University of Twente / ET / CEM / CME

DATA DRIVEN PREDICTION AND REDUCTION OF EXCAVATION DAMAGES

ENGD PROJECT

JIARONG LI J.LI-5@UTWENTE.NL

DR.IR. L.L. OLDE SCHOLTENHUIS L.L.OLDESCHOLTENHUIS@UTWENTE.NL
DR.IR. E.J.A. FOLMER ERWIN.FOLMER@KADASTER.NL
PROF.DR.IR. A.G. DOREE A.G.DOREE@UTWENTE.NL

kadaster

kpn

Rijksinspectie Digitale Infrastructuur
Ministerie van Economische Zaken en Klimaat

UNIVERSITY OF TWENTE.

PROBLEM

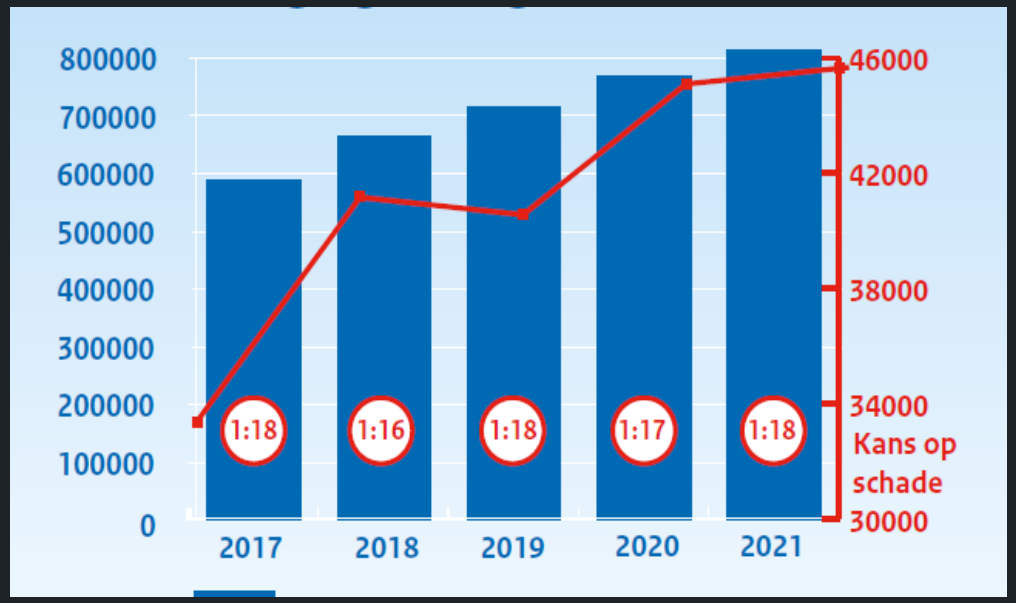


Figure: Number of excavations and excavation damages in the Netherlands from 2017 to 2021

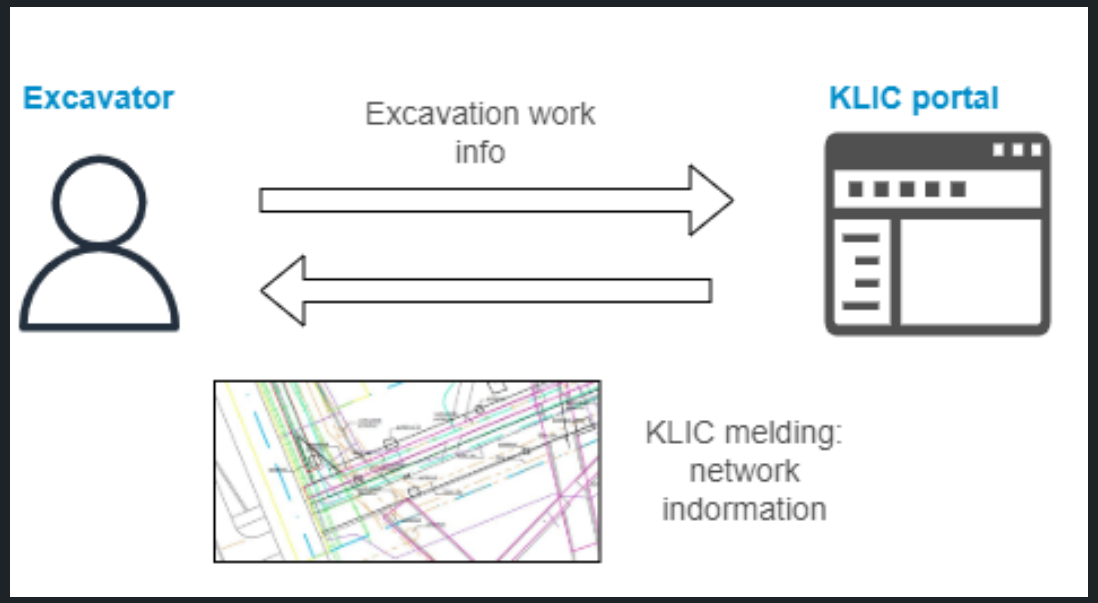


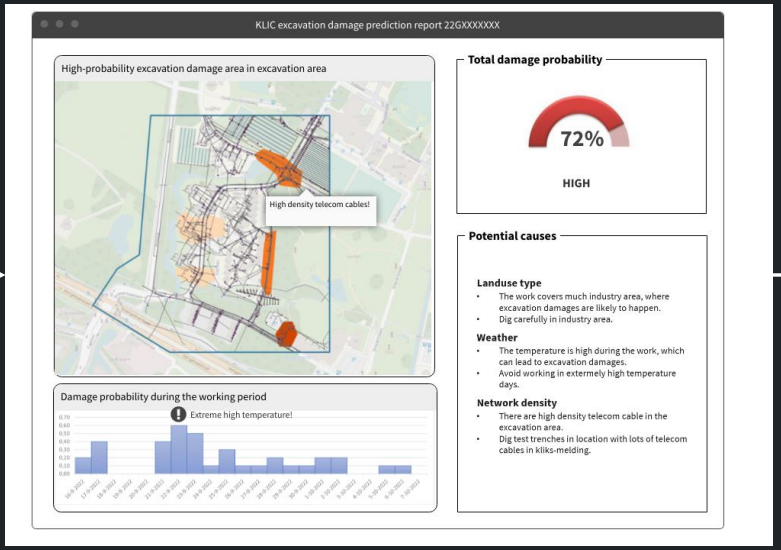
Figure: Current KLIC process

https://www.rdi.nl/documenten/publicaties/2022/05/31/feiten-en-cijfers-over-schade-door-graafwerkzaamheden-2021

Objective: Develop a machine learning model to predict excavation damages and apply this model to reduce damage occurrences.

KLIC request

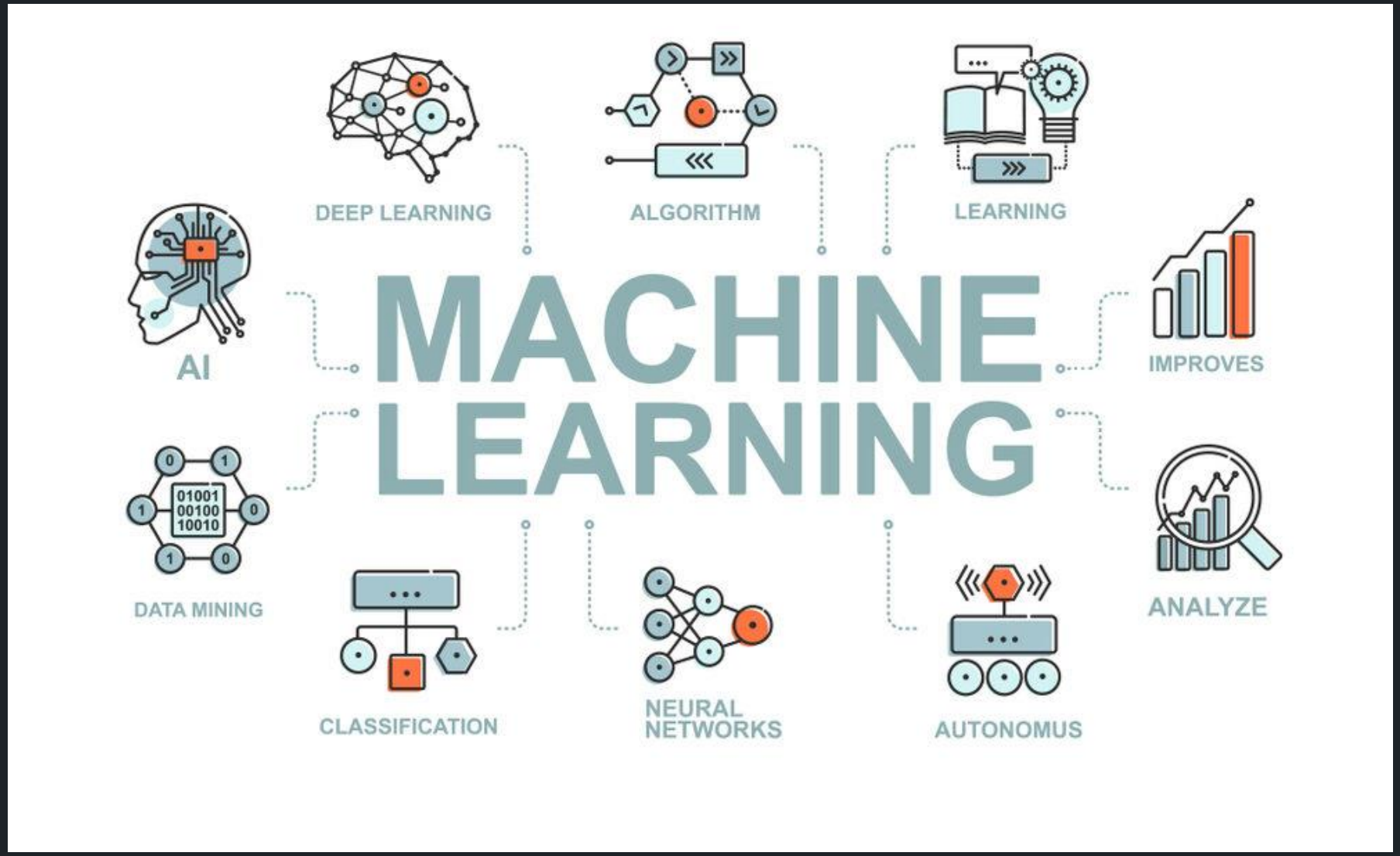
ML model



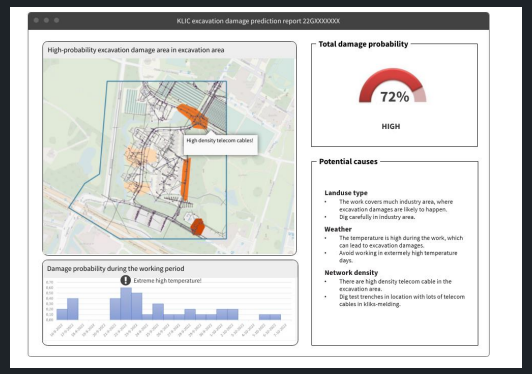
Utility owners

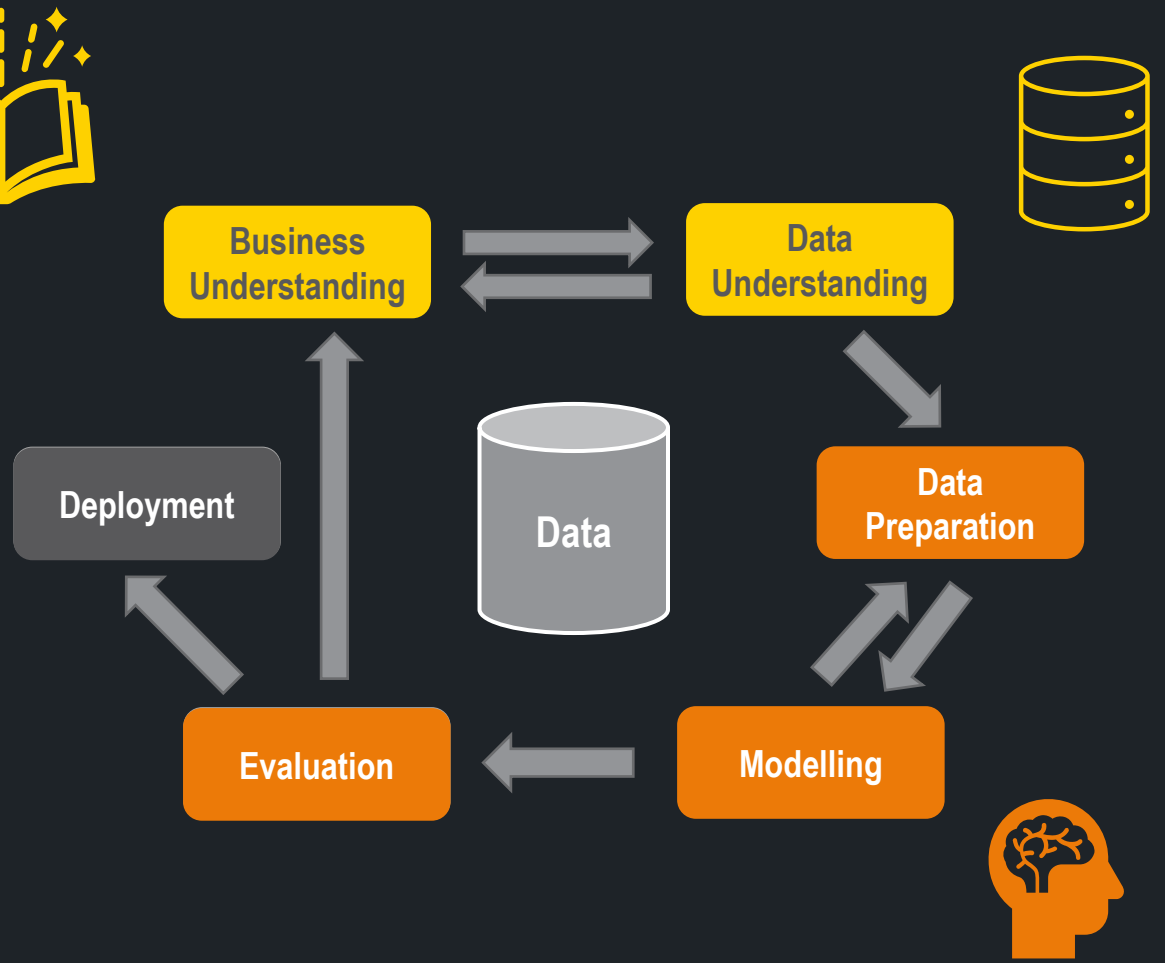
Excavation and client company

Rijksinspectie Digitale Infrastructuur

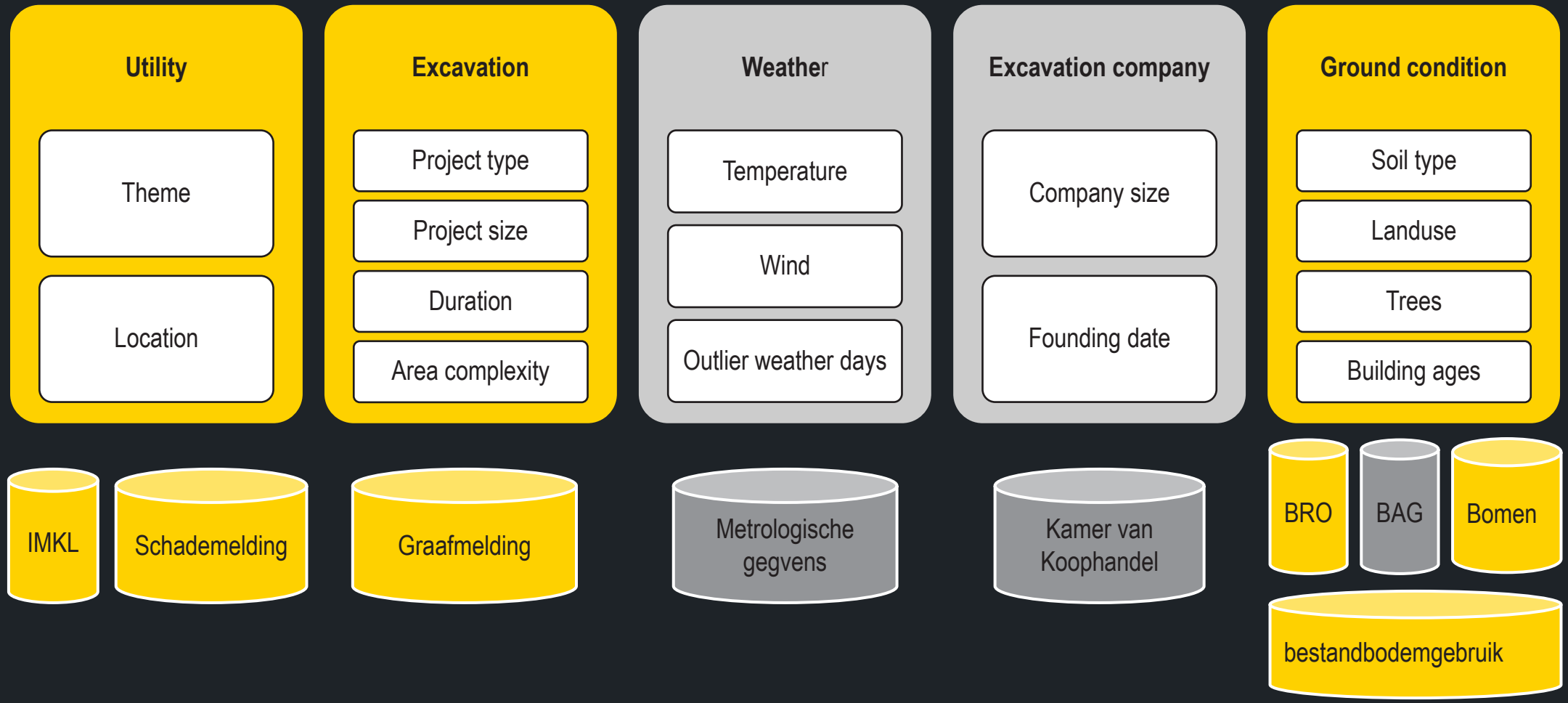

MACHINE LEARNING
DEEP LEARNING
ALGORITHM
LEARNING
AI
IMPROVES
01001
00100
10010
DATA MINING
ANALYZE
CLASSIFICATION
NEURAL NETWORKS
AUTONOMUS

METHODOLOGY: DM-CYCLE





DATA COLLECTION



PRE-PROCESSING

Table 1: Description of datasets for modelling

Table 2: Overview of the modelling data

Attributes	Value
Number of observations	769398
Number of independent variables	125
Number of dependent variables	1
Number of positive cases	18151
Number of negative cases	751247
License	closed data
Interested columns	-

Table 3: Structure of the modelling data

Column name	Description	Data type
EV	indication of containing risk area defined by utility owners	int
Oppervlakte-Woonplaats	area of the municipality	float
Opdrachtgever	client company	int
Opervlakte-Graafbericht	area of the digging polygon	float
Prioriteit	priority of the work	int
Grondroerder	excavation company	int
Berichtsoort	type of the digging request	int
Thema	dummy variables of types of underground cable/pipeline declared by utility owners	int
Typen-werkzaamheden	dummy variables of work type involved	int
m-work	month of the work	int
d-work	day of the work	int
m-request	month of the requesting	int
h-request	hour of the requesting	int
diff-wr	difference of hours between working day and requesting day	int
OPPERVLAKTE	polygon area	float
AANTAL-COORDINATEN	number of coordinates of the polygon	int
LENGTH	perimeter of the polygon	float
Schade	indication of at least one damage occurs during the work	int
Datum-aanvang-str, Datum-aanvraag-str		

MODEL SELECTION

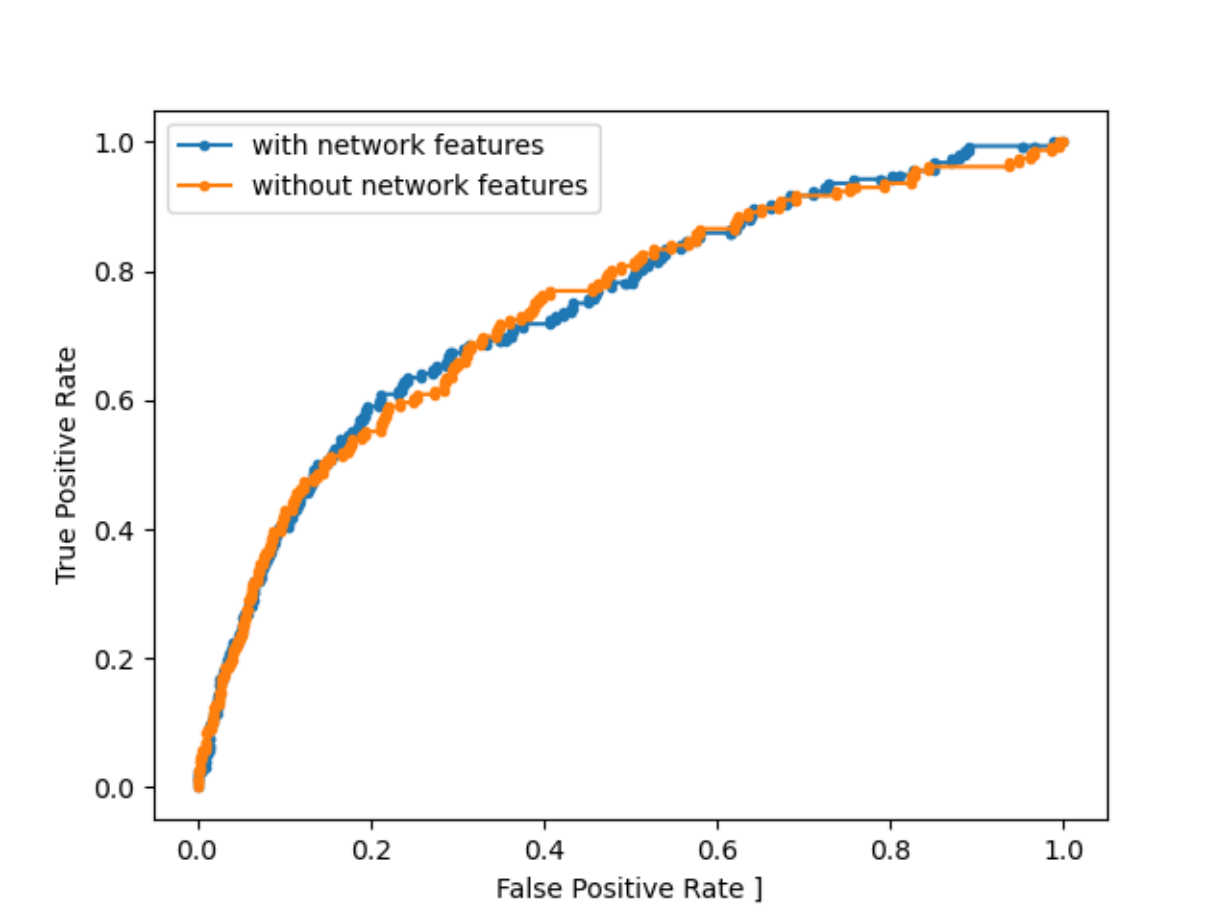
Table 4: Result of Pycaret testing sorted by AUC

Model	AUC	Training Time (s)
Light Gradient Boosting Machine	0.829	4.723
Gradient Boosting Classifier	0.814	116.745
Ada Boost Classifier	0.808	32.700
Random Forest Classifier	0.792	63.348
Extra Trees Classifier	0.786	97.854
Linear Discriminant Analysis	0.782	21.640
Naive Bayes	0.683	3.362
K Neighbors Classifier	0.584	684.615
Decision Tree Classifier	0.554	14.297
Logistic Regression	0.514	78.739
Dummy Classifier	0.500	0.928
Quadratic Discriminant Analysis	0.460	16.962
Ridge Classifier	0.000	4.159
SVM - Linear Kernel	0.000	13.060

Table 5: Evaluation of prediction models

Model	XGBoost	LightGBM	CatBoost
Features used	63	63	125
Training time (s)	28	10	480
Predicting time	2	2	4
AUC score	0.829	0.827	0.833
PR score	0.186	0.185	0.193
Balanced accuracy	0.749	0.743	0.749
Precision	0.090	0.100	0.110
Recall	0.660	0.610	0.620
F1-score	0.160	0.180	0.180

FEATURE SELECTION: NETWORK FEATURES



https://dataexcellence.nl/2020/02/20/enexis/
https://www.techconnect.nl/nieuws/ziggo/8303/nieuwe-zakelijk-pakketten-van-ziggo-met-backup/
https://www.overstappen.nl/internet-bellen-tv/internetaanbieders/kpn/

RESULT: EVALUATION

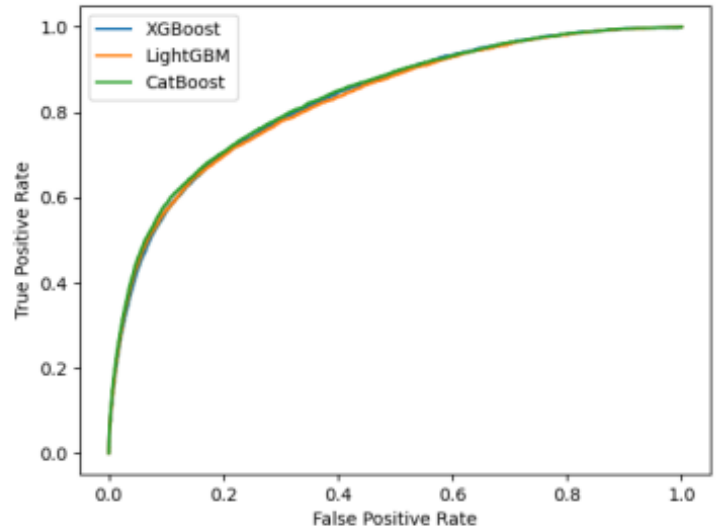


Figure 2: ROC curve of the XGBoost, LightGBM and CatBoost classifiers

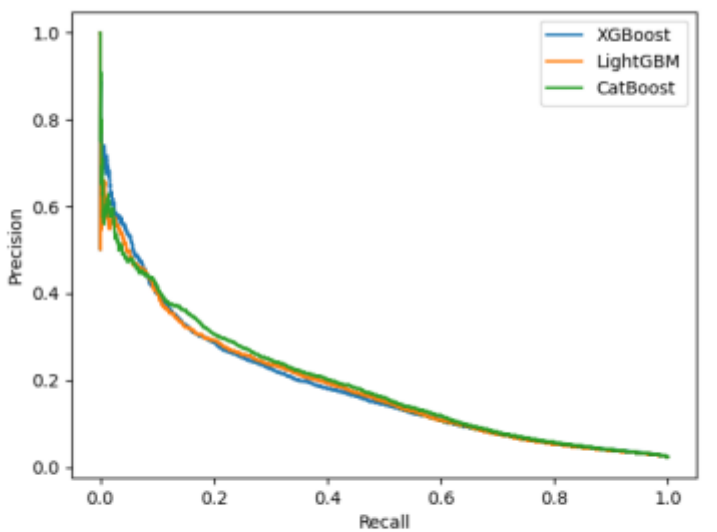


Figure 3: Precision-recall curve of the XGBoost, LightGBM and CatBoost classifiers

Improved XGBoost model: AUC: 0.827, Balanced accuracy: 0.747

Important features: Tree density, Interval between requesting and working, Polygon complexity, Excavation company and Client company, Landuse (build-up area and highway)

RESULT: FEATURE IMPORTANCE

1	tree_mean	Average tree density of the excavation area
2	Diff_wr	Difference of days between requesting day and working day
3	Opdrachtgever	Name of the client company
4	LENGTH	Perimeter of the excavation area
5	Behouwd.exclusief.bedrijfsterrein	Buildup area excluding business area (a landuse type)
6	Grondroerder_count	Number of excavation work taken by the excavation company in 2021
7	Polygon_complex	Perimeter / area
8	Opdrachtgever_count	Number of excavation work taken by the client company in 2021
9	Grondroerder	Name of excavation company
10	Hoofdweg	Highway (a landuse type)

Table: Top 10 most important features by the XGBoost model

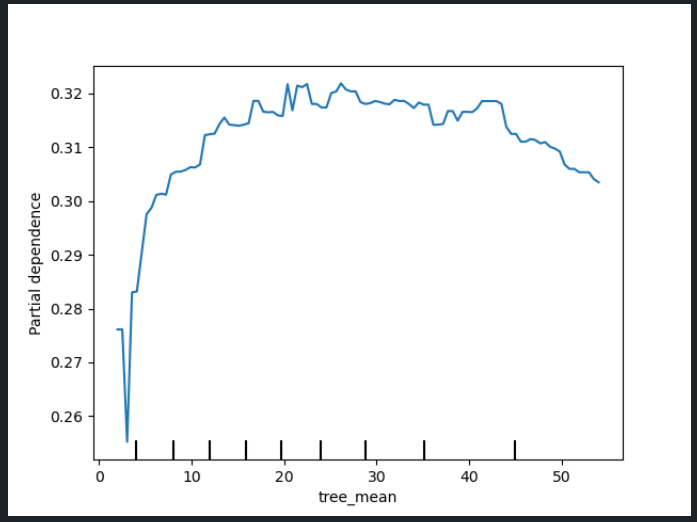
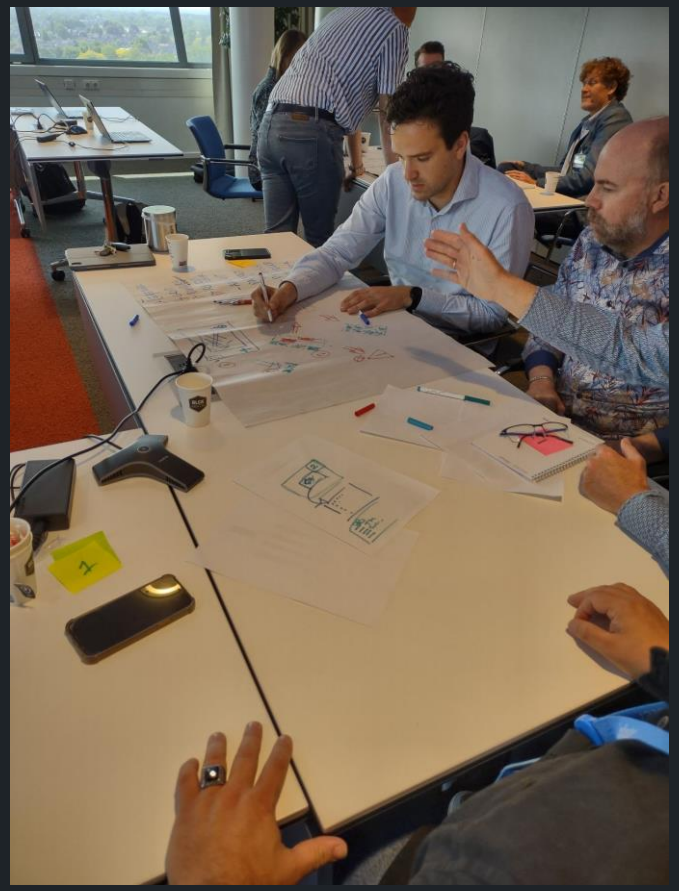


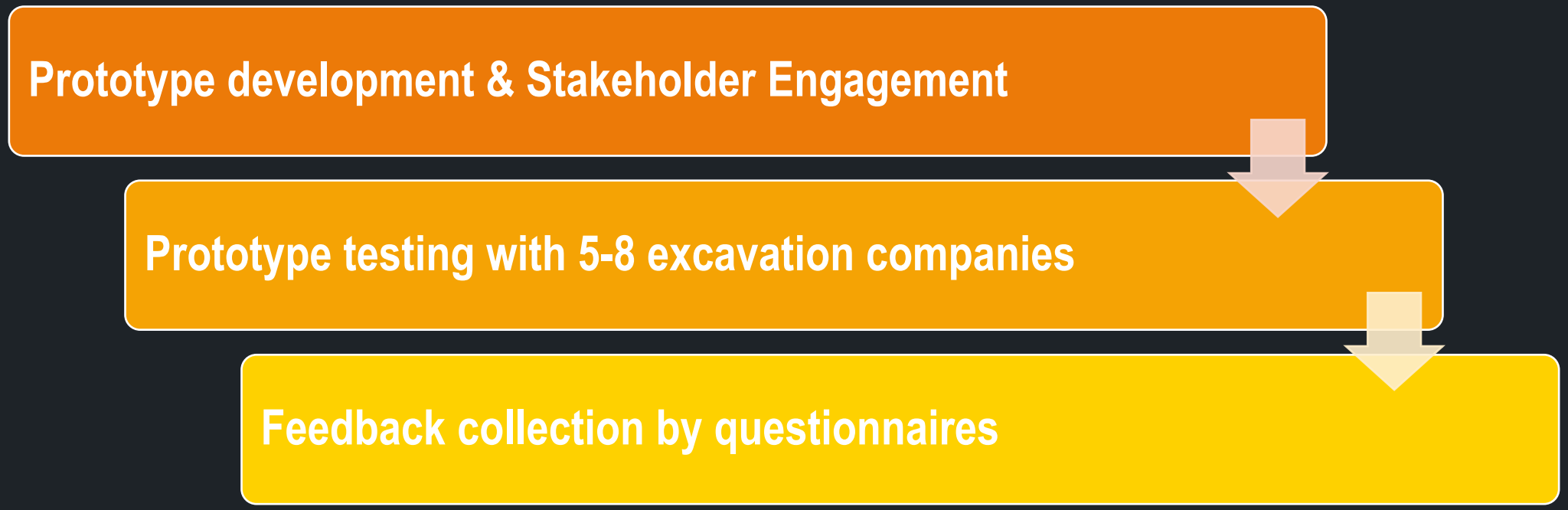
Figure: Partial dependence plot of the tree_mean (tree density) feature

SYSTEMS DESIGN: WORKSHOP



https://www.linkedin.com/feed/update/urn:li:activity:7067815063922864128/

SYSTEMS VALIDATION



THANK YOU!

JIARONG LI
J.LI-5@UTWENTE.NL

Rijksinspectie Digitale Infrastructuur
Ministerie van Economische Zaken en Klimaat

kpn

kadaster

UNIVERSITY OF TWENTE.