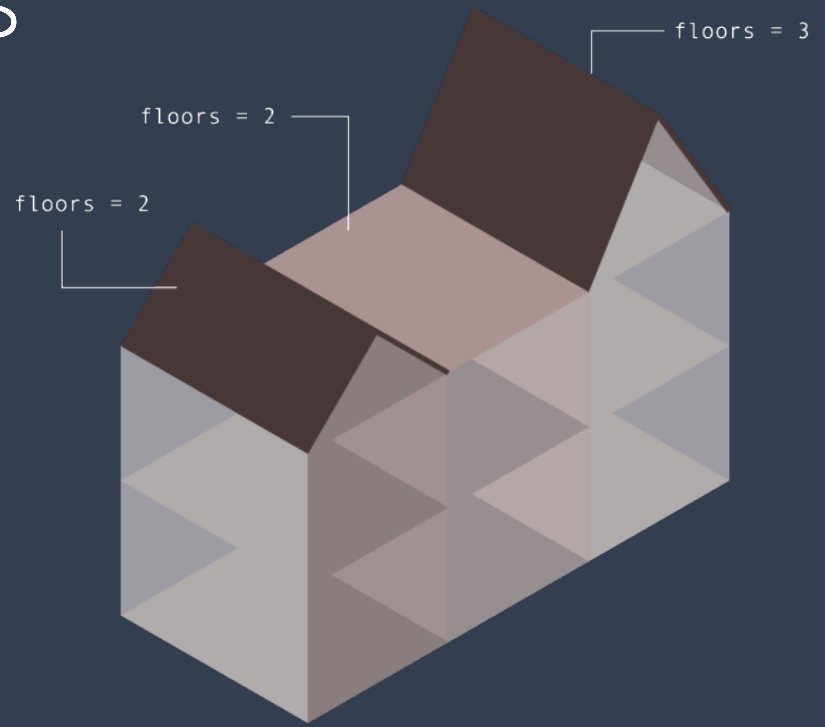
MSc Thesis in Geomatics

# Inferring the number of floors of building footprints in the Netherlands




Ellie Roy

13th January 2022

Delft University of Technology
Faculty of Architecture
and the Built Environment

# Topic relevance

## Energy demand estimation

Image: https://nos.nl/artikel/2352272-europese-commissie-komt-met-renovatiegolf-voor-huizen-vanwege-klimaat.html

## Building population estimation

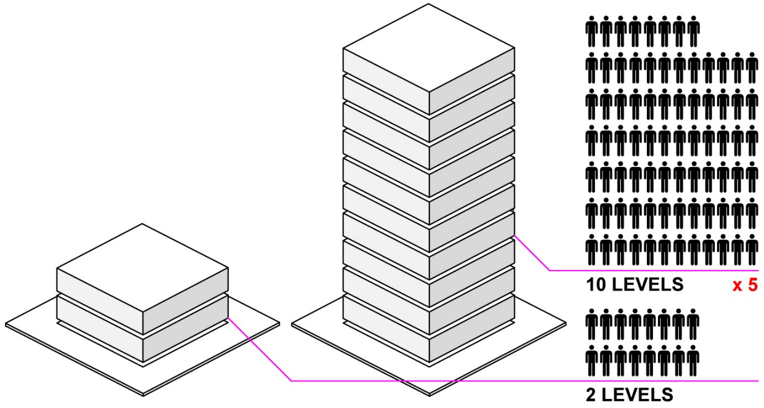
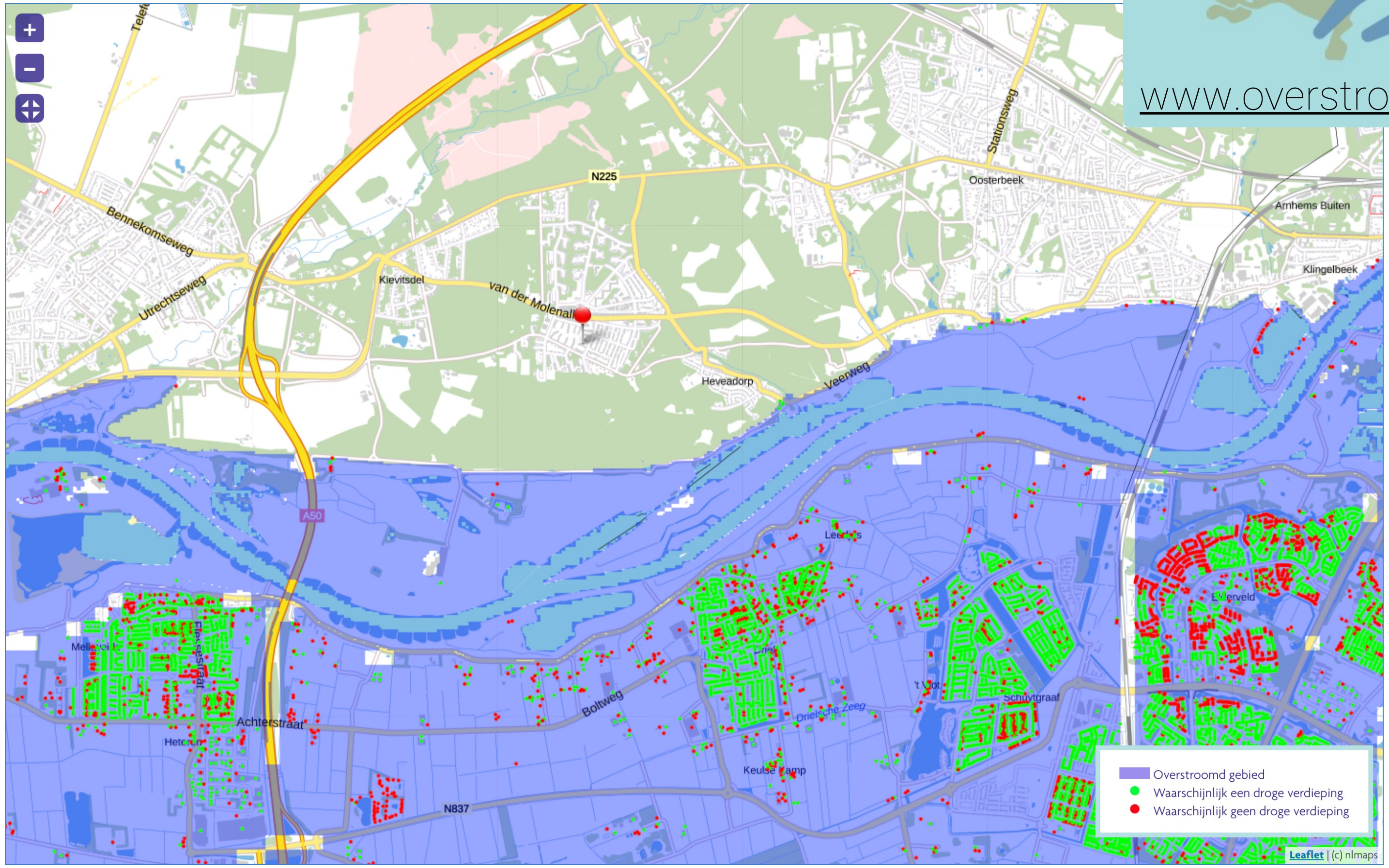



Image: https://densityarchitecture.wordpress.com/2013/01/18/understanding-density/

## Flood response plans

Image: https://www.eoi.es/blogs/imsd/let-the-flood-come-dutch-urban-planning/

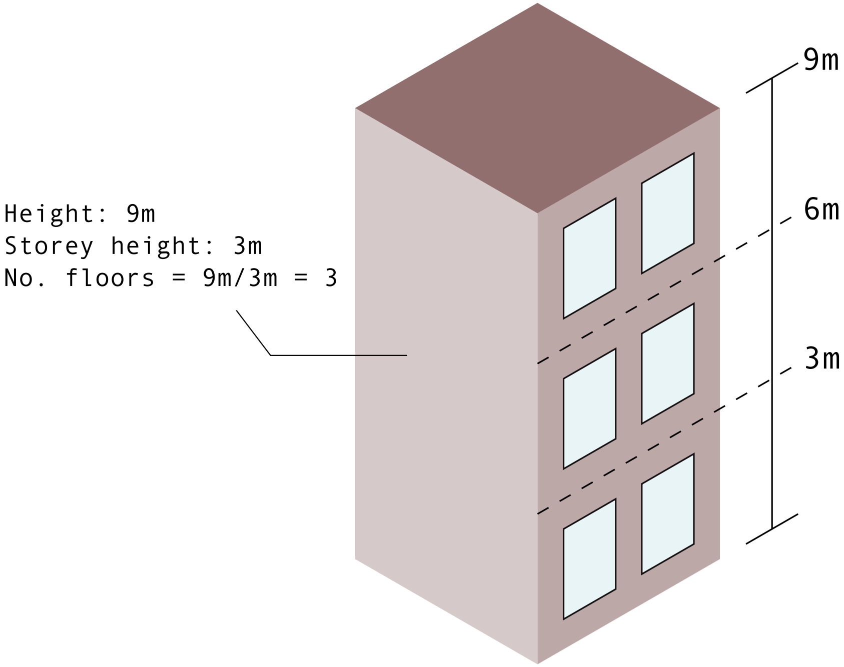
# Topic relevance

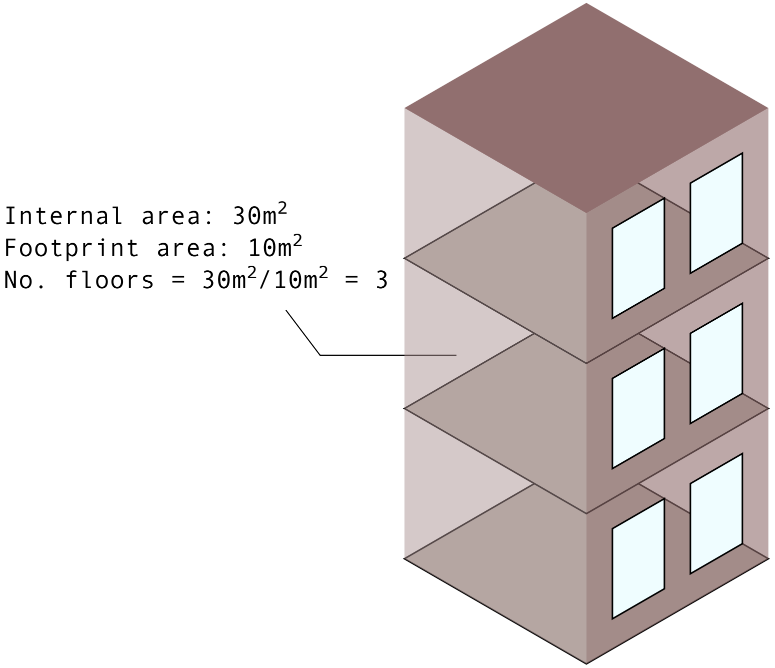

TU Delft

# Current approach

## Height-based




## Area-based



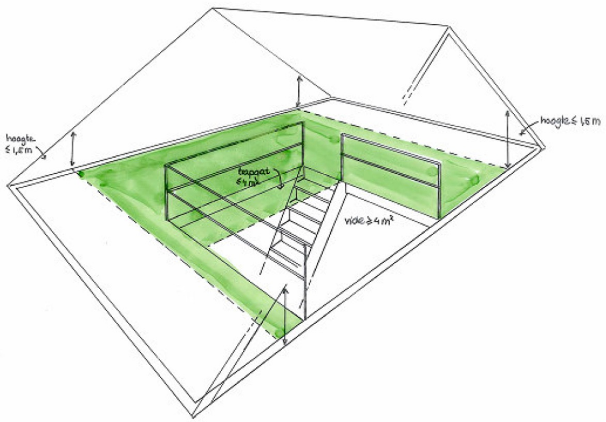
# Limitations of current approach

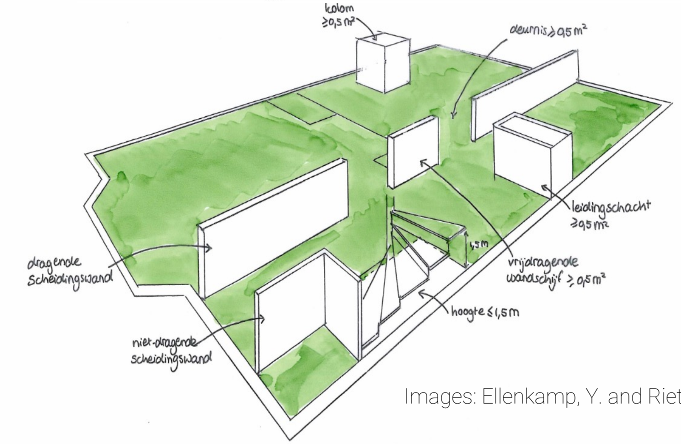
## Height-based

Actual: 1 floor
Calculated: 2 floors

Actual: 8 floors
Calculated: 9 floors

## Area-based




Images: Ellenkamp, Y. and Rietdijk, M. (2010)

# Research objectives

# Research questions

To what extent can **machine learning** provide a **better estimate** of the number of floors than a **purely geometric** approach?

a) Which **features** are related to the number of floors? Is there any **overlap** between these features and which subset yields the **best** results?

b) Which **machine learning algorithm** provides the **best** results? How are the results affected by **feature subsets** that reflect different levels of **data availability**?

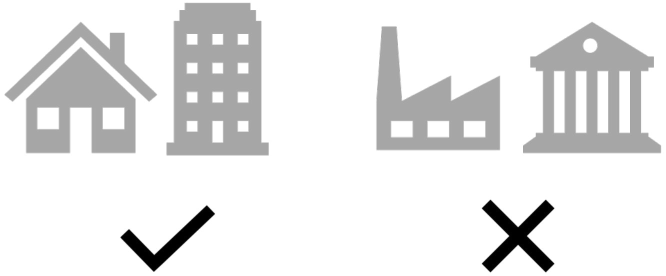
c) What **level of performance** can be achieved compared to a **purely geometric approach**? What types of **gross errors** are present?

d) Since floor count is generally an integer value, is this a **regression or classification** problem? If considered regression, how does **rounding** the predictions affect the results?

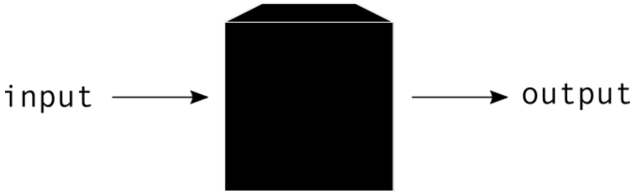
# Scope

(1) NL

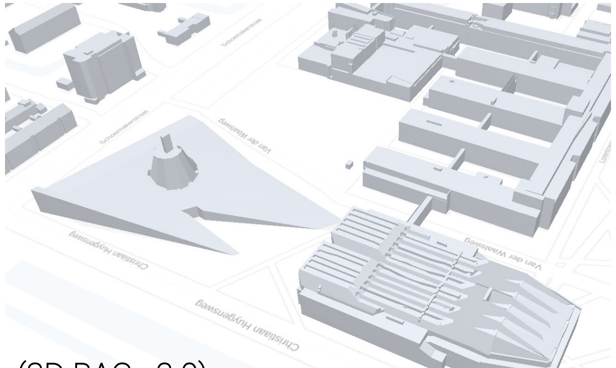
(2) (Mixed-)residential buildings



(3) Algorithm ≠ black-box




(4) No geometric modelling



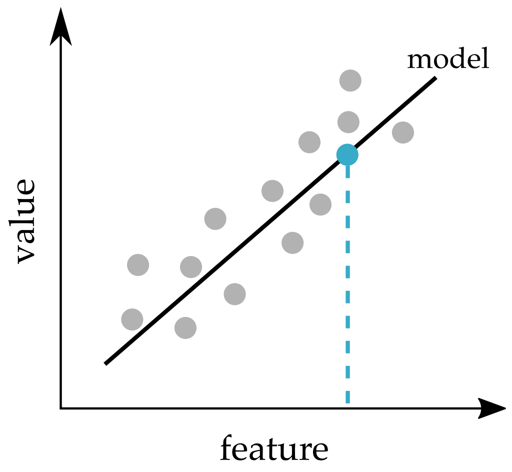
(3D BAG v2.0)

# Background

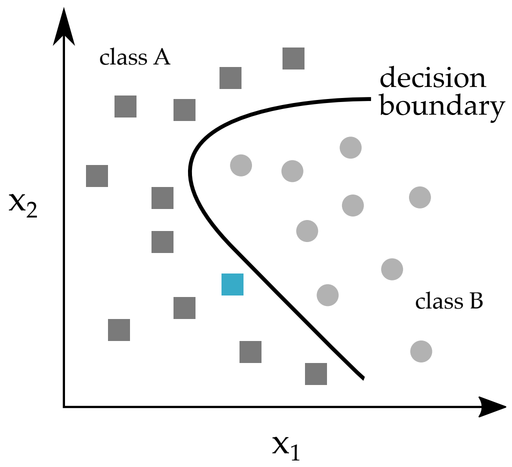
# Machine learning

## Supervised learning

### Regression



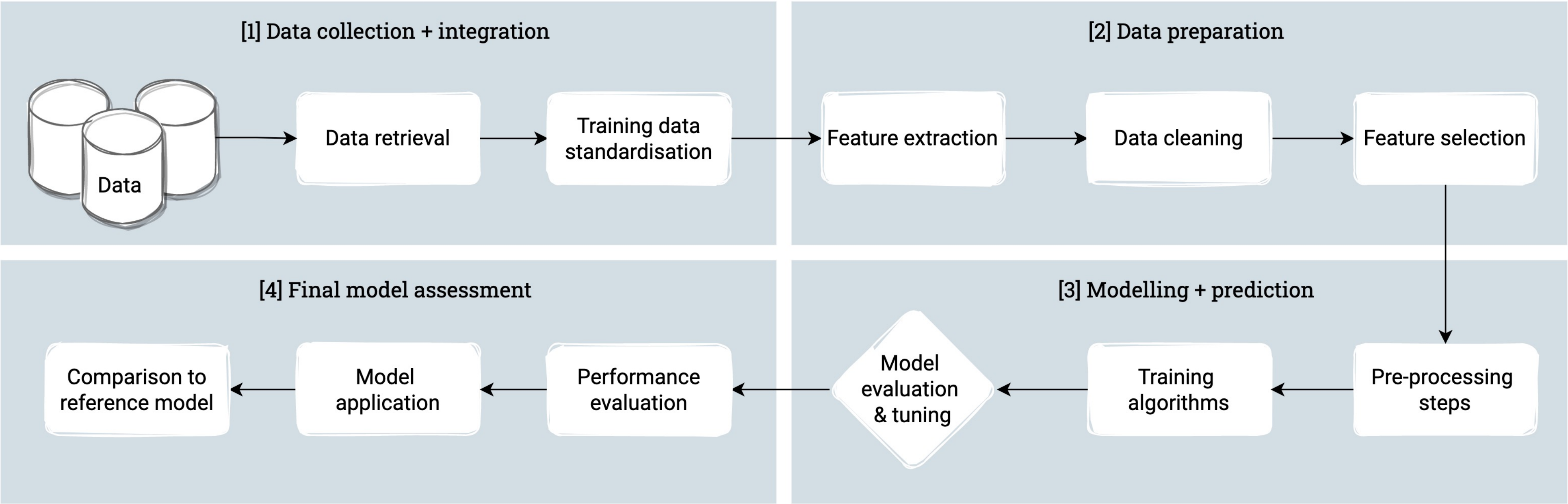

### Classification



# Methodology

- Research objectives
- Background
- Methodology
- Implementation
- Results
- Conclusions

# Overview methodology



# Extracted features

| Cadastral | Census |
|---|---|
| - Construction year | - Population density |
| - Building function | - % multi-household buildings |
| - Net internal area | - Average no. cafes in 1km |
| - No. units | |

| 2D geometric | 3D geometric |
|---|---|
| - Area | - Volume |
| - Perimeter | - Roof surface area |
| - No. vertices | - Wall surface area |
| - No. neighbours in 100m | - Building height |
| - No. adjacent buildings | - Ridge – eave height |
| | - Roof shape |

Image: Biljecki et al. (2014)

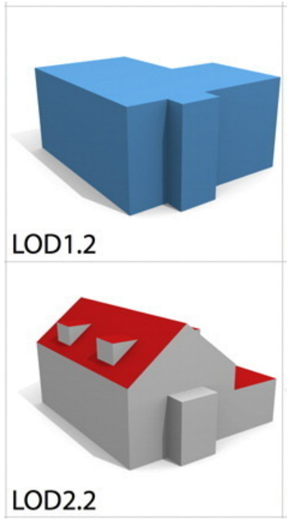
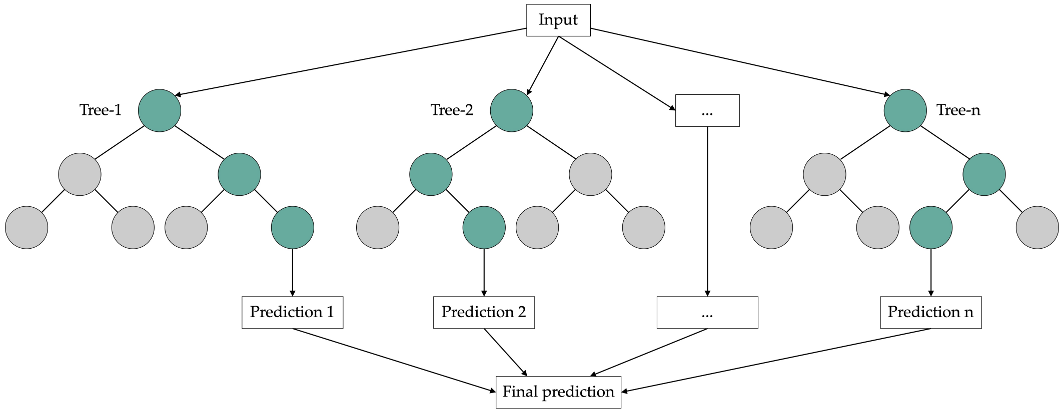


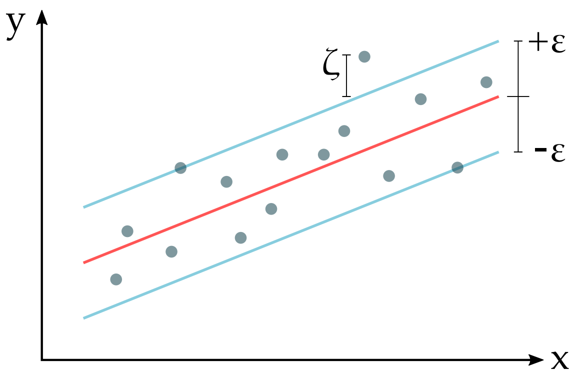

Image: Biljecki et al. (2016)

# Machine learning algorithms

1. Random Forest (RF)
2. Gradient Boosting (GB)
3. Suppor Vector Regression (SVR)




Random Forest




Linear SVR

# Implementation details

# Software & tools

## 1. Database

PostgreSQL

PostGIS

## 2. Processing

python™

pandas

scikit learn

## 3. Visualisation

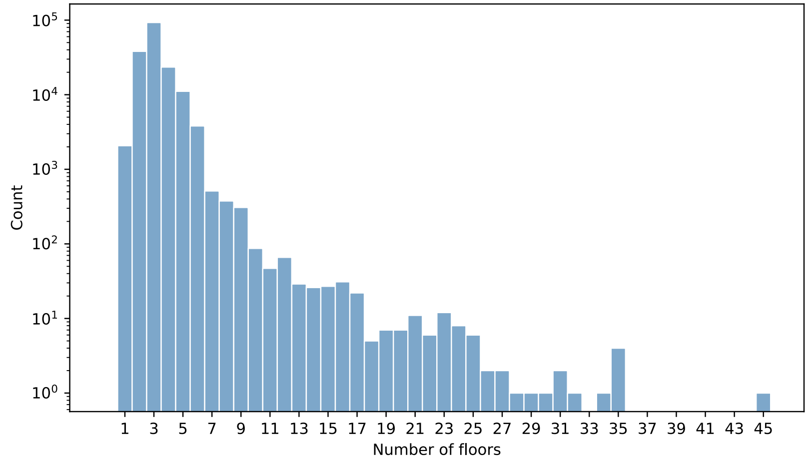
matplotlib

seaborn

# Datasets

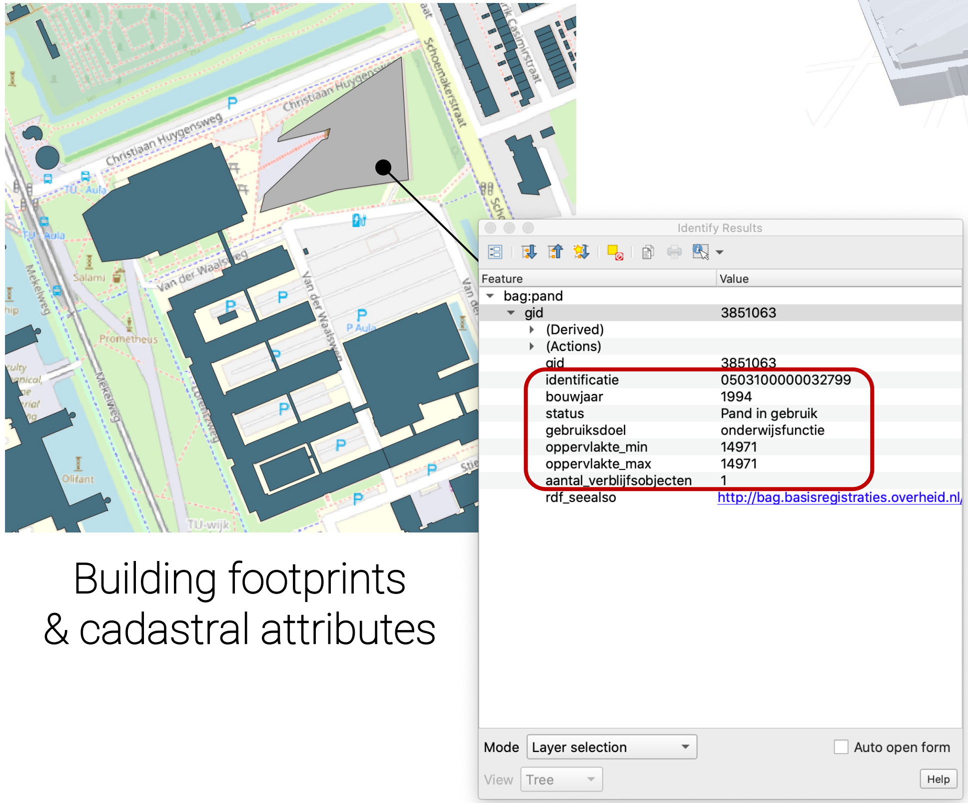
## Data on number of floors

| Municipality | # buildings |
|---|---|
| Amsterdam | 14,341 |
| Rotterdam | 116,638 |
| Den Haag | 53,559 |
| Rijssen-Holten | 11,516 |
| Total | 196,054 |

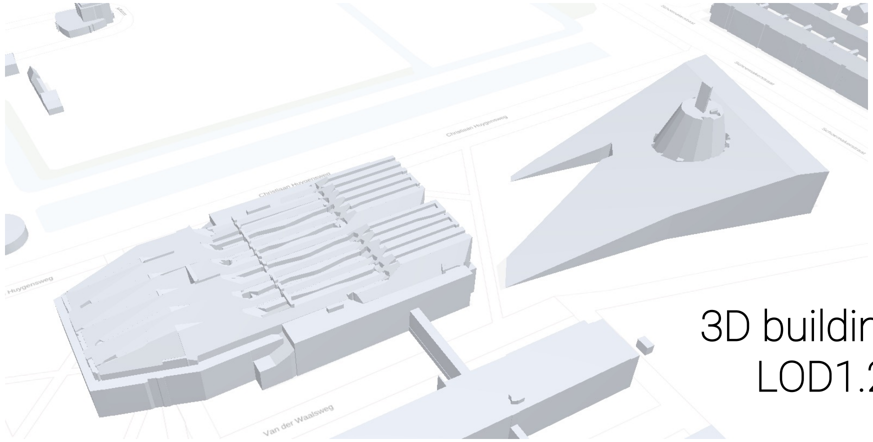
# Datasets

## BAG



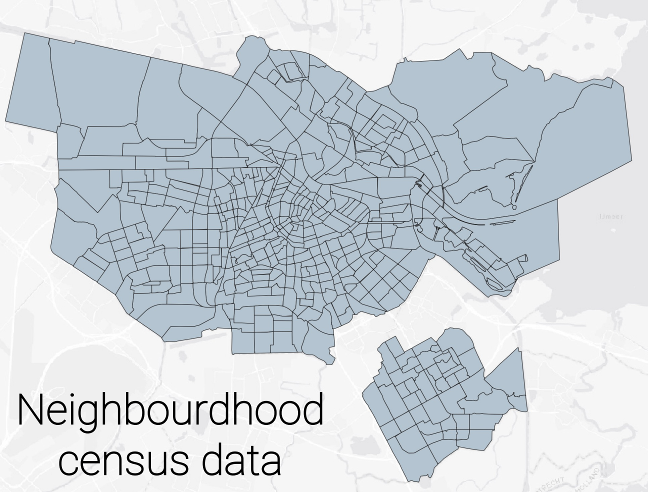

Building footprints
& cadastral attributes

## 3D BAG



3D building models in LOD1.2, 1.3, 2.2

## CBS wijken en buurten



Neighbourhood census data

TU Delft

- Research objectives
- Background
- Methodology
- **Implementation**
- Results
- Conclusions

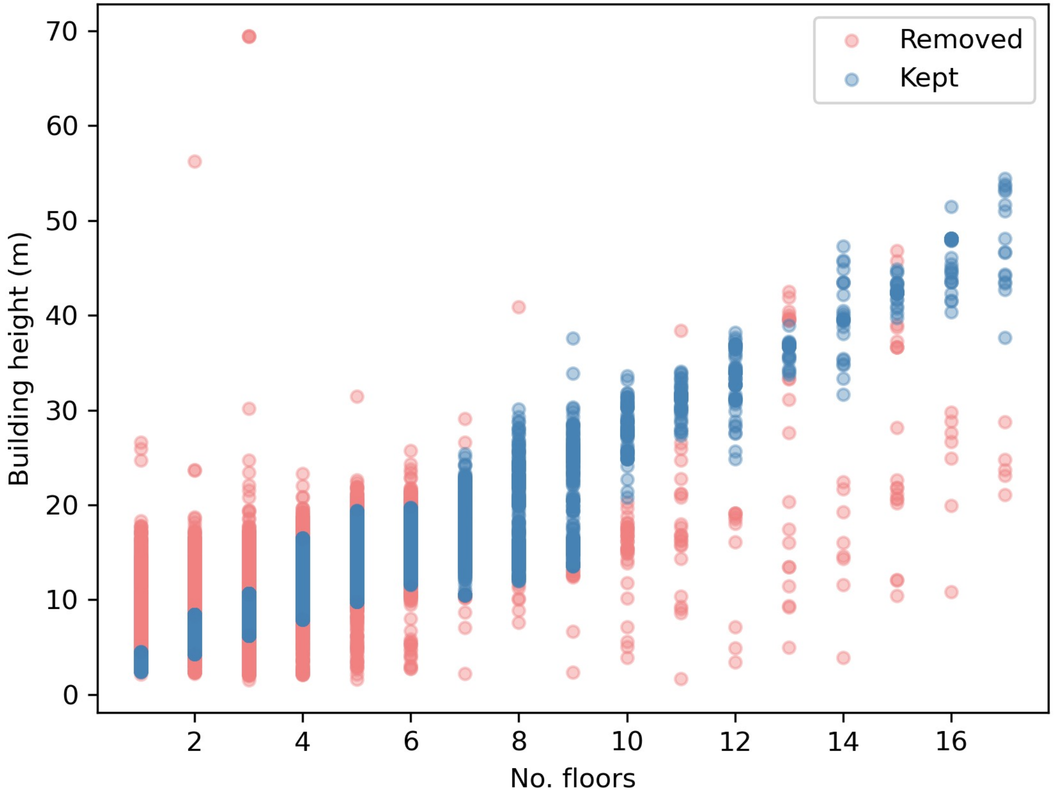
# Data cleaning

Approximately **11%** of mixed-residential buildings removed

| Cleaning step | # buildings removed |
|---|---|
| Automatic | 11,383 |
| Semi-automatic | 11,394 |
| Manual | 27 |
| Total | 22,804 |

TU Delft

# Data cleaning




Actual: 4 floors
Training data: 2 floors

Actual: 5 floors
Training data: 2 floors

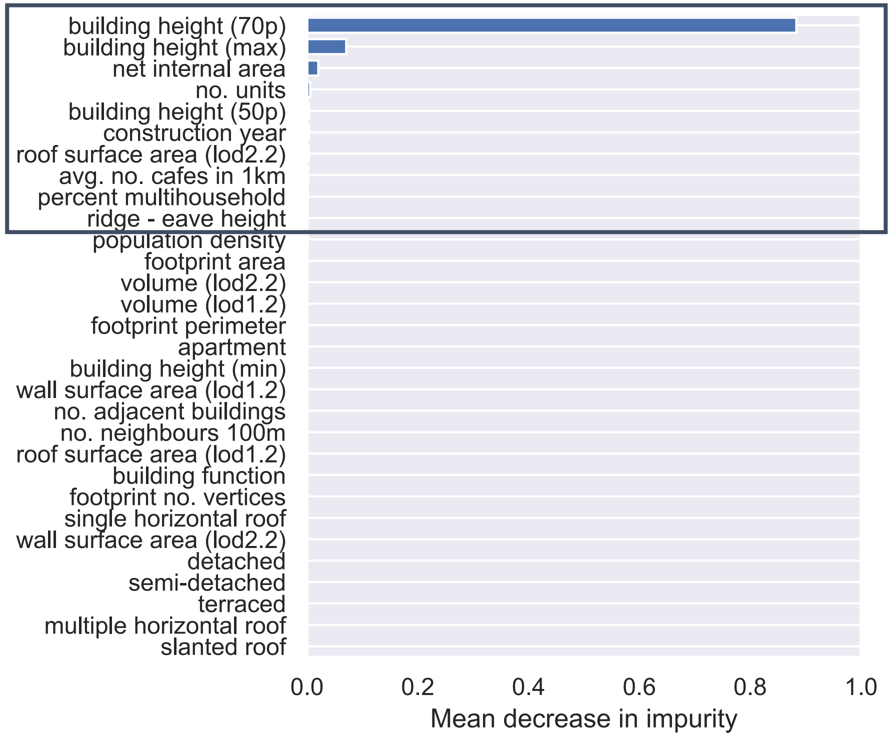
# Feature selection

## Method 1: Filter-based

| | Feature subset |
|---|---|
| 1 | Height (70th) |
| 2 | Height (max) |
| 3 | Height (50th) |
| 4 | Roof area (LOD1.2) |
| 5 | Roof area (LOD2.2) |
| 6 | Net internal area |
| 7 | Volume (LOD1.2) |
| 8 | Volume (LOD2.2) |
| 9 | Population density |
| 10 | % multihousehold |

## Method 2: Embedded

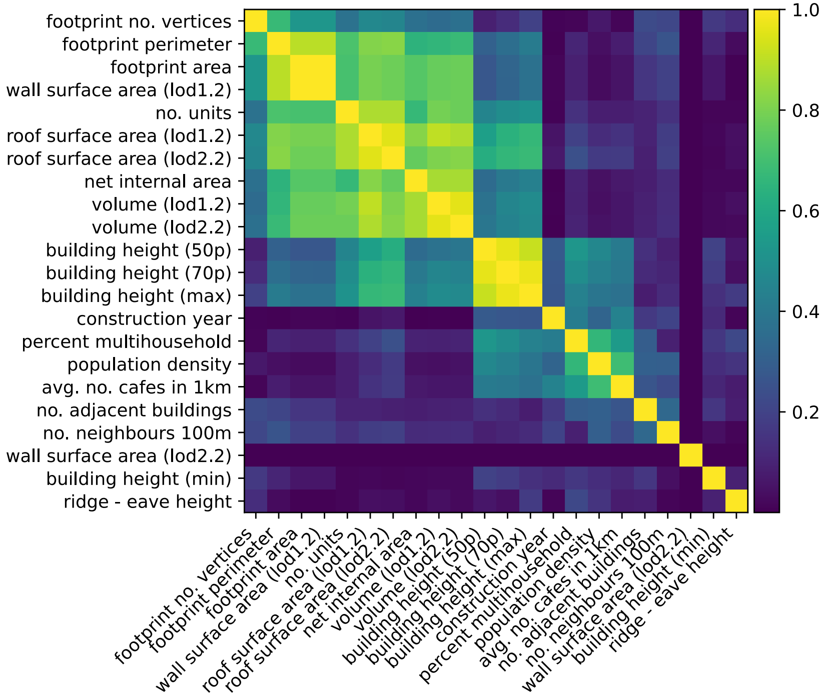
Feature importance GBR



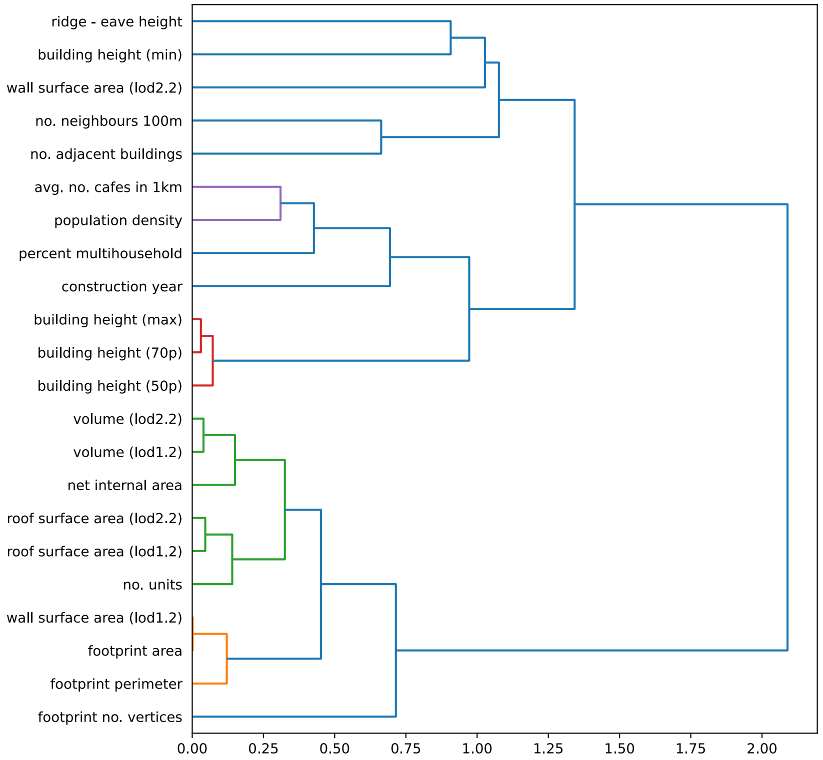

TUDelft

# Feature selection

- Method 3: Multicollinearity reduction




Pearson's correlation coefficient




Hierarchical clustering

# Results

# Model performance

## Gradient Boosting Regression (GBR)

| | MAE | | RMSE | | Max. error | | Accuracy (%) | | Training time (s) |
|---|---|---|---|---|---|---|---|---|---|
| | > 5 | ≤ 5 | > 5 | ≤ 5 | > 5 | ≤ 5 | > 5 | ≤ 5 | |
| All | 0.98 | 0.11 | 1.29 | 0.33 | 4 | 2 | 25.6 | 89.6 | 71.15 |
| Subset 1 | 1.04 | 0.11 | 1.36 | 0.33 | 5 | 2 | 24.1 | 89.4 | 40.71 |
| Subset 2 | 0.98 | 0.11 | 1.31 | 0.33 | 5 | 2 | 27.0 | 89.6 | 31.99 |
| Subset 3 | 1.02 | 0.12 | 1.33 | 0.35 | 4 | 2 | 23.7 | 88.4 | 35.08 |

MAE = Mean Absolute Error

RMSE = Root Mean Square Error

- Research objectives
- Background
- Methodology
- Implementation
- **Results**
- Conclusions

# Model performance

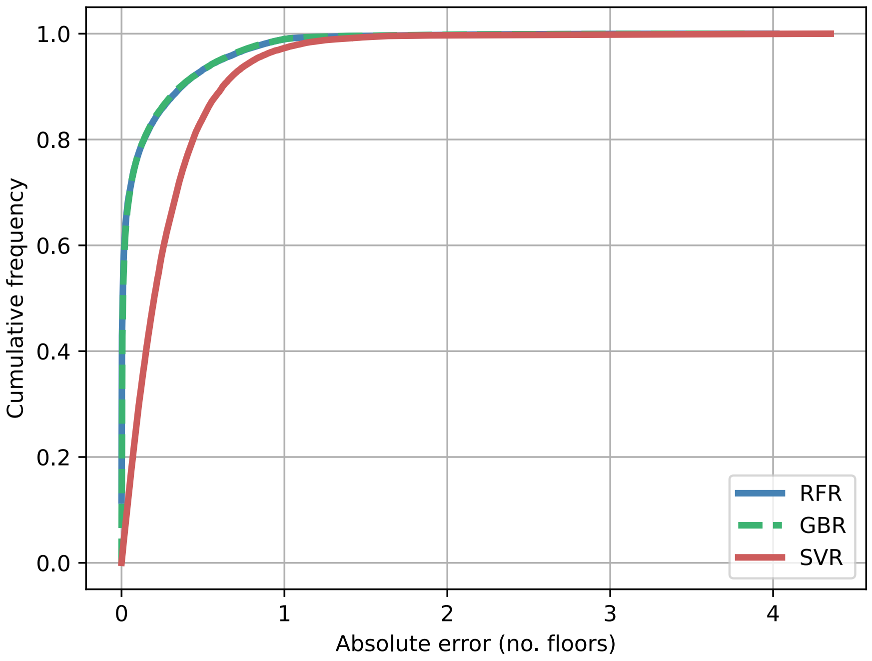
## Before and after hyperparameter tuning

| GBR subset 2 | MAE | | RMSE | | Max. error | | Accuracy | |
|---|---|---|---|---|---|---|---|---|
| | > 5 | ≤ 5 | > 5 | ≤ 5 | > 5 | ≤ 5 | > 5 | ≤ 5 |
| Untuned | 0.98 | 0.11 | 1.31 | 0.33 | 5 | 2 | 27.0 | 89.6 |
| Tuned | 0.62 | 0.06 | 1.00 | 0.24 | 4 | 3 | 52.3 | 94.5 |

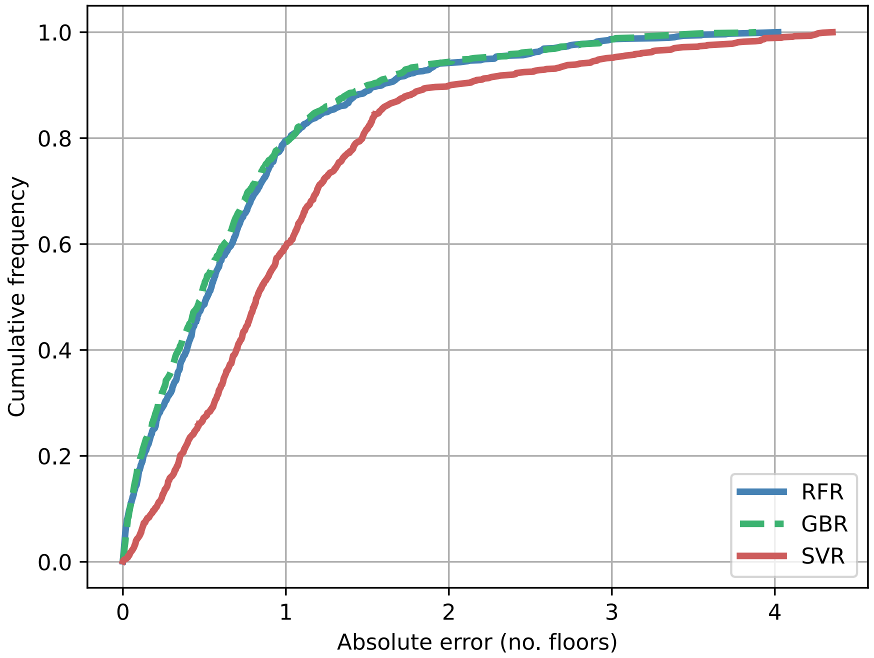
TU Delft

# Cumulative error analysis






All buildings




> 5 floors

TU Delft

# Gross errors

Incorrectly labelled


Actual: 5 floors

Label: 9 floors

ML: 5 floors

Exceptionally tall storeys

Actual: 4 floors

Label: 4 floors

ML: 7 floors

High storey apartments

Actual: 14 floors

Label: 15 floors

ML: 17 floors

TUDelft

# Comparison to geometric approach

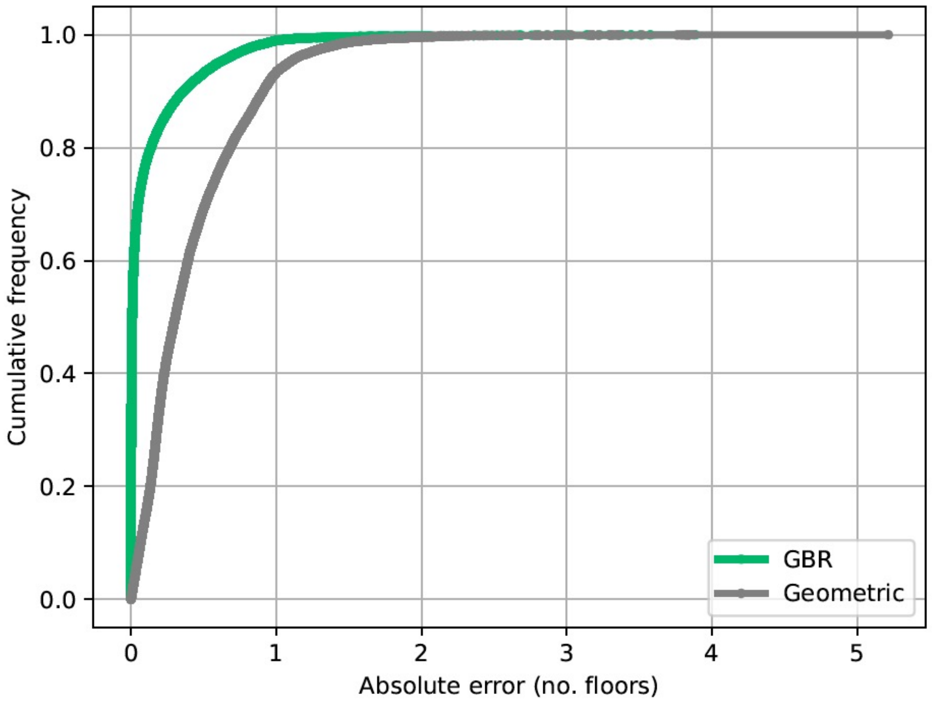
- Buildings with 5 floors or less

| | MAE | RMSE | Max. error | Accuracy (%) |
|---|---|---|---|---|
| GBR | 0.06 | 0.24 | 3 | 94.5 |
| Geometric | 0.31 | 0.31 | 2 | 69.9 |

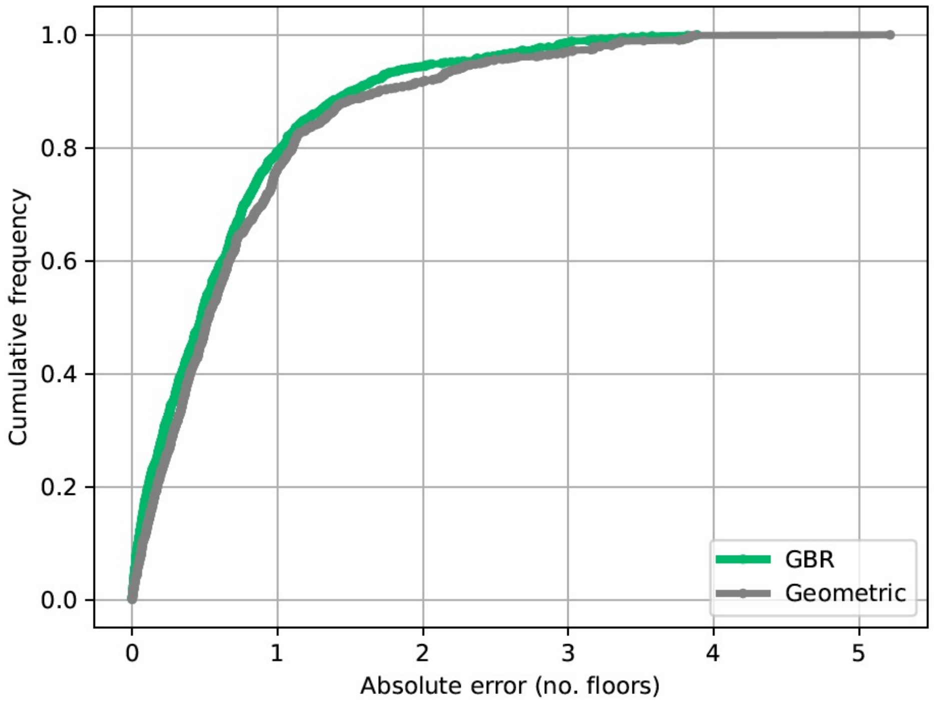
- Buildings above 5 floors

| | MAE | RMSE | Max. error | Accuracy (%) |
|---|---|---|---|---|
| GBR | 0.62 | 1.00 | 4 | 52.3 |
| Geometric | 0.70 | 1.09 | 5 | 47.5 |

# Comparison to geometric approach



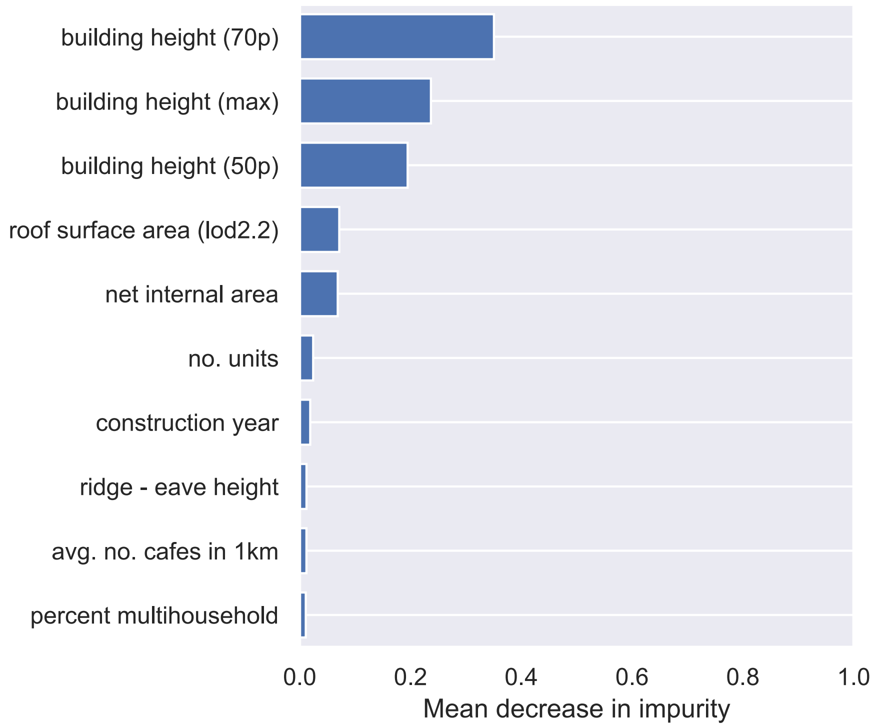

All buildings



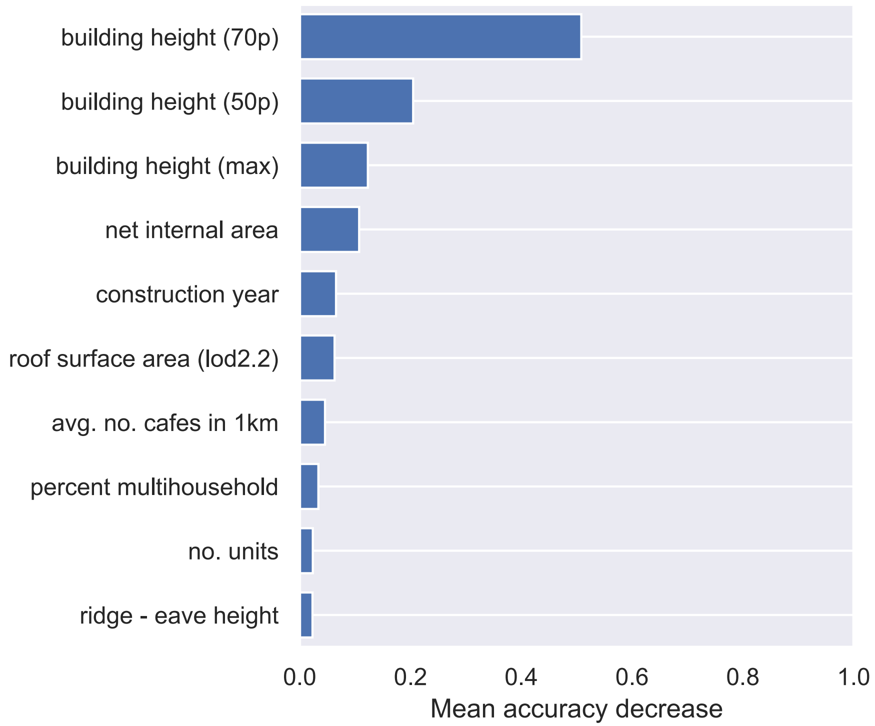

> 5 floors

# Feature contributions

- Research objectives
- Background
- Methodology
- Implementation
- **Results**
- Conclusions




Impurity importance




Permutation importance

TU Delft

# Impact of data availability

| | | MAE | | Accuracy (%) | |
|---|---|---|---|---|---|
| | | > 5 | ≤ 5 | > 5 | ≤ 5 |
| All features | | 0.64 | 0.05 | 51.7 | 94.8 |
| Cadastral | | 1.35 | 0.19 | 25.3 | 82.5 |
| Geometric | 2D | 2.23 | 0.39 | 5.8 | 65.2 |
| | LOD1.2 | 0.89 | 0.10 | 32.5 | 90.1 |
| | LOD2.2 | 0.87 | 0.10 | 34.8 | 90.5 |
| Census | | 2.55 | 0.41 | 3.6 | 61.7 |
| | | | | | |
| Subset 2 | | 0.62 | 0.06 | 52.3 | 94.5 |

TU Delft

# Conclusions

# Main research question

To what extent can **machine learning** provide a **better estimate** of the number of floors than a **purely geometric** approach?

Partially

| Low storey | High storey |
|---|---|
| Substantially better | Little improvement |

TUDelft

# Sub-question (a) – features

- Most relevant feature: building height
- Other relevant features: volume, roof area, net internal area
- Overlaps: height references and 3D geometric features
- Best subset: unclear

TU Delft

# Sub-question (b) – model results



- Best algorithm: Gradient Boosting
- Effect of data availability:
  - Best performance: 3D geometric features (LOD1.2 or LOD2.2)
  - Worst performance: 2D geometric and census features

# Sub-question (c) – comparison to geometric approach

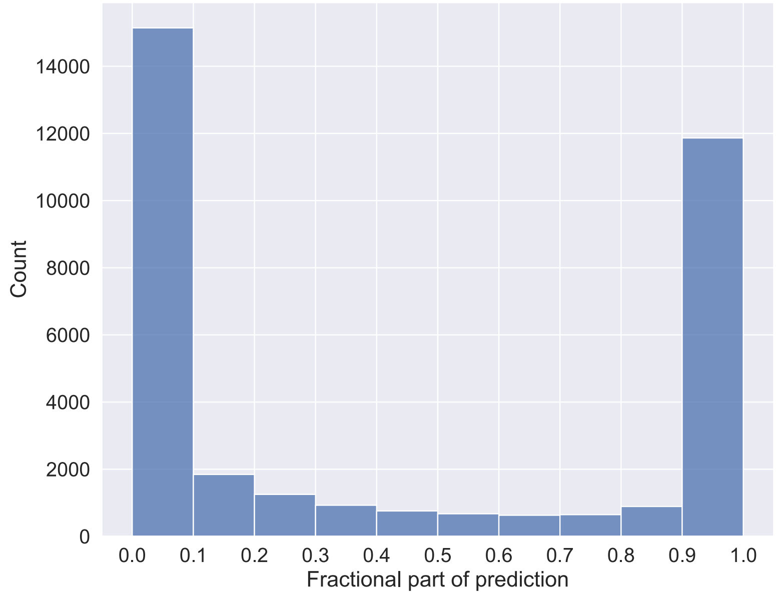
- Research objectives
- Background
- Methodology
- Implementation
- Results
- **Conclusions**

- Buildings ≤ 5 floors: 25% higher accuracy
- Buildings > 5 floors: 5% higher accuracy
- Cumulative error distributions: comparable for errors ≥ 1 floor
- Gross errors: mainly incorrectly labelled buildings

TU Delft

# Sub-question (d) – regression vs. classification

- Regression or classification? Regression
- Impact of rounding? Little impact on results




TU Delft

# Contributions

- Reliability of geometric approach
- Analysis of contributing factors
- Regression vs. classification

TU Delft

# Limitations

- Training data
- Data cleaning
- Feature selection
- Model performance and results

TUDelft

- Research objectives
- Background
- Methodology
- Implementation
- Results
- Conclusions

# Future work

- Wider model applicability
- Automatic data correction
- Analysis of input features
- Improved geometric approach

TU Delft

Thank you for your attention!

# References


Biljecki, F., Ledoux, H., and Stoter, J. (2014). Height references of CityGML LOD1 buildings and their influence on applications.

Biljecki, F., Ledoux, H., and Stoter, J. (2016). An improved LOD specification for 3D building models. Computers, Environment and Urban Systems, 59:25 – 37.

Biljecki, F., Ledoux, H., and Stoter, J. (2017). Generating 3D city models without elevation data. Computers, Environment and Urban Systems, 64:1–18.

Biljecki, F. and Dehbi, Y. (2019). RAISE THE ROOF: TOWARDS GENERATING LOD2 MODELS WITHOUT AERIAL SURVEYS USING MACHINE LEARNING. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, IV-4/W8:27–34.

Géron, A. (2019). Hands-on machine learning with Scikit-learn, Keras, and Tensorflow. O'Reilly Media, Inc.

Ellenkamp, Y. and Rietdijk, M. (2010). Kwaliteit van de basisregistraties adressenen gebouwen. Technical report, Ministerie van Volkshuisvesting, Ruimtelijke Ordening en Milieubeheer.


TU Delft