	days	no	Data not available
74	Delivery strategy	Literature		string	-	no	Feature cannot be categorized or quantified
75	Force majeure	Literature		string	-	no	Feature cannot be categorized or quantified
76	Inadequate site investigation	Literature		string	-	no	Not this feature, is broken down in more (detailed) feature(s)
77	Watch project on forehand	Interviews	If there is watched on forehand	boolean	-	no	Data not available
78	Interest rate	Literature	Interest rate of finance of project	int	percent	no	Data not available
79	Level of quality control	Literature		string	-	no	Feature cannot be categorized or quantified
80	level of understanding contract	Literature		string	-	no	Feature cannot be categorized or quantified
81	Machinery cost	Literature		int	euros	yes	Data is usable and available
82	number of paralel projects	Literature	How many projects are executed at the same time	int	-	no	Data not available
83	pre-contract design	Literature		string	-	no	Feature cannot be categorized or quantified

#	Feature	Origin	Additional description	Coding	Unit	Feature usable	Comment
84	project status	Literature		string	-	no	Feature cannot be categorized or quantified
85	Relation management and labour	Literature		string	-	no	Feature cannot be categorized or quantified
86	risk management	Literature		string	-	no	Feature cannot be categorized or quantified
87	size project team	Literature	Number of workpeople on project	int	-	no	Data not available
88	Supplier manipulation	Literature	If a supplier is manipulated	boolean	-	no	Data not available
89	Transport cost	Literature		int	euros	no	Not this feature, is broken down in more (detailed) feature(s)
90	wastage	Literature		int	kilo's	no	Data not available
91	Project financing	Literature		string	-	no	Feature cannot be categorized or quantified
92	Awarded target cost	Literature	Expected earnings/billed money to client	int	euros	yes	Data is usable and available
93	bureaucracy in tendering	Literature		string	-	no	Feature cannot be categorized or quantified
94	Collaboration of designer and contractor	Literature		string	-	no	Feature cannot be categorized or quantified
95	Consecutive	Literature	Are the different projects consecutive / standalone project	boolean	-	no	Data not available
96	Contamination of ground	Literature		boolean	-	no	The data is not available on a large standardized scale
97	Cost of insurance	Literature		int	euros	no	Data not available
98	cost of reworks	Literature		int	euros	no	Not this feature, is broken down in more (detailed) feature(s)
99	delay in approving drawings	Literature		int	days	no	Data not available
100	delay in information	Literature		int	days	no	Data not available
101	design complexity	Literature		string	-	no	Feature cannot be categorized or quantified
102	Excavation cost	Literature		int	euros	no	Not this feature, is broken down in more (detailed) feature(s)
103	Excavation depth	Literature		float	meters	no	Not this feature, is broken down in more (detailed) feature(s)
104	Execution method	Literature		string	-	no	Feature cannot be categorized or quantified
105	Existing utilities	Literature/Interviews		boolean	-	no	Data not available

#	Feature	Origin	Additional description	Coding	Unit	Feature usable	Comment
106	urbanity of an area	Interviews		int	-	yes	Data is usable and available
107	Fraud	Literature	If there is fraud detected	boolean	-	no	Data not available
108	Impact of foreign companies	Literature		string	-	no	Not relevant for utility projects in NL
109	Installation cost	Literature		int	euros	no	Not this feature, is broken down in more (detailed) feature(s)
110	Insufficient / changed material type	Literature		boolean	-	no	Data not available
111	Lack of contractor performance	Literature		string	-	no	Feature cannot be categorized or quantified
112	lack of flexibility in design	Literature		string	-	no	Feature cannot be categorized or quantified
113	Length of pipe	Literature		int	meters	no	Not this feature, is broken down in more (detailed) feature(s)
114	Lowest bid problems	Literature		string	-	no	Feature cannot be categorized or quantified
115	multidisciplinarity	Literature		string	-	no	Feature cannot be categorized or quantified
116	name of project	Literature	Name or unique code for project	string	-	yes	Data is usable and available
117	Number of manholes	Literature		int	-	no	Not relevant for utility projects in NL
118	pandemic	Literature	If the pandemic was active	boolean	-	yes	Data not directly avaiable, but could be calculated
119	Pipe diameter	Literature/Interviews	Diameter of pipe	int	millimeters	no	Data not available
120	progress percentage of construction	Literature		int	percentage	no	Data not available
121	Project leadership	Literature		string	-	no	Not this feature, is broken down in more (detailed) feature(s)
122	Extern supervisor	Interviews	Quality of extern supervisor	string	-	no	Feature cannot be categorized or quantified
123	Project phases	Literature		int	-	no	Data not available
124	Season	Literature	Season in which project is conducted	string	-	yes	Data not directly avaiable, but could be calculated
125	Testing cost	Literature	Cost for testing the asset	int	euros	no	Not relevant for utility projects in NL
126	Type of formwork	Literature		string	-	no	Not relevant for utility projects in NL

#	Feature	Origin	Additional description	Coding	Unit	Feature usable	Comment
127	Underground water	Literature	Drainage used	boolean	-	yes	Data is usable and available
128	worked hours per week	Literature		int	hours	no	Data not available
129	Kind of utility	Interviews		string	-	yes	Data is usable and available
130	Connections to houses	Interviews		int	-	yes	Data is usable and available
131	Number of drillings	Interviews		int	-	yes	Data is usable and available
132	Initial scheduled work hours	Researcher		int	-	no	Data not available
133	Length cable per utility type	Researcher		int	meters	yes	Data is usable and available
134	Actual project size	Researcher		string	-	no	Not this feature, is broken down in more (detailed) feature(s)
135	Actual project size	Researcher	Actual length of trench	int	meters	yes	Data is usable and available
136	Money billed to client	Researcher		int	euros	yes	Data is usable and available

Table 16: Data availability

Appendix F: Additional information about available data

#	Feature	Feature usable	Comment	Description of actual data	Category	Owner	Interface	Format	Granularity	Granularity level	Availability	Reliability	Link
1	Project duration	yes	Data not directly avaiable, but could be calculated	The difference in days between the first day and the	numerical				per project	3			
6	preperation	yes	Data is usable and available	Cost of preperation works	numerical	Bam	file	.xlsx	per project	3	private	3	SAP
8	Rain	yes	Data is usable and available	Per day the total sum of rain, measured in a metereological staion in The Netherlands	numerical	KNMI	file	.txt	daily	3	open source	3	https://www.knmi.nl/nederland-nu/klimatologie/daggegevens
9	Temperature	yes	Data is usable and available	Mean temperature in 24h	numerical	KNMI	file	.txt	daily	3	open source	3	https://www.knmi.nl/nederland-nu/klimatologie/daggegevens
11	Actual cost	yes	Data not directly avaiable, but could be calculated	Turnover-profit	numerical				per project	3			
12	Location	yes	Data is usable and available	Street, number, town, province are added	textual	Bam	Webservic	.xlsx	Village-base	2	private	3	https://digiflow.baminfra.nl/
20	Group of workpeople that conducted work	yes	Data is usable and available	All building executors	textual	Bam	Webservic	.xlsx	per project	3	private	3	https://digiflow.baminfra.nl/
29	Client	yes	Data is usable and available	Client per project. When project has more clients,	textual	Bam	Webservic	.xlsx	per project	3	private	3	https://digiflow.baminfra.nl/
32	Labour cost	yes	Data is usable and available	The cost of workpeople made	numerical	Bam	file	.xlsx	per project	3	private	3	SAP
34	material cost	yes	Data is usable and available	The cost of material made on the project	numerical	Bam	file	.xlsx	per project	3	private	3	SAP
37	Subcontractors	yes	Data is usable and available	The amount of money spend on subcontractors	numerical	Bam	file	.xlsx	per project	3	private	3	SAP
41	productivity	yes	Data not directly avaiable, but could be calculated	Is productivity per day for all specialisations of the project because trial length is also given in meter times specialisation.	numerical				per project	3			
43	Cost/time overrun	yes	Data not directly avaiable, but could	If the profit was negative, this	binary				per project	3			
48	Experience of project manager	yes	Data is usable and available	Not the experience, but the names of the projectleaders are added	textual	Bam	Webservic	.xlsx	per project	3	private	3	https://digiflow.baminfra.nl/
65	Profit	yes	Data is usable and available	Per project the financial result.	numerical	Bam	file	.xlsx	per project	3	private	3	PowerBI sheets
81	Machinery cost	yes	Data is usable and available	This is part of the added cost categories.	numerical	Bam	file	.xlsx	per project	3	private	3	SAP
92	Awarded target cost	yes	Data is usable and available	Calculation sum (expected amount to bill)	numerical	Bam	Webservic	.xlsx	per project	3	private	3	https://digiflow.baminfra.nl/

#	Feature	Feature usable	Comment	Description of actual data	Category	Owner	Interface	Format	Granularity	Granularity level	Availability	Reliability	Link
106	urbanity of an area	yes	Data is usable and available	Per zip code, the urbanity is given in numbers ranging from 1 (very urban) to 5 (not urban).	numerical	CBS	file	.xlsx	pc5 zipcode	3	open source	3	https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/gegevens-per-postcode
116	name of project	yes	Data is usable and available	Project codes and descriptions of project	numerical	Bam	file	.xlsx	per project	3	private	3	PowerBI sheets
118	pandemic	yes	Data not directly avaiable, but could be calculated	According to the government of The Netherlands, the pandemic was between march 2020 and march 2022. If the center date of the execution timespan was between these dates, this value was set to 1.	binary				per project	3			
124	Season	yes	Data not directly avaiable, but could be calculated	This is calculated using middle dates of execution timespan checking in what season the date was.	texutal				per project	3			
127	Underground water	yes	Data is usable and available	If drainage is used on a project	textual	Bam	file	.xlsx	per project	3	private	3	SAP
129	Kind of utility	yes	Data is usable and available	The discipline of utility involved in the project.	textual	Bam	Webservic	.xlsx	per project	3	private	3	https://digiflow.baminfra.nl/
130	Connections to houses	yes	Data is usable and available	How many house connections are done in the project.	textual	Bam	file	.xlsx	per project	3	private	3	SAP
131	Number of drillings	yes	Data is usable and available	Difference is made between rocket and steered drilling	textual	Bam	file	.xlsx	per project	3	private	3	SAP
133	Length cable per utility type	yes	Data is usable and available	Elektra LS, Elektra MS, Gas LD, Gas HD, Water are added	numerical	Bam	Webservic	.xlsx	per project	3	private	3	https://digiflow.baminfra.nl/
135	Actual project size	yes	Data is usable and available	Is per meter times discipline	numerical	Bam	Webservic	.xlsx	per project	3	private	3	https://digiflow.baminfra.nl/
136	Money billed to client	yes	Data is usable and available	Actual turnover per project	numerical	Bam	file	.xlsx	per project	3	private	3	PowerBI sheets

Table 17: Information about available data

Appendix G: Feature importance

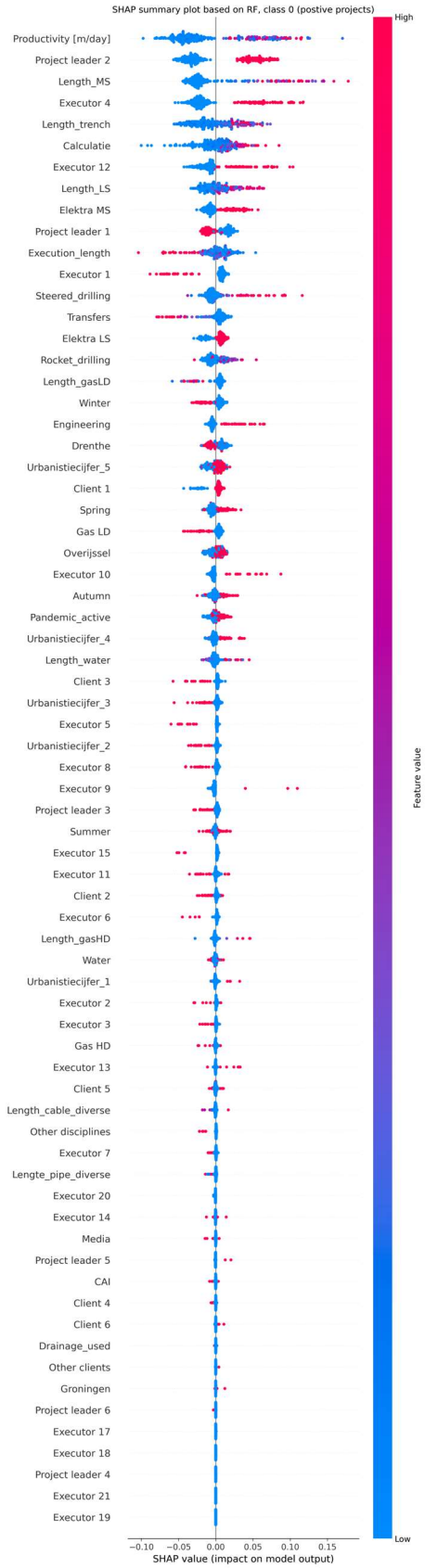

Figure 5: Feature importance of all features

Appendix H: Results of RFE

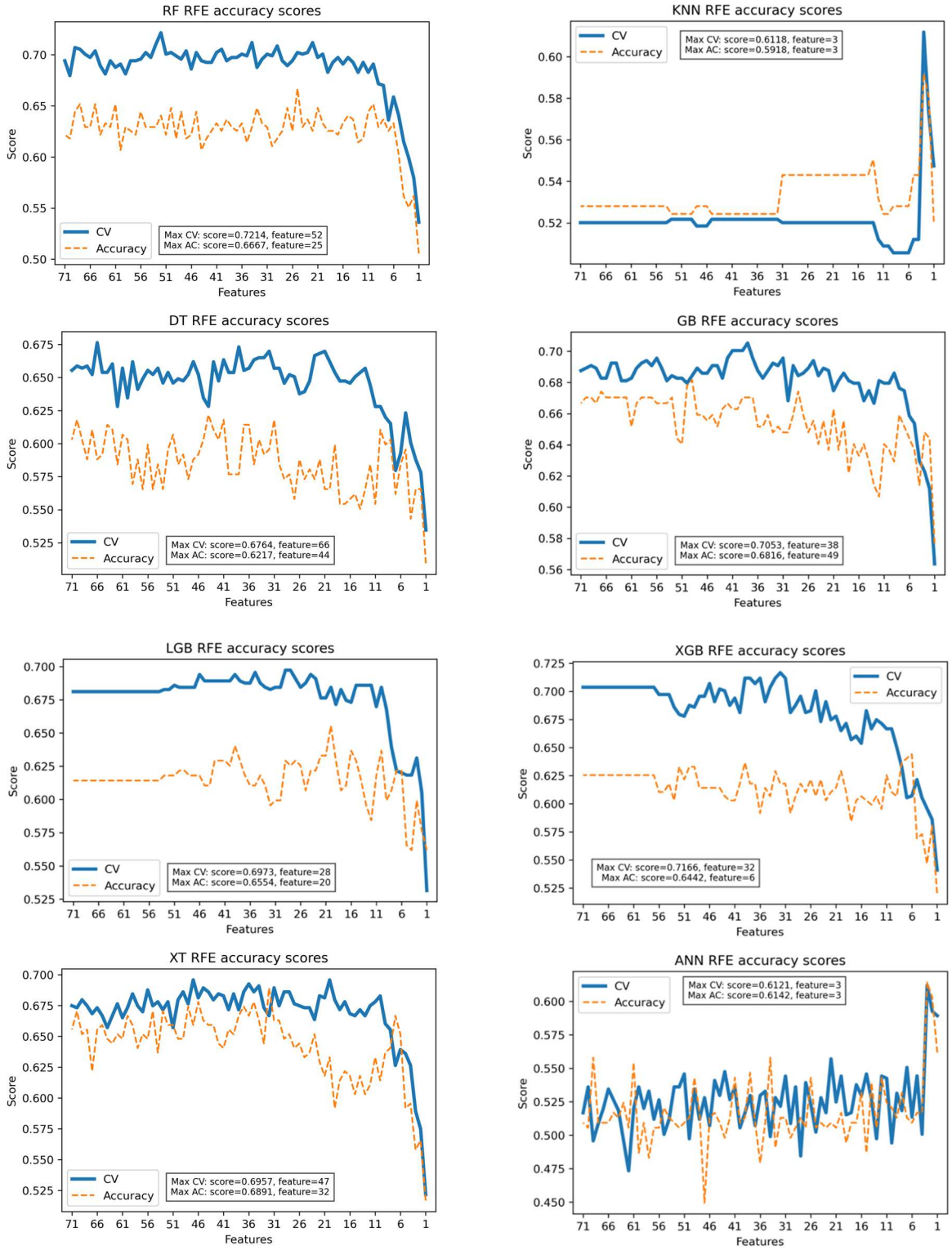

Figure 6: RFE performances of all models

Appendix I: Results of final models (RFE and ROC)

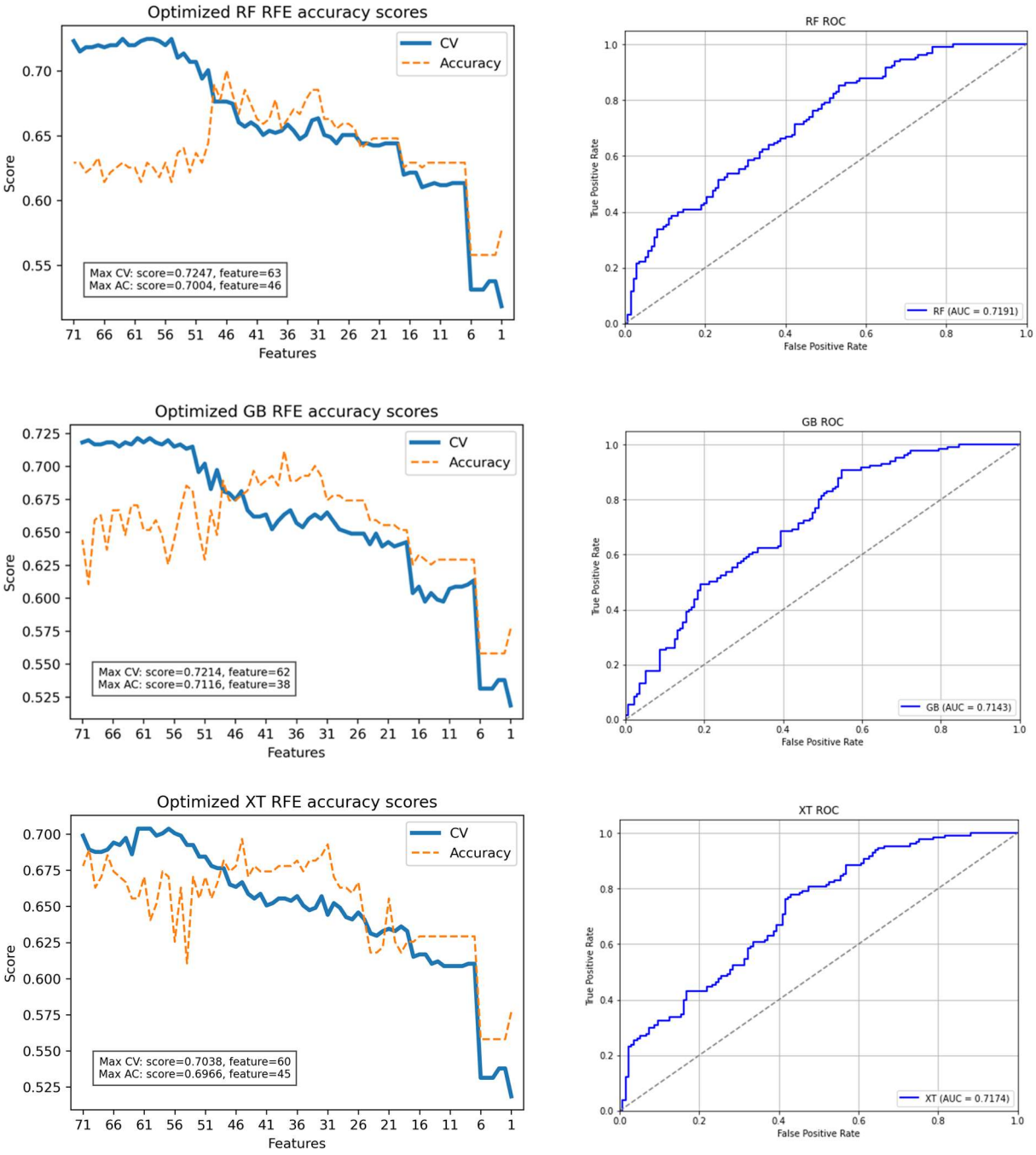

Figure 7: RFE performances and ROC curves of optimised models

Appendix J: Hyperparameter optimization grids

Hyperparameter	**Value**
Max_depth	10
Max_features	None
Min_samples_leaf	1
Min_samples_split	2
N_estimators	200

Table 18: Hyperparameters used for RF final model

Hyperparameter	**Value**
Learning_rate	0.1
Max_depth	3
Max_features	Sqrt
Min_samples_leaf	2
Min_samples_split	2
N_estimators	200
Subsample	1.0

Table 19: Hyperparameters used for GB final model

Hyperparameter	**Value**
Max_depth	20
Max_features	Sqrt
Min_samples_leaf	1
Min_samples_split	10
N_estimators	100

Table 20: Hyperparameters used for XT final model