CHALLENGES IN COLLECTING AND MANAGEING DATA FOR AI APPLICATION IN SMALL AND MEDIUM-SIZED CONSTRUCTION ENTERPRISES

Dominik Steuer[1], Svenja Oprach[1], Hannes Gerber[1] and Shervin Haghsheno[1]
[1] Karlsruher Institute of Technology, Karlsruhe, Germany

Abstract

Demolition projects today face challenges in adopting smart digital tools such as Building Information Modeling (BIM), Big Data, and Artificial Intelligence (AI) within the construction sector. Despite the prevalence of these technologies, their integration into daily construction operations remains limited. This paper presents a case study conducted with a German medium-sized construction company, highlighting the challenges faced in data collection and management practices. The company prided itself on being digital and data-driven, allowing for a comprehensive data collection process across its databases and documents. The collected data was utilized to apply various AI methods in predicting durations for small earthwork and infrastructure projects. However, the analysis of the results revealed that the current data availability and quality were insufficient for effective AI implementation in construction SMEs. Consequently, the paper provides implications to enable SMEs to harness the benefits of AI methods in the future.

Introduction

"AI is important for small and medium-sized businesses (SMEs) because it can increase efficiency and productivity, improve customer experience, provide a competitive advantage, reduce costs, and support data-driven decision making, which can help SMEs to become more efficient, competitive, and profitable. By leveraging AI, SMEs can gain a strategic advantage and stay ahead of the competition in today's fast-paced and constantly changing business environment."

ChatGPT for the input: "Why is AI important for SMEs in two sentences?"

Across the 27-EU countries about 88% of employees work in the narrow construction sub-sector in companies with less than 250 employees in so called micro, small and medium-sized enterprises (SMEs) (European Commission 2023). How can these important companies profit from the rise of digital technology especially artificial intelligence? In this paper a case study in a German medium-sized construction company was conducted to analyze the status quo of data collection and management in the SME for applying AI methods. The case study exposed the need for a structured framework within the dimensions of technology, people, and business models to enable SMEs to profit from AI potentials. Following the CRISP-DM process model (Wirth et al. 2000) data is collected, prepared, and modeled for different business implications. The results are evaluated and discussed. The gap that is formed by the given data is shown and suggestions are made to prepare SMEs for a profiting use of AI applications.

Theoretical Background

The case study follows the scheme of the CRISP-DM (Görz et al. 2020; Matzka 2021) model shown in Figure 1. The choice of machine learning methods is based upon literature research. Here three models are applied on the data set: Artificial Neural Networks, Support Vector Regression, and a Random Forest Regression.

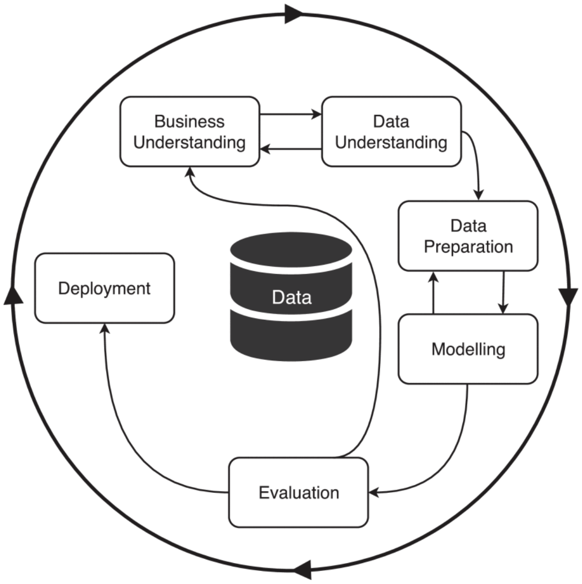

Figure 1: CRISP-DM Model (Martínez-Plumed, F. et al. 2019)

CRISP-DM

Cross Industry Standard Process for Data Mining defines a process model that provides a framework for data mining projects. The model sets a standard to translate business problems into data mining tasks, suggest data transformations and mining techniques, and provides means for evaluating the results (Wirth et al. 2000).

Artificial Neural Networks

The algorithm, that is used by the most in related research is the Artificial Neural Network (ANN) (An et al. 2021; Sharma et al. 2021). One of the main reasons is its applicability onto nearly all functions and patterns. The suitable network-type for the regression problem of this case study is the feedforward neural network (FNN).

Its application is easy but has gained many promising results in related studies (Elmousalami 2021; Fan et al. 2021).

The structure of the network consists of one input layer with input neurons, which have usually the exact same number as the features. So, every datapoint passes the information of each feature to a different input neuron. Then there are one or more hidden layers containing a predefined number of hidden neurons per layer. These are connected to the last layer, which is called output layer. As a learning mechanism for the network the most popular backpropagation algorithm was used (Ertel 2021).

While searching for the best for purpose ML method, the algorithms have been tested with different hyperparameters.

Despite the wide range of applications and the wide usage of the algorithm, especially the hyperparameters that are defined before training are a main weakness. This is because they have a strong influence on the performance of the network while being chosen by a trial-and-error approach or as a reference out of similar studies. Unfortunately, the hyperparameters are not part of any optimization algorithm and need to be selected cautious in advance. Another characteristic is the large amount of datapoints that is needed to train a network properly, which turns out to be a disadvantage in this case and most likely in a lot of cases in construction SMEs, due to the lack of large datasets. Nevertheless, FNNs are a promising choice for the case study because of their wide applicability.

Support Vector Regression

In contrast to FNN, Support Vector Regression (SVR) work with less predefined parameters and merged out of the original classification context of the Support Vector Machines. Both are based on the kernel transformation idea, which implements that linear inseparable datapoints can be separable in a higher feature space after a certain transformation. For this case Study we used the ε-loss-SVR with soft margin (Smola und Schölkopf 2004). One of the largest advantages is that the SVR gets robust results with less datapoints compared to the FNNs (An et al. 2021). Cause of the operating principle only the datapoints representing the support vectors are crucial for the training. The remaining datapoints are not included in the training calculations. Therefore, these models can be trained in a short period of time while providing robust results against outliers. Particularly this resistance can be of great potential in the construction industry with from time-to-time appearing projects that are falling out of the grid. Furthermore, the SVRs are the second most common algorithms in similar studies, where they often produce excellent results even likely to be better than the FNNs (Darko et al. 2020; Peško et al. 2017; Son et al. 2012). Hence the SVRs are an excellent candidate for improving the forecasting accuracy of project durations.

Random Forest Regression

The last algorithm was chosen as a promising although not often used method in related research. Generally, the research on forecasting project duration or cots is severely limited to FNNs and SVRs (Elfaki et al. 2014; Sharma et al. 2021). Whereas other methods are getting close to none research attention. Random Forest Regressions (RFRs) are ensemble learning models, which are consisting of many regression trees as the basic model. They confront the issue of overfitting with combining the forecast of the individual regressors. Precisely a RFR uses the mean of the predictions, consequently it's a bagging algorithm. In addition, the individual regression trees are only learning on a randomized part of the whole dataset, so the risk of overfitting is reduced. RFR has gained good results in its few implementations (Elmousalami 2021) and because of the reduced tendency of overfitting is suitable for an introduction on the subject of project duration forecasting.

Relevant project attributes

The attributes are defined based on a literature analysis of 13 studies conducted in a comparable field. In total 330 attributes are considered and aggregated in categories (Table 1).

Table 1: starting attributes derived from literature

Category	Frequency
size	96
type	46
setting of site	39
customer	35
cost	32
process steps	16
construction management	16
geographical location	15
material	14
weather	7
complexity	5
number of construction sites per project	4
economical data	4
year	1
result	330

Based on the literature research, information regarding project and land size are used alongside project type, cost

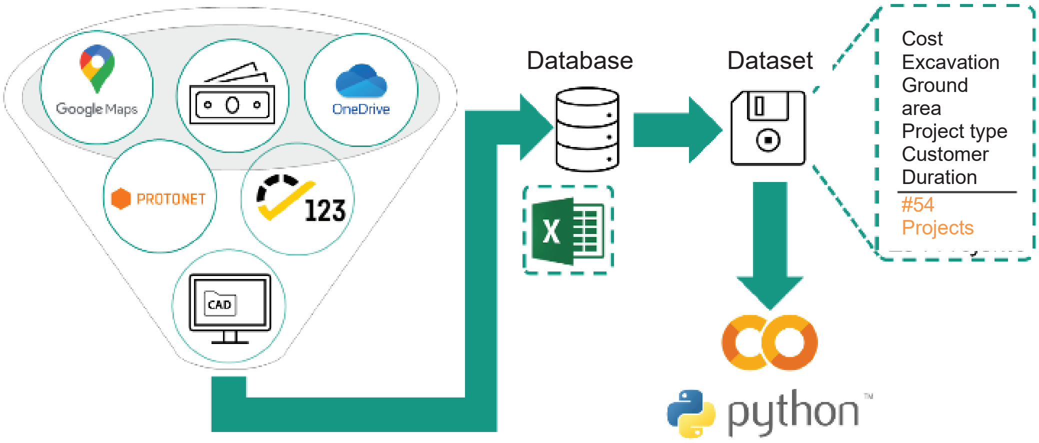

Figure 2: Data preparation and application of ML methods

as well as satisfaction with design and management to determine project outcomes. In contrast weather and economic data are used more seldomly.

Performance metrics

A common way of assessing the quality of a model is via error functions. In the following the performance metrics MSE, RSME, MAE, MAPE and R^2 are described as error functions. (Hyndmann et al. 2006)
In the case of regression, the mean squared error (MSE) described is often used.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (1)$$

However, since this quantity is available in the squared unit of the target variable, the root mean squared error (RMSE) is also frequently used due to its better interpretability.

$$RSME = \sqrt[2]{MSE} \qquad (2)$$

The MAE serves as metric to compare the absolute prediction error. It calculates the difference between the predicted value $\hat{y}$ and output values y. The MAE measures the absolute quality.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (3)$$

The MAPE is the ratio of the difference between the actual output value y and the predicted value $\hat{y}$ to the actual output value y over all data points.

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \qquad (4)$$

The $R^2$is the coefficient of determination and is also frequently used to evaluate a regression model. Contrary to the previous metrics, this is not a metric to determine the deviation, but to evaluate the goodness of fit.

$$R^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \qquad (5)$$

Case Study

The aim of the case study is to predict duration of small earthwork and infrastructure projects more accurately, then solely based on personal or company wide experience from past projects. In collaboration with a German construction company data from all business divisions was collected and structured to be used in AI methods. In the following the data preparation with the data collection and reduction as well as the application of ML methods and its evaluation are presented.

Data collection

The data set originates from a medium-sized company with a predominant area of activity in southern Germany. The construction company is active in both civil engineering and building construction, whereby the work is limited to the data of the subsidiary specialized in civil engineering. The company is making many efforts towards digitalization, such as the digital recording of working hours via a web tool, but it should be noted that the data recording is not error-free or complete and that there is a need for improvement.
The collected dataset has been extracted out of CSV Data, pictures, Google Maps satellites, price indexes from Federal Statistical Office of Germany, CAD-Drawings, final bills as PDFs. The process of data acquisition and combination is described in the following:
Data is aggregated from different platforms used by the company. The data baseline is collected from a time registration software, where the accurate duration of projects and the addresses are derived. To extend the dataset project cost were added, analyzing data from outgoing invoices. Defining unique datapoints that are distinctly connected to a data set let to a reduction of data sets from over 1000 to just above 100. A literature analysis showed the relevance of project size, this was added by surface size and material excavated. The setting of the site is derived from GPS locations. Additionally, information regarding project type is added by manually analyzing drawings, offers, pictures and bill of quantities.

Nevertheless, the weather and economic data is easily accessible and therefore also collected.

Dataset reduction

Studies show that a reduction of attributes can lead to improved results and processing speed (Matzka 2021). Using ANOVA relevant nominal attributes are defined (Sarkar et al. 2018). Using correlation analysis cardinal scaled attributes are defined.

Using ANOVA the relevance of attributes is analyzed and resulted in a non-significance of price indices, season, and setting of site to project duration. This was confirmed applying ML-methods, which let to better results not using the analyzed attributes. Therefore, data was aggregated using this preliminary research.

Overall, the major challenge of data aggregation is preparing a dataset with relevant datapoints from various sources, software tools and data types. This caused a tremendous manual workload, which makes the approach very costly.

Data preparation was done in MS Excel, duplets, missing, and false data was extracted. Due to the small dataset imputation procedures were not applied.

The result of data collection and aggregation lead to a final complete dataset of 54 projects with five starting attributes X ∈ {X_1, ..., X_5} and one resulting attribute y (table 2).

Table 2: metric attributes of the dataset

Feature	Name	Scale	Mean	Min	Max	σ
y	Duration in hours	metric	487	41	3,526	589
X_1	Costs in €	metric	39,825	886	173,654	41,996
X_2	Excavation in m3	metric	503	5	4,000	734
X_3	Ground area in m2	metric	686	8	5,000	1,054

The attributes are characterized by a high variance and a wide interval between minimal and maximal value. The variance needs to be considered when analyzing the standard variance σ. Additional to the metric scale two nominal starting attributes with two to three values are added (table 3).

Table 3: nominal attributes in dataset (translated)

Feature	Name	Scale	Value	Quantity	Distribution
X_4	project type	nominal	no_Housing	32	58%
			new_Housing	23	42%
X_5	customer	nominal	commercial	10	18%
			private	32	58%
			public	13	24%

Table 4 shows an exemplary dataset, that was derived from the various sources.

Table 4: detail from dataset (translated)

Costs in €	Excavation in m3	Ground Area in m2	Projecttpy	Contractor	Duration in hours
42,418	900	680	new_Housing	private	200
29,593	1,040	340	new_Housing	private	314
74,156	410	330	no_Housing	public	1025
6,744	200	400	new_Housing	commercial	64

Application of ML-methods

For this specific case study, the machine learning libraries scikit-learn and Kreas are used both based on the backend framework tensorflow (Sarkar et al. 2018). Both applications are open source. The SVR model is implemented using scikit-learn, for the FNN model tensorflow is used due to its GPU integration and therefore faster processing times.

Data preprocessing is executed with the panda's library. For numerical operations NumPy is applied (Sarkar et al. 2018)

For visual data presentation matplotlib with functions from seaborn is chosen (Sarkar et al. 2018). Statistical parameters are derived from statsmodels and SciPy (Sarkar et al. 2018, Frochte 2021). The implementation is done using python.

The dataset is divided in training and test data with a 70 to 30 ratio (Frochte 2021). For all methods a hyperparameter optimization and a fivefold cross-validation (k = 5) was conducted. The hyperparameters as shown in table 5 are applied.

Table 5: Hyperparameter for machine learning procedure

method	hyperparameter	values	best
SVR	Kernel function	linear; sigmoid; rbf; poly	rbf
	polynomial kernel	1; 2; 3; 4; 5; 6	1
	margin ε	0,0001; 0,001; 0,1; 0,2	0,2
	cost parameter	1; 10; 100; 1000; 10000	100
	kernel parameter ɣ	0,0001; 0,001 ; 0,01; 0,1; 0,2; 0,5; 0,6; 0,9	0,0001
RFR	Trees per model	10; 100; 200; 400; 600; 800; 1000; 1200; 1400;1600;1800; 2000	200
	Max depth	3; 4; 5; 6; 7; 10; 15; None	6
	attributes p‘	1,0; sqrt; log2	sqrt
	min knots	1; 2; 4	1
	min split	2; 5; 10	2
KNN	covered layers	1; 2	1
	covered neurons	8; 16; 32	8
	Activation function	relu; sigmoid; tanh	sigmoid
	batch size	5; 10; 15; 20	15

Evaluation of the results

Table 6 shows the results derived from the model and the dataset. Regarding the mean absolute percentage error (MAPE) of 60% and the root square mean error (RMSE) of 350 hours, the prediction accuracy is not satisfactory. Following Elmousalami (2021) a MAPE of less than 20% should be achieved.

Similar conclusion can be drawn from the mean absolute error (MAE) of 250 hours and R-squared (R^2) of, in the best case, 42%. Therefore only 42% of the variance are described through the defined model.

Overall, the FNN model delivers the best results only in case of MAPE, SVR is slightly better.

Table 6: Results of ML methods

	SVR	RFR	FNN
MAPE	65.96	80.54	68.27
MSE	155,389.60	140,466.64	128,164.67
RMSE	394.19	374.79	358.00
MAE	267.97	261.35	243.43
R^2	0.30	0.37	0.42

Discussion of results

In general, there are a lot of sources for bad performing machine learning applications. From the wrong choice of hyperparameter to not fit for purpose algorithms. Either of these points was dealt with based on an in-depth literature review. In this case study the root cause for bad processing outcome is due to poor dataset quality and the overall number of complete datasets. Therefore, further implications for construction SMES are discussed.

Here we provide implications for SMES to set the baseline for the application of ML methods in the future in the fields of technology, people and business.

Technology – Common Data Environment (CDE)

There are studies that work with small datasets as provided in this paper, but the amount is too small especially when working with ANNs (Elmousalami 2021) where thousands of datasets are recommended. In this case study the availability of datapoints was not the main issue for poor quality in datasets. The main issues are naming and unique identifiers for project data to connect then, usage of different platforms to store and collect data. Data silos between different data types, (i.e., ERP data, CAD data and machine control data). For future improvement it is recommended to develop a data strategy with a common data environment to enable easy access to datasets, without vast amounts of preprocessing. Additionally, there could be autonomous methods to collect process data.

People – Data Collection

Within the construction environment a lot of data is collected manually or even on paper and drawings. This implies human error in the data collection process. This error could be handled in different ways but must be addressed to improve the quality on analysis (Burchard 2011). Data collection can be standardized, processes can be clearly defined and trained via employee education. Strict rules need to be defined to structure the collected data. Another improvement could be achieved by restricting the possible data entries to standardized lists or supporting the employee with additional sensors from mobile devices or surveying equipment. Continuous and consistent data collection process need to be established, to set the baseline for AI applications.

Business Model

Construction SMES need to develop a data-driven strategy to profit from AI applications in terms of business success. Therefore, a cascade of four steps needs to be developed (Hofstadler and Motzko 2021):

1) A digital vision in alignment to the overall company goals is established and defined with key performance indicators (KPIs).
2) After the vision, strategic fields of action are defined with analysis of the industry and the market position. With a focus on fields that are disrupted by digitization.
3) Strategic action fields are broken down to functional action fields. Here specific areas in the business are defined, where digitization can be implemented.
4) Development of an execution plan, operational tasks, and processes, where machine learning methods could be used, are defined.

Here the case study provides a good example of company, that describes its approach as digital-first, but misses the overall strategy to align different solutions to create a reliable dataset. Cornerstones for data management and storage could be defined as follows (Hofstadler und Motzko 2021):

- **Completeness**, data is collected all along the relevant process steps and could even be further developed for the operation of buildings afterwards to create a closed loop in the building's life cycle.
- **Correctness**, data is collected without failure, because wrong dataset led to false predictions or a vast amount of preprocessing effort.
- **Up-to-dateness**, data is as real-time as possible to generate the most valuable outputs.
- **Structure and Traceability**, data is structure in a clear understandable and biunique way.
- **Immutability**, data like identifiers cannot be changed.
- **Confidentiality**, data is protected throughout the system to protect companies' knowledge and employee's personal data.
- **Integrity**, data environments need to store the data consistent and trustworthy
- **Availability**, data is accessibly throughout construction processes for all the relevant people and processes.
- **Access and liability**, data is reliable and accessible to the right persons and processes.

Conclusion

Due growing data collecting practices, there is a high demand of methods for data analytics and especially for methods in the field of Artificial Intelligence. In this paper we elaborate the challenges and implications for SMEs collecting and managing their data. This focus is first set since 88 % of the employees in the 27-EU countries work in SMEs.

In an exemplary case study, data from a SME is pre-processed, linked, and lastly analyzed with the target to predict the costs and the duration of construction projects. Here, the results of a FNN, SVR and RFR were compared. Even the best model only shows a MAPE of 65 %. The results show that the use of AI methods is currently not possible even with a lot of manual data preparation and cleaning.

Nevertheless, the target of this paper is rather in defining data collecting and managing recommendations than in selecting the best model (data versus model-centric focus). SME construction companies must establish:

1. a CDE system (Technology)
2. standardized processes for a continuous and consistent data collection (People)
3. a data-driven strategy (Business model)

By including the employees in the digitalization strategy and in improving the daily data management practices, the use of AI for SME will become more possible in the future.

Due to the examined dataset and the manually set hyperparameters there are limitations in this research. In the future research the prediction accuracy:

- of further ML models (e.g., XGBoost, Ensemble methods),
- of common project management practices with ML models
- by setting different hyperparameters,
- and by documenting more features in the relevant data categories

should be analyzed and compared.

Acknowledgments

We express our gratitude to Steuer Tiefbau GmbH for facilitating and supporting this research. The authors acknowledge the financial support by the Federal Ministry for Economic Affairs and Climate Action of Germany in the project Smart Design and Construction (project number 01MK20016F).

References

An, Y., Li, H., Su, T. und Wang, Y. (2021) Determining Uncertainties in AI Applications in AEC Sector and their Corresponding Mitigation Strategies. Automation in Construction, 131, p.pp.103883. DOI: https://doi.org/10.1016/j.autcon.2021.103883.

Barchard, K. A., & Pace, L. A. (2011). Preventing human error: The impact of data entry methods on data accuracy and statistical results. Computers in Human Behavior, 27(5), 1834-1839.

Darko, A., Chan, A. P., Adabre, M. A., Edwards, D. J., Hosseini, M. R. und Ameyaw, E. E. (2020) Artificial intelligence in the AEC industry: Scientometric analysis and visualization of research activities. Automation in Construction, 112, p.pp.103081. DOI: https://doi.org/10.1016/j.autcon.2020.103081.

Elfaki, A. O., Alatawi, S. und Abushandi, E. (2014) "Using Intelligent Techniques in Construction Project Cost Estimation: 10-Year Survey. Advances in Civil Engineering, 2014, p.pp.1–11. DOI: https://doi.org/10.1155/2014/107926.

Elmousalami, H. H. (2021) Comparison of Artificial Intelligence Techniques for Project Conceptual Cost Prediction: A Case Study and Comparative Analysis. IEEE Transactions on Engineering Management, 68(1), p.pp.183–196. DOI: https://doi.org/10.1109/TEM.2020.2972078.

Ertel, W. (2021) Grundkurs Künstliche Intelligenz: Eine praxisorientierte Einführung [Basic Course in Artificial Intelligence: A Practice-Oriented Introduction], Wiesbaden: Springer Fachmedien Wiesbaden; Springer Vieweg, p.pp.313.

European Commission (2023, 10. January) Internal Market, Industry, Entrepreneurship and SMEs. https://single-market-economy.ec.europa.eu/sectors/construction/observatory/data-mapper_en

Fan, S.-L., Yeh, I.-C. und Chi, W.-S. (2021) Improvement in Estimating Durations for Building Projects Using Artificial Neural Network and Sensitivity Analysis. Journal of Construction Engineering and Management, 147(7), p.pp.04021050. DOI: https://doi.org/10.1061/(ASCE)CO.1943-7862.0002036.

Frochte, J. (2021) Maschinelles Lernen: Grundlagen und Algorithmen in Python [Machine Learning: Basics and Algorithms in Python], München: Carl Hanser Verlag GmbH & Co. KG, p.pp.51,85.

Görz, G., Schmid, U., und Braun, T. (2020) Handbuch der Künstlichen Intelligenz [Handbook of Artificial Intelligence], Berlin, Boston: De Gruyter Oldenbourg, p.pp432.

Hofstadler, C. und Motzko, C., Hrsg. (2021) Agile Digitalisierung im Baubetrieb: Grundlagen, Innovationen, Disruptionen und Best Practices [Agile digitalization in construction operations: fundamentals, innovations, disruptions and best practices], Wiesbaden: Springer Vieweg, p.pp. 53, 96.

Hyndman, R.J. & Koehler, A.B. (2006) Another look at measures of forecast accuracy. International Journal of Forecasting, 22.4, p.pp.679–688.

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., ... & Flach, P. (2019) CRISP-DM twenty years later: From data mining processes to data science trajectories. IEEE Transactions on Knowledge and Data Engineering, 33(8), p.pp.3048-3061.

Matzka, S. (2021) Künstliche Intelligenz in den Ingenieurwissenschaften: Maschinelles Lernen verstehen und bewerten [Artificial intelligence in engineering: understanding and evaluating machine learning], Wiesbaden: Springer Vieweg, p.pp.18,162.

Peško, I., Mučenski, V., Šešlija, M., Radović, N., Vujkov, A., Bibić, D. und Krklješ, M. (2017) Estimation of Costs and Durations of Construction of Urban Roads Using ANN and SVM. Complexity, 2017, p.pp.1–13. DOI: https://doi.org/10.1155/2017/2450370.

Sarkar, D., Bali, R., und Sharma, T. (2018) Practical machine learning with Python: A Problem-Solver's Guide to Building Real-World Intelligent Systems, Berkeley, CA: Apress, p.pp.75-85, 109, 116, 181.

Sharma, S., Ahmed, S., Naseem, M., Alnumay, W. S., Singh, S. und Cho, G. H. (2021) A Survey on Applications of Artificial Intelligence for Pre-Parametric Project Cost and Soil Shear-Strength Estimation in Construction and Geotechnical Engineering. Sensors, 21(2), p.pp.463. DOI: https://doi.org/10.3390/s21020463.

Smola, A. J. und Schölkopf, B. (2004) A tutorial on support vector regression. Statistics and Computing, 14, p.pp.199–222. DOI: https://doi.org/10.1023/B:STCO.0000035301.49549.88.

Son, H., Kim, C. und Kim, C. (2012) Hybrid principal component analysis and support vector machine model for predicting the cost performance of commercial building projects using pre-project planning variables. Automation in Construction, 27, p.pp.60–66. DOI: https://doi.org/10.1016/j.autcon.2012.05.013.

Tofallis, C. (2015) A better measure of relative prediction accuracy for model selection and model estimation. Journal of the Operational Research Society, 66, p.pp.1352–1362. DOI: https://doi.org/10.1057/jors.2014.103.

Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (Vol. 1, pp. 29-39).