CHALMERS | GÖTEBORGS UNIVERSITET

MASTER'S THESIS

Machine Learning for NCC's Concrete Pile Production

Konrad Pohl

Department of Mathematical Statistics
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2021

Thesis for the degree of Master of Science

Machine Learning for NCC's Concrete Pile Production

Konrad Pohl

Supervisors: Holger Rootzen, University of Gothenburg
Rasmus Rempling, NCC
Jonas Magnusson, NCC

Collaborator: Deco Josephson, Karlstad University

Department of Mathematical Statistics
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2021

Abstract

In this thesis, the usefulness of machine learning (ML) is evaluated for the processes of NCC's subsidiary company Hercules. It is evaluated with regard to ML's ability to assist reduction of CO_2 footprint and costs. The work comprises analysis of Hercules' processes and analysis of data from these processes as well as a search for an appropriate ML model for predicting compressive strength of concrete. Results show that gradient boosted trees through the CatBoost library is a suitable ML model. However, additional data is needed to develop any such an ML model that is fit for use. A general example of how the CatBoost library can be used to predict strength of concrete is given. This example can be used as a starting point for future work on predicting compressive strength of concrete and for other ML problems at NCC. It was also found that Hercules' logistical system would benefit from further investigation. Short order times stresses the organisation and there may be a case for ML to improve the logistical system.

Acknowledgements

I wish to express my deepest gratitude for all support recieved during this project.

Thank you Holger Rootzen for your insightful and compassionate supervision. Thank you Rasmus Rempling and Jonas Magnusson for having this interesting project and for your support. Thank you Deco Josephson for our collaboration and times together. Thank you Umberto Picchini for examining me. Thank you Peter Elmgren, Iad Saleh and workers on the shop floor for taking time to answer questions about the pile production. Thank you Louise Westin Elger for your continuous and untiring support during this project. Finally, thanks to all friends and family that have supported me with questions about the proceedings of the project.

I also wish to commemorate this work to loved friends and family that have passed or have been close to pass away.

Notations and Abbreviations

Mathematical notations

j - index of predictive variable
i - row index, specifies an observation
$\hat{i}$ - estimated improvement
R - Region
P - Set of all predictive variables
p - Size of P
2^P - Power set of P
S - Subset of P
s - size of S
v - value function in a coalition game
x - observed values of predictive variable
y - observed values of outcome variable
$\hat{y}$ - estimated values of outcome
$\bar{y}$ - mean of y
$\emptyset$ - The empty set
$\sum$ - sum operator, sums all variables which fulfills a logical condition as specified in its indices.

Mathematical abbreviations

ANN - Artificial Neural Network, algorithm
CART - Classification AND Regression Tree
DNN - Deep Neural Network
GA-ANFIS - Generic Algorithm-Adaptive Neuro Fuzzy Inference System
kNN - k Nearest Neighbours
MAPE - Mean Absolute Percentage Error
MARS - Multivariate Adaptive Regression Splines
MLP - Multilayer Perceptron
MLR - Multiple Linear Regression
Resnet - Residual Neural Network
RF - Random Forest
RMSE - Root Mean Square Error
R^2 - Coefficient of determination
SHAP - SHapley Additive exPlanations
SVM - Support Vector Machine

Further notations

f_c - Compressive strength of concrete
f_{ck} - Characteristic strength of concrete
$\bar{f}_{ct}$ - Mean value of strength test from the last t days
Fillers - Small size particles in concrete. Rock dust is a type of filler that is usually unwanted while fly ash and blast furnace slag are added intentionally.
Hercules - NCC's subsidiary company specialized in building foundations.
Internal rate - Expected return on invested money in the company.
Metodia - Software used by Hercules to estimate compressive strength development based on temperature development in the concrete.
Tied up capital - Expenses of a company that are yet to be realized as revenue. Tied up capital is regarded as expenses through the company's *internal rate* and the time the capital is tied up.
Ucklum - Production site of concrete piles.

Further abbreviations

CA - Coarse Aggregates
FA - Fine Aggregates
ML - Machine Learning
TQM - Total Quality Management
W/C - Water Cement ratio
W/P - Water Particle ratio

Contents

1 Introduction

As concrete has a large climate footprint and today's technologies provide means for sophisticated data analysis, this project has investigated these two subjects at one of the largest producers of concrete in Sweden.

1.1 Background

NCC's subsidiary company Hercules is the leading concrete pile producer in Sweden. They have production lines in two plants, one in Ucklum and one in Västerås. The production of the piles is to a large extent automated and thereby data is automatically gathered from the processes. Controlling the relationship between the production's in-parameters and quality measures of the piles is of importance for development of the production line of tomorrow. Enabling prediction of quality measures would help create better quality at a lower cost.

NCC was also interested in examining the possibilities of predicting quality measures of piles of theoretical material compositions for possible future use. Such predictions would give means to reduce the proportion of cement in the concrete piles. Cement is an expensive material, both with regards to climate footprint and costs. Cement comprises up to 90% of CO_2-emissions of concrete [6] and the cement industry is liable for 7% of the global CO_2-emissions [4]. However, the carbon footprint of cement can be lowered by partially substituting cement with waste products from other industries such as fly ash and blast furnace slag. Cement also comprises 2/3 of the material costs in the concrete mixed in Hercules Ucklum. However, the principal motivation for the project was the climate aspect as "reducing the climate footprint is essential for NCC's survival" [1].

1.1.1 Financial reference points

Hercules' production plant in Ucklum was studied in this report. They had the following financial highlights 2020:

- Carbon footprint of cement: 3 700 tonnes CO_2-eq.[2]
- Total carbon footprint of piles: 6 700 tonnes CO_2-eq [3]
- Material costs of cement: 5 Msek
- Net billing: 55 Msek

[1] Rasmus Rempling, coordinator of the project

[2] Carbon footprint of cement = 0.581 CO_2-eq/tonne [10]. Mass of cement obtained from Mixomemory in Ucklum.

[3] Carbon footprint of piles = (24.7+33.9)/2 CO_2-eq/m [21].

For reference, Sweden's total carbon footprint of 2020 has preliminary been estimated to 47,4 Mtonnes CO_2-eq [22]. Hercules Ucklum's carbon footprints then makes up 0.008% and 0.014% of Sweden's carbon footprint, for cement only and total emissions from produced piles.

1.1.2 Concrete piles

Concrete piles are typically between 5-14 m long, and 235-350 mm wide, steel reinforced elements used to stabilize building foundations. The piles are hammered down the ground to transfer the load of a building deep into the earth, preferably all the way to the bedrock. Piles may be joined together to reach depths of up to 70 m. Figure 1 and figure 2 show concrete piles, ready for delivery and at installation.

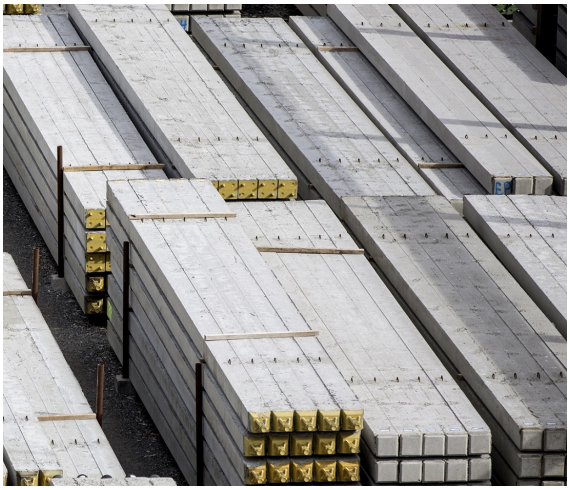


Figure 1: Concrete piles in storage [1].

Figure 2: Concrete piles at installation [1].

Concrete piles are generally the most cost efficient type of piles for deep foundation. Raw materials are inexpensive and the procedure of forming concrete is advantageous compared to other materials with similar load bearing properties. Concrete mixtures contain water, cement and aggregates and solidify in about 24 hours. For specific physical properties of the concrete, admixtures can be added. Admixtures are chemical compounds which can be used to control the rate of strength development and air content of concrete. See [12] for further reference.

1.1.3 Outline of the concrete pile production

The pile production revolves around two major activities, mixing concrete and steel assembly. An overview of the processes and material flow of the concrete pile production is presented in fig. 3

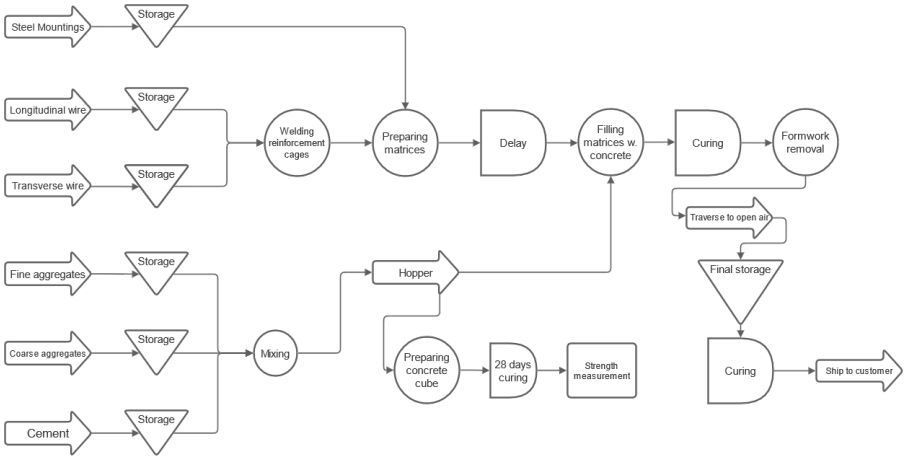


Figure 3: Overview of the processes and material flow of the concrete pile production.

1.2 Purpose

The purpose of this project is for NCC and Hercules to gain knowledge of parameters from the production of concrete piles that influence the quality of the piles. This is to enable control of the piles' carbon footprint and costs. Influential parameters were expected to originate from the following areas; material composition, manufacturing environment, manufacturing processes and the installment process of the concrete piles. Knowledge about these parameters can act as a basis for making data driven, or fact-based decisions and contribute to improved quality and lower CO_2-emissions and lower costs of the concrete piles.

1.3 Aims

The aim of the project was to:

1. Find a machine learning (ML)-model that predicts quality measures related to the production of concrete piles.
2. Enable virtual testing of possible new material compositions, with focus on compositions where fly ash partly substitutes cement.

1.4 Specified purpose

The purpose is specified by the following questions:

1. What measures are relevant for concrete pile quality?
2. What is the present state of data collection in Hercules' processes?

3. Which ML-algorithms are appropriate to predict relevant quality measures in Hercules plant in Ucklum?

4. Does the present state of data collection in Hercules' processes allow for ML?

2 Method

The whole project was composed as a proposal and demonstration of using machine learning for quality control of NCC's processes.

2.1 Methods for the specified purpose

Work methods used are literature studies, unstructured interviews and observations in the plant in Ucklum. The following list relates the work methods used to answer the questions in the specified purpose.

1. What measures are relevant for concrete pile quality?
   -Literature studies, unstructured interviews, observations in the plant in Ucklum

2. What properties of an ML-model would benefit NCC and Hercules?
   -Unstructured interviews

3. What is the present-day state of data collection in Hercules' processes today?
   -Unstructured interviews, observations in the plant in Ucklum

   (a) What relations are viable in this data?
   -Unstructured interviews, observations in the plant in Ucklum

   (b) What effects the reliability of this data?
   -Literature studies, unstructured interviews, observations in the plant in Ucklum.

4. Which ML-algorithms are appropriate to predict relevant quality measures?
   -Literature studies

2.2 Work procedure

Projects of this type, with many interacting uncertainties are most suitable to agile project management. Therefore, an iterative cycle of methods was used: Unstructured interviews enabled data acquisition, which in turn enabled data analysis, which in turn enabled development of ML-models. Completed ML-models were then evaluated and the performance of ML-models provided ground for further interviews. Literature studies were carried out in parallel to interviews, data analysis, ML-model development and ML-model evaluation. This cycle is visualized in fig 5.

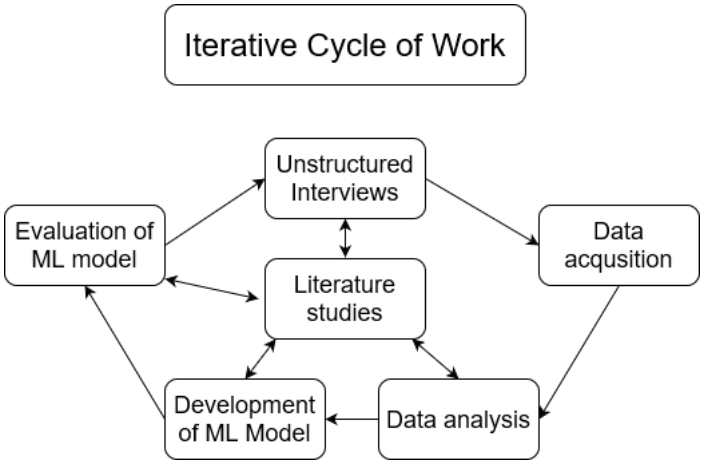


Figure 4: Iterative cycle of work methods for the project.

2.3 Participants

The project was organized as a joint master's thesis project, with one student from mathematical statistics and one from engineering mechanics. The following people participated in the project:

- Student from engineering mechanics. [12] refers to his work. (Karlstad University)
- Coordinator of ML-development (NCC)
- Senior Technical Specialists in concrete structures (NCC)
- Plant manager from Ucklum (Hercules)

3 Theory

This chapter contains the theoretical background for machine learning, a summary of the theory of concrete mixtures and an introduction to the subject of quality. The chapter begins with a framework for ML, and describes four important aspects to consider when dealing with ML; *data types, extrapolation, overfitting* and *learning speed*. Then, the mathematical theory behind two simple ML models, *linear models* and *classification and regression trees (CART)* are described. These models are part of *base learners*, simple ML models which may be refined through further algorithms. *Boosting* is such an algorithm, and section 3.4 accounts for how a CART model can be drastically improved through the *gradient boosted trees* algorithm. At the end of the chapter, section 3.7 describes how concrete piles are certified through the concrete's compressive strength. Lastly, section 3.8 gives an introduction to the theory of *quality* and *lean production*, a widespread production philosophy which NCC adopts.

3.1 Mathematical framework of ML

With some data, one wants to find a relation between some *features*, or *predictive variables/predictors* $X = (X_1, X_2, ...)$ and one *output variable* Y as this could increase understanding of a real-world phenomenon. One has a set of observed values (x, y) for the predictive variables and the output variable [4]. So each data point can be described as a vector of observations of a set of variables (x, y). The objective of ML is to describe the relation between predictor and output through some function[5] g:

$$y = g(x) \tag{1}$$

But, two important aspects must be considered. Firstly, the model needs to be useful. Just any mathematical function may not be a suitable representation of real-world phenomena. Secondly, the predictive data may not be reliable. One cannot expect to find a model which perfectly represents the real world. Both these aspects are dealt with by introducing an error term, ϵ:

$$y = g(x) + \epsilon \tag{2}$$

Then, separate the relevant relation g from the errors ϵ by acknowledging that we estimate y from x. Estimates of y are denoted $\hat{y}$ and are described by:

$$\hat{y} = \hat{g}(x) = y - \hat{\epsilon} \quad \Longleftrightarrow \quad \hat{\epsilon} = y - \hat{y} \tag{3}$$

[4]The setting with access to observed output values is called *supervised ML* and ML is not necessary constrained to settings where observed output values are available.

[5]Refer to g as "General ML model"

The facts that there is a variety of models out of which no one describes reality perfectly and that perfect data does not exist is commonly depicted by the following quote:

"All models are wrong, but some are useful"-George Box

Note now that the error term ϵ allows for evaluation of the usefulness of a model. By some quantifying function $q(\epsilon)$, the impact of the errors can be investigated and model usefulness can be assessed.

Notations used further down are:

- x denotes observed values of the predictive variable(s) X, in other words realizations of the random variable or vector X. In the same way, y are regarded realizations of Y.
- X_j, $p = 1, 2, ..., p$ denotes the j:th subscript of X if X is a vector. Similarly for x_j.
- (x_i, y_i), $i = 1, 2, ..., N$ denotes the i:th pair of observations (x, y).
- $\boldsymbol{x}$ denotes a $N \times p$ matrix of N observations and p features.
- "hat-values", e.g. $\hat{y}$ denote a parameter or function approximated from data.
- m denotes indices of various models g.

Concepts to consider when working with ML are:

- Data types. Measurements may be *qualitative* or *quantitative*. Consider two variables X_1 and X_2. X_1 describes a material type and takes values on $\{steel,\ plastic\}$, whereas X_2 describes temperature in an environment and may take values on $[10, 40]$. Then, X_1 holds qualitative information and X_2 holds quantitative information. The two types of data create two main types of problems in machine learning, *classification* problems and *regression* problems. In classification problems, the output variable is of qualitative data type and in regression problems, the output variable is quantitative.
- *Training* and *test* data. In order to improve on the model's ability to describe reality, the data is partitioned into training and testing sets: Data $= \{\boldsymbol{x}_{train}, y_{train}, \boldsymbol{x}_{test}, y_{test}\}$. Then, a set $\{\hat{g}_m\}$ of models are created by $\{\boldsymbol{x}_{train}, y_{train}\}$. Thereafter, models are tested by calculating $\{\hat{g}_m(\boldsymbol{x}_{test})\} = \{\hat{y}_{test\ m}\}$. This makes it possible to obtain approximations of the models' $\{\hat{g}_m\}$ ability to predict on "real-world-data", since $\hat{\epsilon}_{test} = y_{test} - \hat{y}_{test}$. It is also easy to calculate $\hat{\epsilon}_{train} = y_{train} - \hat{y}_{train}$, and by comparing $\hat{\epsilon}_{train}$ to $\hat{\epsilon}_{test}$, one can see if a model is *overfitted*,

i.e. the complexity of a model is unjustified, which leads to decreased training errors but increased test errors in comparison to a less complex model.

- *Extrapolation* is when a model is used on new data which does not lie in the interval of the training data, $\hat{g}(x_{new})$ and $x_{new} \notin [min(x_{train}), max(x_{train})]$. This can be dangerous as this model may not be useful in the region of x_{new}.

- *Learning speed* denotes how much data is needed for a model to perform optimally. There is usually a trade-off in model selection between how much data the model requires for finding its optimal fit and how well the model may perform with an abundance of data. In cases where data is expensive, it is desired to have a model which performs well with a small training data set.

3.2 Linear regression

Linear regression is a well-established starting point for working with ML. Due to its mathematical simplicity, it is easy to work with and easy to interpret [11]. *Simple linear regression* is fitting a straight line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = g(x) \tag{4}$$

to a data set, so that the distance between the line and the data points are minimized. The following notions can be derived from this concept:

- $RSS = \sum_i (\hat{y}_i - y_i)^2 = q(\epsilon)$. The smaller the residual sum of squares, the squared distance between observed and predicted values, the better the model fits the data. The distance between one observed and predicted value, $|\hat{y} - y|$, is called a *residual*.

- Finding the set of parameters $(\hat{\beta}_0, \hat{\beta}_1)$ that minimizes the RSS is done by differentiating RSS and setting the derivative equal to zero, $RSS'(\beta_0) = 0, RSS'(\beta_1) = 0$.

- One can see that observations far away from the center of the x_i-s will have greater influence on the estimated parameters $(\hat{\beta}_0, \hat{\beta}_1)$ than observations closer to the center.

General linear regression, can be viewed as an extension of the concept of simple linear regression. This extension may take three directions.

Firstly, *multiple linear regression* can be done using multiple predictive variables. With n predictive variables, this creates expressions for $\hat{y}$ as:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_n x_n \tag{5}$$

so, $\hat{y}$ is approximated by a n-dimensional *hyperplane* instead of a line.

Secondly, linear regression allows for transformations of either the predictive or the response variables, or both. For example, in the case of an exponential model where

$$log(y) = \beta_0 + \beta_1 x \tag{6}$$

is assumed, $\hat{y}$ may be estimated as

$$\hat{y} = e^{\hat{\beta}_0 + \hat{\beta}_1 x} \tag{7}$$

Thirdly, multiple linear regression may be used to do *polynomial regression*. Thereby, model complexity may be increased by increasing the degree of the polynomial. For a polynomial of degree p, this gives the expression:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_1 x^2 + ... + \hat{\beta}_p x^p \tag{8}$$

Note that equations 5 and 8 both describes multiple linear regression. In conclusion, as all these three directions of increased complexity may be combined, linear regression is not limited to two dimensions nor to straight lines or straight hyperplanes. Any regression where the underlying model is linear in the parameters, is a linear regression.

3.2.1 Pearson's correlation coefficient

Prior to starting with any ML problem, it is useful to visualize a summary of the data one intends to work with. One such visualization is to plot the estimated *correlation* between variables. Pearson's correlation coefficient ρ_{X_1,X_2} is commonly used. It is defined as:

$$\rho_{X_1,X_2} = \frac{\text{cov}(X_1, X_2)}{\sigma_{X1}, \sigma_{X2}} \tag{9}$$

for two random variables X_1 and X_2. Here, $\text{cov}(X_1, X_2)$ refers to the covariance defined as:

$$\text{cov}(X_1, X_2) = E[(X_1 - E[X_1])(X_2 - E[X_2])] = E[X_1 X_2] - E[X_1]E[X_2] \tag{10}$$

So ρ_{X_1,X_2} is the covariance normalized with regard to the standard deviations of X_1 and X_2. ρ_{X_1,X_2} takes values on $[-1, 1]$ and can be estimated from an existing data set. $\hat{\rho}_{X_1,X_2} = -1$ or $\hat{\rho}_{X_1,X_2} = 1$ implies that all data points lie perfectly on a line, a simple linear regression model describes the relation between X_1 and X_1 perfectly. $\hat{\rho}_{X_1,X_1} = 0$ implies that a simple linear model does not explain the relation between X_1 and X_1 at all. Low correlation is generally desired between predictive variables. Highly correlated variables usually contain the same underlying information and may cause problems in ML algorithms. Hence, adding an predictive variable which correlates with another predictive variable is undesirable.

3.3 Decision Trees - CART

Classification and Regression Trees (CART) or *decision trees* is another type of simple ML method. CART is also considered an ideal base for Boosting algorithms. ML problems may constitute a classification or regression problem based on *qualitative* or *quantitative* input data. One advantage of CART models are that they can be used on all such problems and that it easily transitions between them. [11]. Let A be a logical statement and I be the indicator function:

$$I(A) = 1 \text{ If A is a true statement, } I(A_{false}) = 0 \text{ f A is false.} \tag{11}$$

further, let $\{R_m\}$ be a set of disjoint rectangular regions and $\{c_m\}$ be a set of constants for some index set m. Then CART can conceptually be described by the following estimation of the response variable:

$$\hat{Y} = \sum_m c_m I(X \in R_m) \tag{12}$$

This translates to: "If the predictive variable X is in the rectangular region R_m, set $\hat{Y} = c_m$".

However, the regions R_m and the constants c_m must be obtained according to some optimization of a performance measure before a CART model constitutes meaningful ML. This section describes how decision trees are created by first describing them in their simplest forms, *decision stumps*, and then how the stumps can be joined into decision trees.

3.3.1 Decision stumps

In their simplest forms, decision trees only split the feature space into two regions: R_1 and R_2, i.e $m = 2$, through one single *split* in the feature space. Such decision trees are called decision stumps. The method for determining the optimal regions R_1 and R_2 is to :

1. Split the feature space into two regions R_1 and R_2:

$$R_1(j, s) = \{X | X_j \leq s\} \text{ and } R_2(j, s) = \{X | X_j > s\} \tag{13}$$

   for a splitting variable j and a split point s.

2. Evaluate the regions according to some performance measure, typically the *squared loss* for regression trees, see section 3.3.3 further down. See e.g. *Gini index* in [11] for classification trees.

3. Select the regions R_1 and R_2 with the best performance measure.

3.3.2 Decision trees

Knowing how to make a decision stump, or split, it is possible to create many splits and hence to create decision trees. Decision trees grow in complexity with increased amount of splits, but also with the amount of input parameters and the amount of data points, as these lead to more regions to evaluate and more calculations to make. However, the procedure is in principle a repetitive process of creating decision stumps. The amount of splits in the tree is usually determined by some *stopping criteria* for the algorithm, this could be a predetermined amount of splits or a minimum amount of data in created regions or something else. The following pseudo-algorithm describes how to build a decision tree top-down:

1. Initiate a counting variable: $a = 1$
2. Consider the the space of the input variables as a set of regions as at step q: $\{R_{ma}\}$. ma is the index set of the regions at step a.
3. While(stopping criteria is not reached)
   (a) set $a = a + 1$
   (b) Let the set of input regions $\{R_{ma}\}$ be the set of regions created by making decision stumps in each of the regions in $\{R_{ma-1}\}$.

Note that by this algorithm, the splits created are dependent on splits made in previous iterations. This implies that the algorithm is not guaranteed to find the set of regions $\{R_m\}$ with the best global fit to the data, but that it is proceeding according to the best local fit in each iteration. To improve the tree further, techniques for removing redundant splits can be applied. Such techniques are referred to as *pruning*.

3.3.3 Regression trees and squared loss

In regression, the constants c_m in equation 12 are set as the mean value of y in region R_m:

$$\hat{y}_m = c_m = \frac{1}{N_m} \sum_{y_i \in R_m} y_i \tag{14}$$

Further, the regions R_m are selected to minimize the selected loss function. The squared loss function, which is commonly used, is:

$$L(y, \hat{y}) = \sum_{y_i \in R_m} (y_i - \hat{y}_i)^2 \tag{15}$$

which is the same as the RSS seen in linear regression.

3.4 Boosting

The main idea that makes up *boosting* is developing "a set of weak learners that create a single strong learner" and that each recursion in the algorithm should learn from previous steps. This section aims to give an idea of how this works through describing the *gradient boosted trees* algorithm.

3.5 Gradient Boosting

Gradient Boosting builds on the idea that parameters of a regression or classification model can be generated by moving iteratively in the direction in which errors are decreasing in the fastest pace.

Consider a variable γ, a regression tree denoted $g(x)$ as a base learner, with $J-1$ allowed splits in the tree and a loss function $L(y_i, g(x_i)) = \frac{1}{2}(y_i - g(x_i))^2$. The negative derivative of the loss function is then equal to residuals $-\partial L(y_i, g(x_i))/\partial g(x_i) = r_i = y_i - g(x_i)$. Let M be the number of iterations for the algorithm and N the amount of observations. Gradient boosted trees for regression are then constructed as follows:

1. Initialize $g_0(x) = arg\, min_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$

2. For $m = 1$ to M:

   (a) For $i = 1$ to N, compute residuals:

   $$r_{im} = -[\partial L(y_i, g_{m-1}(x_i))/\partial g_{m-1}(x_i)] = y_i - g_{m-1}(x_i)$$

   (b) Fit a regression tree to the residuals r_{im}, giving terminal regions R_{jm}, $j = 1$ to J

   (c) For $j = 1$ to J, compute

   $$\gamma_{jm} = arg\, min_\gamma \sum_{x_i \in R_{jm}}^{N} L(y_i, g_{m-1}(x_i) + \gamma)$$

   (d) Update $g_m(x) = g_{m-1}(x) + \sum_{j=1}^{J} \gamma_{jm} I(x_i \in R_{jm})$

3. Output $g(x) = g_M(x)$

Note that by this algorithm, $g(x)$ gets constructed through sums of different γ-values, and the γ-values are just the means of the residuals in the regions they minimize the loss in, R_{jm}. The algorithm starts with the first tree predicting the output equal to the mean of all the observed output values and then continues to add different γ-values to different regions.

Note also that this description of the Gradient boosting algorithm is simplified by

- Stating J as constant. However, J is usually defined as the maximum allowed number of splits in the tree and may therefore vary. J is commonly in the interval $4 \leq J \leq 8$ and variations of J are then caused by early stopping of the tree building.

- Not accounting for general types of loss functions. Although $L(y_i, g(x_i)) = L(y_i, g(x_i)) = \frac{1}{2}(y_i - g(x_i))^2$ is commonly used, the algorithm allows for other loss functions too. Residuals then gets exchanged to "pseudo residuals", still fulfilling the same purpose and still calculated as $r_i = -[\partial L(y_i, g(x_i))/\partial g(x_i)]$.

- Not accounting for *shrinkage*. The test error of gradient boosting models gets improved by slowing down the rate by which the algorithm gets fitted to the training data. This is carried out by scaling the step size in line 2 (d) by a factor $\nu \in [0, 1]$:

$$g_m(x) = g_{m-1}(x) + \nu \sum_{j=1}^{J} \gamma_j I(x_i \in R_{jm}) \tag{16}$$

  Where $\nu < 0.1$ is common in practise. However, M and ν are not independent. The smaller ν, the more iterations of the algorithm are needed.

[11]

Relative variable importance One tool for interpreting models is through calculations of *relative variable importance*. This measure describes how much each of the predictive variables contributes to model output. For a predictive variable X_ℓ in a tree T, the variable importance is calculated as:

$$\mathcal{I}_\ell^2(T) = \sum_{t=1}^{J-1} \hat{i}_t^2 I(v(t) = \ell) \tag{17}$$

Here, J is the number of splits in the tree. $\hat{i}_t^2$ is the improvement in performance measure, squared error loss, from split t in the tree. $v(t)$ is a function which outputs the index of the predictive variable that gives the best improvement $\hat{i}_t^2$. Then for boosted trees, the variable importance of variable ℓ is calculated as:

$$\mathcal{I}_\ell^2 = \frac{1}{M} \sum_{m=1}^{M} \mathcal{I}_\ell^2(T)_m \tag{18}$$

[11] A visualization of outputted relative variable importance is given in fig. 30. Such plots describes how much each type of input data has contributed to creating the model. When inferring about the importance

of used input variables, note that a relative variable importance plot sums up to 1, or 100%. A variable with high reported variable importance may not be important if the model's predictive ability is low. Note also that a relative variable importance plot depends on the distribution of the data provided to the model. A narrow distribution will cause a lower relative variable importance, resulting in potentially undeserved low variable importance values. Or vice versa for variables with wide distributions in the input data.

3.6 SHAP values

A recently developed tool for model interpretation are SHapley Additive exPlanation values (SHAP)[15]. This section aims to introduce this concept. SHAP values are, just as the relative variable importance, a measure of each of the predictive variable's contribution to the model's output. However, they make deeper understanding of the model possible by describing how much each observation x_i contributes to the models output $\hat{y}_i$ in the native unit of the output y. SHAP values can be based on *Shapley values* alone, but may require combinations of Shapley values and other methods due to issues with computational complexity. Shapley values are derived from *game theory* and are constructed as follows:

$$\phi_j(g,x) = \sum_{S \subseteq P \setminus j} \frac{s!(p-s-1)!}{p!}[g_x(S \cup j) - g_x(S)] \tag{19}$$

This looks a bit difficult, but the following list will break down the expression into parts and then describe it with an intuitive interpretation:

- $\phi_i(g,x)$ is the Shapley values of the j:th feature, in learning model g with observations of the predictive variables x.
- P is the set of all input features, $|P|= p$.
- S is the set of non-zero feature indices (the features that are being observed and not unknown), $|S|= s$.
- Recall $s!$ and $p!$ as the total amounts of possible *permutations without repetitions*. So,
$$\frac{s!(p-s-1)!}{p!}$$
is the total amount of permutations of size s, averaged over the total amount of permutations $p!$.
- $g_x(S)$ is the output of the model, with the subset of predictive variables $x_j : j \in S$.

Shapley values for feature j are constructed by creating the learning model on all possible subsets of features which does not include j, and then adding j and creating all those models as well. The contribution of feature j is then the average of the difference in model output between each set with and without j.

A SHAP plot depicts predictive variables in a descending order according to variable importance on separate SHAP-axes, this shows each observation's x_{ij}s contribution to model output in relation to the mean of observed outputs $\bar{y}$. The distribution of each variable is represented with a color heat scale. A SHAP plot is shown in fig. 32.

3.7 Concrete

This section summarizes theory of concrete strength.

3.7.1 Certifications

Industries in Europe are highly regulated by a system of certifications. Almost any product produced and sold on an industrial scale must achieve a set of standards determined either by the Swedish or European institute for standards. In the case of concrete piles, the most important requirements of these standards are related to the compressive strength of concrete.

Compressive strength of concrete is certified through the following quantities:

- f_c - Compressive strength of a single measurement
- f_{ck} - Characteristic strength of concrete. Denotes a certain strength class, a threshold which the strength measurements should surpass. $f_{ck} = 60$ MPa in this project, see further down.
- $\bar{f}_{ct}$ - Mean value of strength test from the last t days.
- $\hat{\sigma}_{\bar{f}_c t}$ - Estimated standard deviation of $\bar{f}_{ct}$.

The class of concrete is denoted by two numbers. The concrete in this project, "C50/60", means that the concrete are supposed to measure a compressive strength $f_c > 50$ MPa if tested on cylinders and a compressive strength $f_c > 60$ MPa if tested on cubes. [12]. However, for continuous production of concrete, the certification requirements [20] are redefined to:

$$\bar{f}_c \geq f_{ck} - 4 \ \ (\mathrm{N/mm^2}) \tag{20}$$

$$\bar{f}_{ct} \geq f_{ck} + 1,48\hat{\sigma}_{f_c t} \ \ (\mathrm{N/mm^2}) \tag{21}$$

Moreover, Two additional conditions are made for the concrete piles; $f_c >= 25$ MPa at the time of removal of piles from the formwork, $f_c > 50$

MPa at the time of delivery of piles to customer. Formwork removal and delivery to customer are currently set to one and seven days after casting.

3.7.2 Components of concrete mixtures

The development of compressive strength of concrete is affected by a many factors and the following parameters of a concrete recipe influences the quality and compressive strength of concrete and must be considered for an ML model:

- Age of concrete
- Cement: Type and amount
- Water: Amount
- Water/cement ratio
- Aggregates: Grain size, grain shapes and amount
- Fillers: Type and amount
- Admixtures: Type and amount
- Temperature in curing environment
- Humidity in curing environment

For further reference of concrete theory, see[12].

3.8 Quality

Quality is, much like *probability* a complex subject. Almost anyone has an intuition of what quality and probability is, but when studied as scientific subjects they have both required vast efforts of the academic community to define. [5] Introduces quality through 9 established definitions and continues to create their own definition:

Quality of a product is its ability to satisfy and preferably surpass the customers' requirements and expectations

It should be noted that further reasoning in [5] sees every stakeholder to the producer as a customer. So, quality is a complex subject, and seeing quality as a main attraction for buyers, it is useful to identify ML's possible role in quality improvement. The subject of quality is far more extensive than what is possible to cover in this report and is mainly been described by an introduction of *Total Quality Management* (TQM) in section 3.8.1. Moreover, section 3.8.2 describes variation in production processes as problematic.

3.8.1 Machine learning for quality

TQM is a strategy for proactively working with all aspects of quality in an enterprise. A comprehensive strategy for quality management is necessary as quality lays the foundation for a customer's long term attraction to a business. A quick summary of the keystones in TQM is to [5]:

1. Have customers in the centre of the enterprise's focus. Customers are defined as the group the enterprise aims to create value for, not only the consumers of products.

2. Work with *processes*. Processes being the activities which are repeated by the enterprise.

3. Work with continuous improvements.

4. Enable participation.

5. Make fact-based decisions.

6. Have a management committed to strengthen the above-mentioned values in the organisation.

The fifth principle, to make fact-based decisions, is in TQM taught to be achieved by using a set of 14 analytical tools. 7 of these analytical tools are based on numerical and statistical analysis. [5].

With ML regarded as advanced data analysis, it should be regarded as a development of the tools for making fact-based decisions. Even though TQM promotes the simplicity of its tools for numerical analysis as an advantage, ML's position in TQM must remain unchanged with regard to the above-stated principles. ML should be regarded as a tool which enables mainly one of the six important principles for achieving quality. At the same time, the basis for making fact-based decisions exists in symbiosis with the other principles. As work with ML demands complex computations, it requires software development. Software development is a subject found to benefit especially from the principle of working with processes. The Software Engineering Institute has produced a model for evaluation of an enterprise's process of software development, the *Capability Maturity Model* (CMM) [5]. CMM identifies five distinct levels of organisational maturity, with the two extremes: "The enterprise does not view software development as a process" and "The enterprise continuously works with improving the process of software development from all aspects of TQM". At each of these five levels, CMM provides a set of questions that are aimed at aiding the organisation to reach the next level. Here, it is only beneficial to work with the set of questions corresponding to the actual level of the company, as questions corresponding to other levels are not expected to benefit the enterprise.

3.8.2 Variation in processes

Recall processes as the activities which are repeated by an enterprise. Variation in production processes is regarded as problematic in many, if not all schools of production management, which aim to achieve improved quality. *Lean production* is a management strategy which aims to improve an enterprise by identifying and eliminating all factors in an enterprise which do not create value. Such factors are described in lean production's origin, *Toyota production system* (TPS) to be:

- *Over-straining the enterprise*
- *Inconsistent processes*
- *Unnecessary activities* or *wastes*

and are developed in [13] to a set of 14 principles to eliminate redundant factors, in which principle 4 promotes the importance of a *level workload*, and principle 6 to have *standardized processes*. In conclusion, variation in processes is undesired in a lean production system as inconsistent processes are difficult to control and may over-strain the enterprise. It should be noted that a successful implementation of lean production requires a comprehensive implementation of all principles of lean production. Many attempts to work according to lean production fails or halt due to an unbalanced focus on the principles.

4 Results

This chapter describes the findings from interviews, observations and literature studies and also from data acquisition and data analysis.

4.1 Interviews

This section contains the main results of the interviews conducted during the project.

4.1.1 First round of interviews: Requirements for an ML model

The main requirements on the ML model expressed by NCC were that it:

- Contributes to reduction of the concrete pile's climate footprint, as measured in CO_2.
- Enables financial analysis, so that costs can be reduced.
- Is easy to use and implement, and can be implemented in Microsoft Azure.
- Can act as a showcase that demonstrates the usefulness of data and ML for both NCC and Hercules.

4.1.2 Later interviews with stakeholders

The area within concrete pile production where machine learning is the most feasible is in the relation between concrete-production processes and the quality of concrete. This area produces most of the data and includes the least uncontrolled variables. Other possible areas are failure frequency of piles during installation and performance of the steel parts of the piles.

The constraints created by standards and processes described in section 3.7.1 are not completely fixed. f_{ck} as the characteristic strength is fixed, but the time tresholds may be possible to vary. If sufficient compressive strength requires more time than 1, 7 or 28 days when removing piles from forms after 1 day, delivering to customers after 7 days or even after the main certification test after 28 days, it can be of interest to allow for more time in order to obtain a lower climate footprint.

NCC means to adopt and implement lean production. The implementation may not have come as far in Hercules, but they mean to adopt lean production as well.

A possible reason for variability in compressive strength may be varying quality of cement, which supposedly decrease during summer.

The plant in Västerås was able to reduce their cement use by 12 % by switching the source of aggregates, i.e. gravel mine.

4.1.3 Interviews in Hercules plant in Ucklum

One of the biggest problem in the plant is related to the logistics. Customers of Hercules occasionally ask for deliveries within the next day, whereas the concrete needs to cure for seven days. This phenomenon is most common during the summer, when demand is high in general, but occurs occasionally throughout the year. At the same time, the permitted size of the stockpile is limited since storing concrete piles in heaps higher than 2 meters is considered a safety hazard. Hercules also produces many variations of pile dimensions, so maximizing stockpile costs a substantial amount through *tied up capital*. Tied up capital is expenses of a company that are yet to be realized as revenue. Tied up capital is regarded as expenses through the company's *internal rate* and the time the capital is tied up. This problem is emphasized by the plant manager and also pointed out by a worker in the plant. Two workers were, independently of the manager, asked: "Where would you consider an ML model to contribute the most in the production?". To which they replied "the logistics situation" and "I do not know".

A possible reason for decreased compressive strength during the summer was proposed during interviews with the workers. The summer heat increases the amount of filler in the aggregate. Fillers are small size particles, here rock dust but may refer to other small size particles as well. This causes several problems. Fillers reduce the compressive strength of the concrete according to the experience of workers, interviews with technical specialists at NCC and literature studies[12]. Fillers also increases the viscosity of the concrete mixture, tricking workers to believe that more water is needed. When high viscosity is caused by fillers, it is, however, not helpful to add more water to the mixture. Instead, some addmixtures are needed.

It was also stated that no data exists on faulty concrete mixtures, as such mixtures do not occur.

With regard to the reinforcement data, it was said that there is data showing a problem with differing strength measurements, but the cause of this problem is hard to persuade.

4.2 Observations

This section accounts for observations made in the production plant.

4.2.1 Mapping of data collection

A visualization of where various data sets are collected in Hercules' processes is presented in fig. 5, which is an expended version of fig. 39

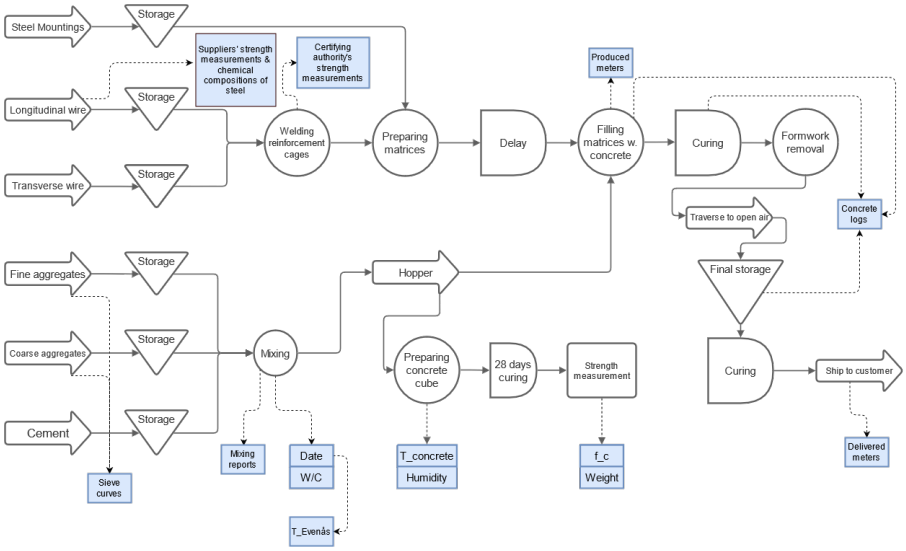


Figure 5: Overview of where in Hercules' processes the data sets found during this project are collected. Data sets are marked by blue sharp-edged rectangles. See section 4.4 for further description of the data.

4.2.2 Reliability of concrete cube measurements

The creation and measurement processes of concrete cubes was inspected and the following was found:

The size of the cube sides are not supposed to vary more than 150 $\pm 0,75$ mm and are continuously reported as 150 mm. However, produced cubes were inspected until one faulty cube was found, and a cube of 146×150 mm was found at the fourth draw, see fig. 6. This variation is likely due to two reasons. Firstly, operators follow different procedures in packing concrete into the cube form. Some pack the form in three layers, some fill the form in one movement. Secondly, the form gets vibrated after being filled up, the precision in the vibration machine is very rough. Overvibrating is therefore likely, which leads to shrunken volume and possibly also segregation in the concrete cube. This affects the reliability of compressive strength measurements as these are carried out through testing the load at which a cube cracks. The load is then divided by a constant area of 150×150 mm but this area is actually a variable.

Figure 6: The sizes of the concrete cubes vary. This affects the compressive strength measurements which depend on the surface area.

4.3 Literature studies

4.3.1 Other works with ML for concrete strength

This section summarizes the findings of literature studies regarding predicting of compressive strength of concrete using machine learning. The main findings are that tree-based methods and neural networks are commonly used and that large data sets are scarce.

Yeh's data

Three papers works with a data set on compressive strength provided by Yeh[24], (for further data description, see section 4.4.4):

Feng et al. [9] evaluates AdaBoost on Yeh's data set and presents performance measures after cross-validation: $R^2 = 0.982$, RMSE (MPa) = 2.20, MAPE = 6.78 %, MAE (MPa) = 1.64. This performance is from CART based AdaBoost and is reported to be superior to *Artificial Neural Networks (ANN), Support Vector Machines (SVM), Classification and Regression Trees (CART), Linear Regression (LR)*, stacking with CART+SVM+LR and Gradient boosted ANN.

Farooq et al. [8] compares other methods to Feng et al.'s work with AdaBoost and finds *Random Forest* (RF) and *Decision Trees with bagging* to be the best methods.

Nguyen et al. [16] find XGBoost and *Gradient Boosted Regressor* (GBR) to be appropriate for Yeh's data set. Both methods reports $R^2 = 0.97$ after cross-validation. These methods are superior to SVM and *Multilayer Perceptron* (MLP), according to [16].

Other findings

Sevim et al. [19] reports that *Generic Algorithm-Adaptive Neuro Fuzzy Inference System* (GA-ANFIS) is more appropriate than ANN's and Multiple Linear Regression (MLR) to predict compressive strength of concrete based on fly ash content. This investigation is mainly focused on investigating the potential for substituting cement with fly ash. This is based on their own data where fly ash substitutes cement with 10-40%. $R^2 = 0.946$ for GA-ANFIS on the test data set.

Ouyang et al. [18] evaluate the learning curves of polynomial regression (PR), ANN and RF and find all three methods to be useful but points especially to the trade-off between learning quickly, with few data points, and achieving high accuracy.

Nguyen et al. [17] reports that the ANN's Residual Neural Network (Resnet) and Deep Neural Network (DNN) are appropriate for predicting compressive strength of green fly ash based geopolymers, which is not concrete but similar to concrete. They gather their own data base through experiments and use the chemical composition of the fly ash as covariates.

Findings without predictive performance measure

Reference [3] contains another data set with 2400 observations of 16 features out of which 12 are relevant for this project.

Chopra [7] finds the the optimal polynomial model to be a multivariate power equation, based on [14]:

$$f_{c28} = A_0 \left(\frac{w}{cm}\right)^{A_1} \left(\frac{FA}{cm}\right)^{A_2} \left(\frac{CA_1}{cm}\right)^{A_3} \left(\frac{CA_2}{cm}\right)^{A_4} \tag{22}$$

This yields $R^2 = 0.9905$ and $MSE = 0.0236$ on concrete with medium workability and unknown size of the data. However, it is not mentioned whether performance is measured on a training or test set. Parameters are water (w), cement content (cm), fine aggregates (FA) size < 10 mm, coarse aggregates > 20 (CA_1) and coarse aggregates > 10 (CA_2). Measurements are distributed over 15 cubes from each of 30 different mixtures.

Ziolkowski and Niedostatkiewicz [25] evaluate a 4-5-1 ANN with the input variables: weight of {cement, water, fine aggregates, coarse aggregates}. No general performance measure of the ANN is presented, but the

authors claim ANN to be a reliable method. The utilized data set contains 741 measurements.

4.3.2 "Off-the-shelf" ML algorithms

This section summarizes the findings of literature studies regarding appropriate ML algorithms from a general perspective.

ML problems usually contains some characteristic challenges.

Firstly, data for ML problems are generally messy; containing both quantitative and qualitative data. Distributions of variables are often long-tailed, skewed and data points might be missing. Secondly, it is often unknown which variables are relevant for predictions of the outcome. Thirdly, some ML algorithms performs well, but are hard to interpret. Many times it is desirable to understand the relationship between input and output variables. Finally, it is desirable for the algorithm to be suited to the amount of data. When the data is expensive, a quick learning speed can reduce the cost of data collection. With large amounts of data, complex algorithms may require an unfeasible amount of computations.

Decision trees have been compared [11] to *neural networks, support vector machines (SVM), multivariate adaptive regression splines (MARS), k-nearest neighbours (kNN)* and performs well with regards to:

- Mixed data types
- Missing values
- Outliers in input space
- Monotone transformations of data
- Large data sets
- Irrelevant input variables

None of the other algorithms, Neural networks, SVM, MARS, kNN, performs well in all the areas listed above. However, decision trees fall short in regard to:

- Ability to extract linear combination of features
- Interpretability of model
- Predicitve ability

However, boosted trees often have a dramatically better predictive ability than ordinary trees. This comes however with the sacrifice of learning speed and interpretability. As boosted trees generally require more calculations which requires more time to train the model. Hence, boosted trees are deemed the best "off-the-shelf" ML algorithm.

4.3.3 Programming libraries

"*CatBoost* is a high-performance open source library for gradient boosting on decision trees" built by Yandex, initially released in 2017 and fully released in 2020. The CatBoost-website reports superior performance measures compared to its competitors *LightGBM, XGBoost* and *H2O* on a set of data sets from old ML-competitions. Such a representation is of course biased, but indicates none the less that CatBoost is a well-functioning library. CatBoost also claims to be easy to use, as desired according to interviews. CatBoost is supposed to be implementable without considering usual problems in ML, specifically CatBoost claims [23] to:

- Not demand parameter tuning, although parameter tuning can lead to small improvements in performance.
- Handle both categorical and numerical data automatically.
- Deal with overfitting automatically.
- Have a fast learning speed, requiring little time to render gradient boosted trees.

Other packages used in this project were:

- The SHAP-package, to generate SHAP plots. (Python)
- The gbm-package, to use gradient boosted trees. (R)

4.4 Data

This section describes the data sets obtained during the project. Each data set is presented by descriptive statistics, its relation to the production system of concrete piles, its format and the process which generated the data.

4.4.1 First round of data: Compressive strength, W/C and Dates

The primary data set, which initiated this project, contains information on the compressive strength, f_c, of the concrete, the water-cement ratio, W/C, and the dates, $Date$, of casting the concrete. For the year of 2020, this comprised a 214×3 matrix.

The main parameter of interest is f_c by which the quality of the concrete piles is determined and certified. A cube of 15×15 cm is taken, stored in an environment of controlled temperature and humidity, and thereafter its ultimate loading capacity is tested after 28 days. A histogram of f_c is shown in 7 [12].

W/C is supposed to be the most influential parameter on f_c, and therefore it must be controlled and measured daily in order to certify the quality of the piles [12]. The W/C is measured in the concrete mixer through a moisture metre with precision of two decimals.

Date in itself is not meant to be a parameter with direct influence on compressive strength. Yet, dates are essential information in case customers would report insufficient strength in the piles. *Date* could also correlate with other influential parameters, such as temperature, humidity, or others.

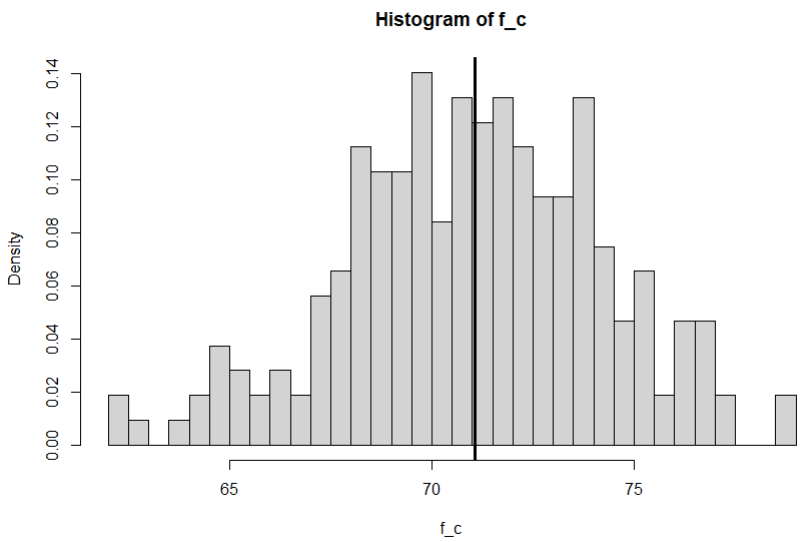


Figure 7: Histogram of compressive strength, f_c, measurements of 2020. The black line marks the mean value.

4.4.2 Second round of data: Temperature, Weight and Humidity

Some additional data related to the concrete cubes was found to be collected. This data was also required for the quality certification of the concrete piles. The parameters of this data was: Concrete temperature at the moment of pouring it into the mold, $T_{concrete}$, weight of the cubes, $Weight$, relative humidity in the lab at day 0 for the cube, $Humidity$. In addition, temperature data for Evenås, a village some kilometers away from Hercules Ucklum, was available at [2]. Hence, a final feature, $T_{Evenås}$, was included in the data, resulting in a total of 6 predictors to work with. Furthermore, data

on f_c, W/C, $Date$ was obtained for the period 2015-2020 and for the additional parameters, only for 2020. This rendered matrices of sizes 1225×3 and 214×7.

$T_{concrete}$ is of interest as temperature and temperature development in the concrete influences f_c in several ways. See [12] for further reference and note that $T_{concrete}$ only provides a point measure and says little about temperature development during the time frame during which temperature affects the strength development. $T_{Evenås}$ was obtained as an attempt to obtain more information on the temperature processes in the concrete.

$Weight$ could hold information on density of the concrete, and primarily indicates air content, which is expected to influence f_c [12]. The relation between density of the concrete would also be affected by the ratios of concrete ingredients in each specific concrete mixture, complicating the usefulness of $Weight$ as a predictor for air content. $Humidity$ is supposed to have a negative relation to f_c [12]. The full data from 2020 is visualized in fig. 10 and 11.

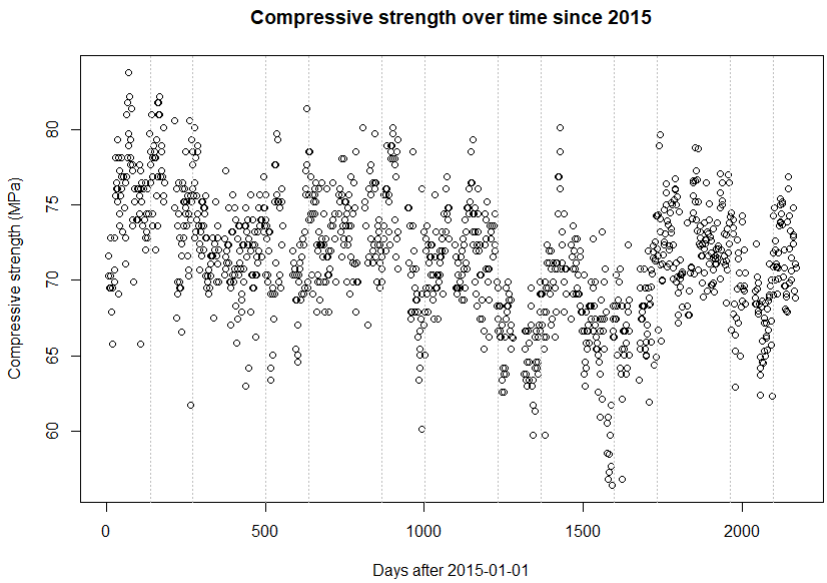


Figure 8: Compressive strength over time for the period 2015 to 2020. Recall that f_c was seen to drop during the summer in 2020 as in fig. 22, dashed lines marks the same summer period in previous years. The drop during summer is reoccurring.

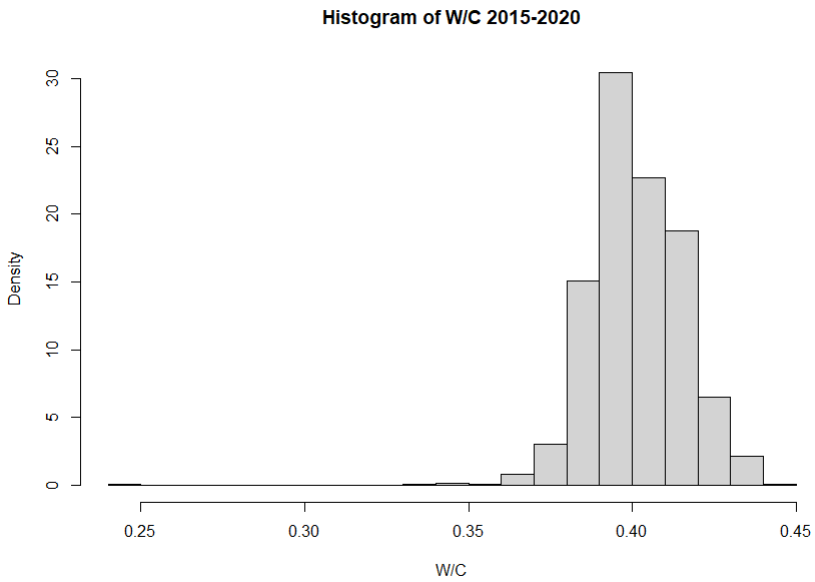


Figure 9: Histogram of W/C values from 2015-2020. Again, a large majority are distributed over a short interval. One notable outlier is present at 0.25.

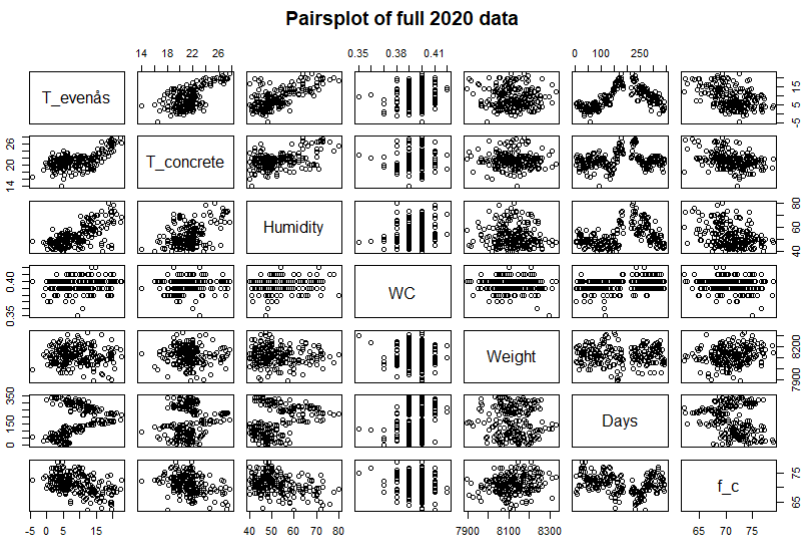


Figure 10: Pairsplot of the full 2020 data. An enlarged version of this figure is found in appendix 7.1.

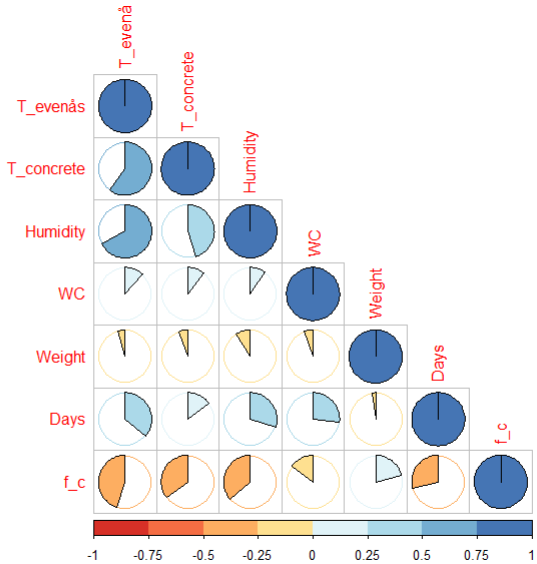


Figure 11: Correlation plot, a vizualization of the correlation matrix of the of the full 2020 data. Red scale colors represent a negative correlation and blue scale colors represent a positive correlation. Pie sizes shows the absolute value of the Pearson correlation coefficient.

4.4.3 Compressive strength - unaccessible data

This section describes the types of data that was found to be collected with consistency, but hard to utilize due to issues with the way it is collected.

Sieve curves are reported for each batch of aggregates. The proportions of different sizes of aggregates influence the compressive strength of concrete. However, the amount of information in these curves is hard to evaluate further than with visual inspection of the curves, as only 20 curves were obtained for 2020. Furthermore, no reasonable way to relate curves to concrete cubes was found except to visually inspect if curves from the summer differ from the rest of the year. The curves are also reported in pdf files, which complicates the process of obtaining numerical values.

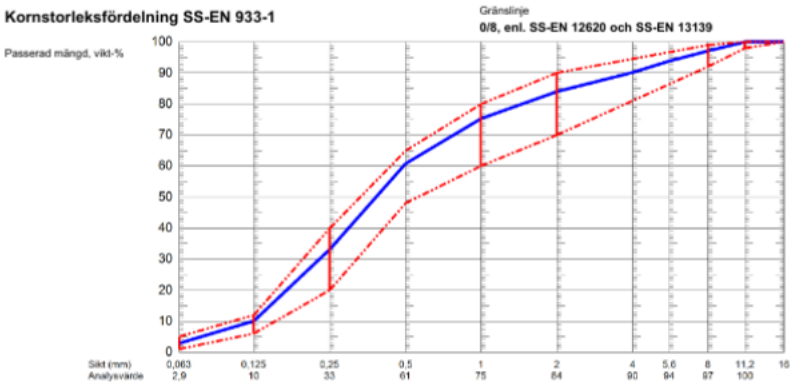


Figure 12: An example of a sieve curve. Although the type of information in the curve is expected to be informative, the data is hard to use.

Mixing reports are created for each batch of mixed concrete, containing information on:

- Amount of total concrete (kg and m^3)
- W/C with 3 decimals
- W/P with 3 decimals, the ratio between water W and small particles P
- Amount of fine aggregates, size 0-8 mm (kg)
- Amount of coarse aggregates, size 8-16 mm (kg)
- Amount of large aggregates, size 30-40 mm (kg)
- Amount of cement (kg)
- Amount of water (kg)
- Time stamps for the actions that the mixer carries out (hh:mm:ss)
- A batch id
- General information about what concrete is being mixed
- A curve for the ampere of the mixer while mixing the concrete

The mixing reports are accessible as pdf files, but the software which generates them does not allow to export multiple files at a time.

Blanderapport

Blander	1	Recept	17-25 Viscocrete	Ordredato	2020-10-14 11:30:19
Chargenr.	63008	Miljøklasse	Passiv	Blandetid	175 sek.
Udleveringssted	11	Eksponeringsklasse	XF1	Doser	11:32:26 (00:02:07)
		Vand ±	-3,0%	Blander	11:36:01 (00:05:42)
		Konsistens	0,00	Klar	11:38:31 (00:08:12)
		Flydespænding	0,00	Tøm	11:39:27 (00:09:08)
		Luftindhold	2,00%	Slut	11:39:37 (00:09:18)

	kg	m³	V/C	V/P
Bør	3623	1,493	0,384	0,384
Er	3622	1,508	0,397	0,397

Materiale		Silo	Density	Bør kg	Er kg	Afv%	Vand kg	Vand%	kg/m³	Type	Fyldt i blander
2	8-16 sten	3	2690	1435,470	1436,000	0,04	8,522	0,60	952,3	Auto	11:36:01 (00:05:42)
1	0-8 sand	1	2690	1345,390	1346,000	0,05	62,488	4,88	892,6	Auto	11:36:01 (00:05:42)
31	Komposit	31	3100	652,706	651,000	-0,26	0,000	0,00	431,7	Auto	11:36:06 (00:05:47)
21	Kall vatt en	21	1000	173,682	173,662	-0,01	173,662	100,00	115,2	Quick (1)	11:37:01 (00:06:42)
48	30/40CA	43	1090	4,895	4,890	-0,11	2,934	60,00	3,2		11:37:16 (00:06:57)
21	Kall vatt en	21	1000	(10,347)	10,347	0,00	10,347	100,00	6,9	M-Vand	
Materiale/spulevand				0,255	0,255		0,255	100,00	0,2		
	Total			3622,745	3622,154		258,208		2402,0		

Quick (1) Recept kg: 169,1 Tør: 6,7 Fugtmåling skalafaktor: 17,010 Bør: 12,4 Er: 7,0

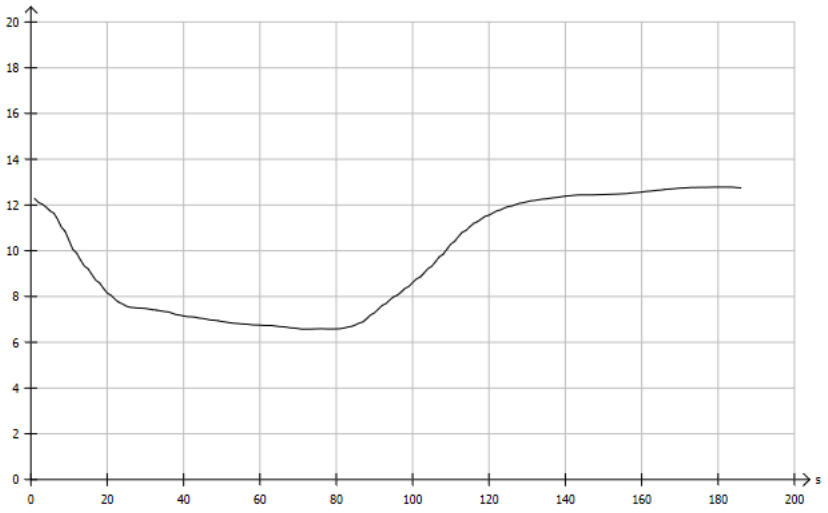
Alarmer

Opstået	Kvitteret	Skip/Genstart	
11:38:42	11:38:57 (00:00:15)	Manuel genstart 11:38:57 (00:00:15)	BLANDARÖPPN. 11, Öppning/Stängning för långsam, BLANDARE 1, Väg ej öppnad

2021-03-16 14:57:27 Chargenr. 63008 Side 1

Figure 13: An example of page 1 of a mixing report.

Blanderapport



2021-03-16 14:57:27 Chargenr. 63008 Side 2

Figure 14: An example of page 2 of a mixing report.

Concrete logs During the first 7 days, core and ambient temperature are measured in the concrete piles. This data is collected by a software named

Metodia and written in .mxl files where these temperature curves acts as a basis for estimating the compressive strength of the concrete. This way of working is allowed by certifying authorities. Although, the .mxl files look a lot like excel files, it was not possible to extract their data except for one file at a time. As this was a time consuming process, only 12 data sets from August and from November-December was obtained. This data is plotted in fig. 15 and 16.

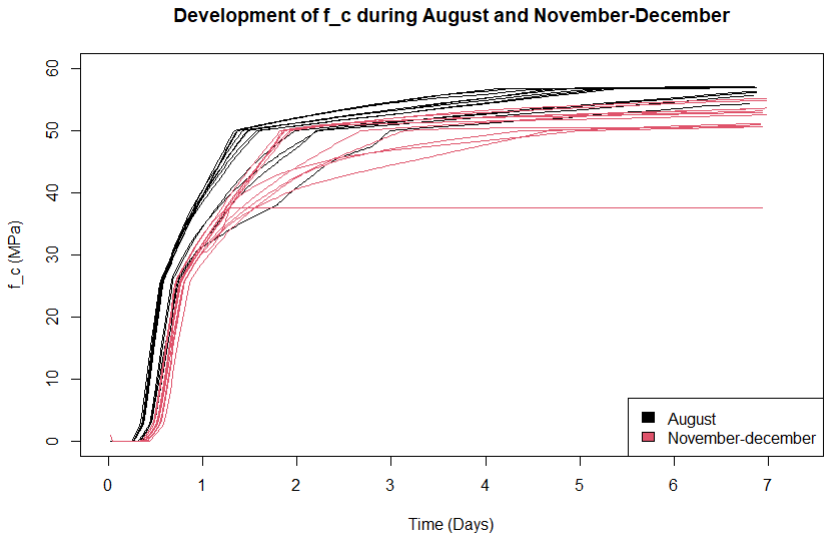


Figure 15: 24 samples of development of compressive strength of piles over the first 7 days as estimated based on temperature development, by the metodia software. Black show curves from August, Red from November-December.

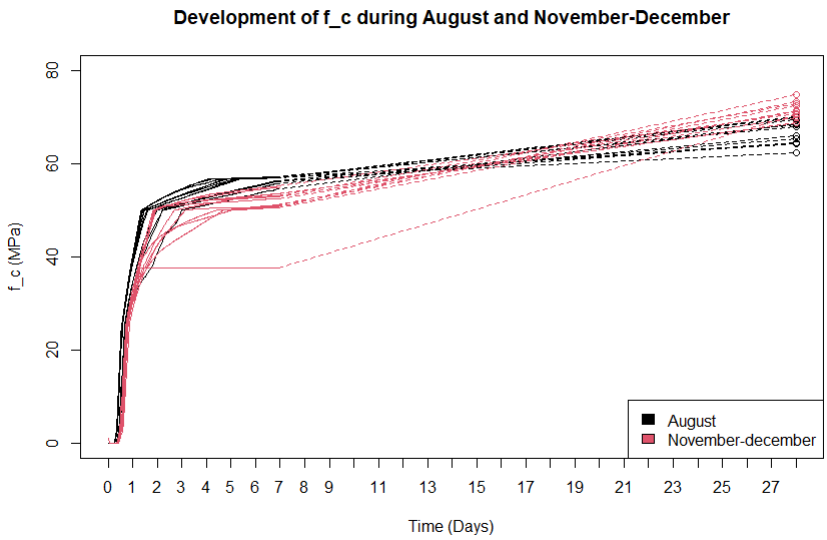


Figure 16: The same 24 curves as in fig 15. Dotted lines connect each sample to its related cube's compressive strength.

4.4.4 I-Cheng Yeh's data

This data set was used to evaluate the appropriateness of the CatBoost library for a general ML application whose primary aim is to enable CO_2 reduction of a concrete recipe. The data does not relate to Hercules' production and analysis of this data is an unsuitable basis for changing Hercules' concrete composition. The data contains 1030 measurements on compressive strength and the 8 quantitative input parameters:

- Cement (kg in a m3 mixture)
- Blast Furnace Slag (kg in a m3 mixture)
- Fly Ash (kg in a m3 mixture)
- Water (kg in a m3 mixture)
- Superplasticizer (kg in a m3 mixture)
- Coarse Aggregae (kg in a m3 mixture)
- Fine Aggregate (kg in a m3 mixture)
- Age (Days)

Note that this data is from Taiwan, 1998. A pairsplot and correlation plot is shown in fig. 18

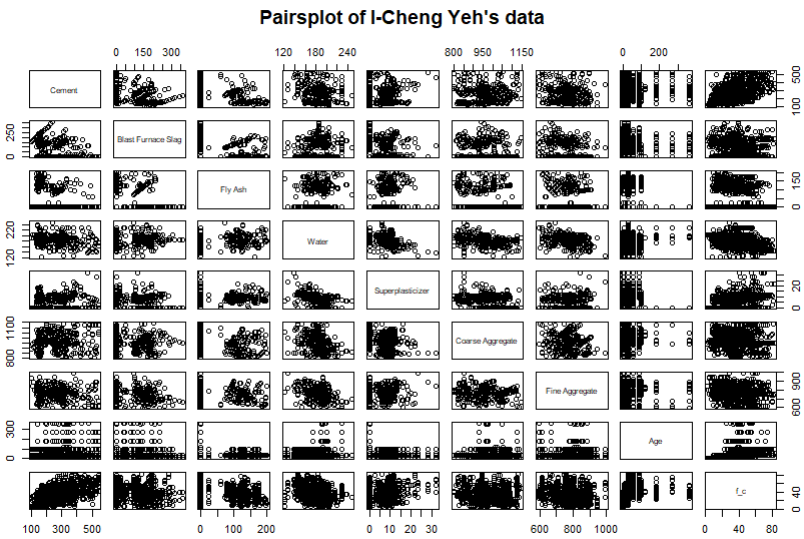


Figure 17: Pairsplot of I-Cheng Yeh's data. f_c is represented at the bottom right. An enlarged version of this figure is found in appendix 7.1.

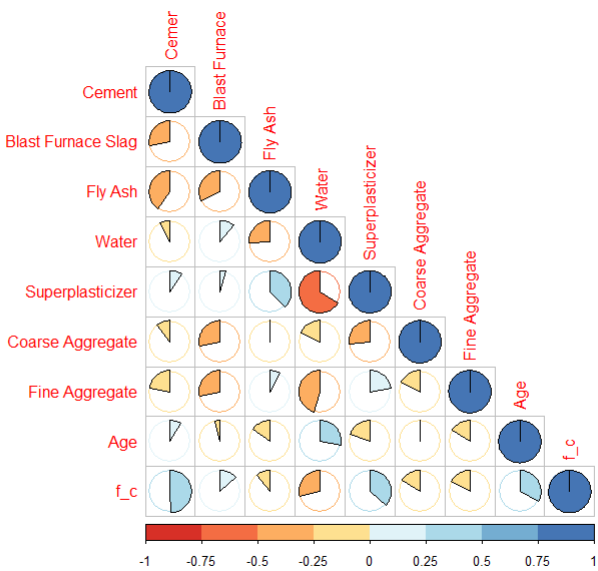


Figure 18: Correlation plot of I-Cheng Yeh's data.

4.4.5 Reinforcement data

Two types of data exists which relate to steel reinforcement, one accessible and one unaccessible. The accessible data describes measurements of yield strength (Re) and ultimate tensile strength (Rm) of the reinforcement steel.

The suppliers measurements and the measurements ordered by Hercules to be carried out by a certifying authority in Sweden vary more than what should be expected when accounting for strain hardening, see fig. 19.

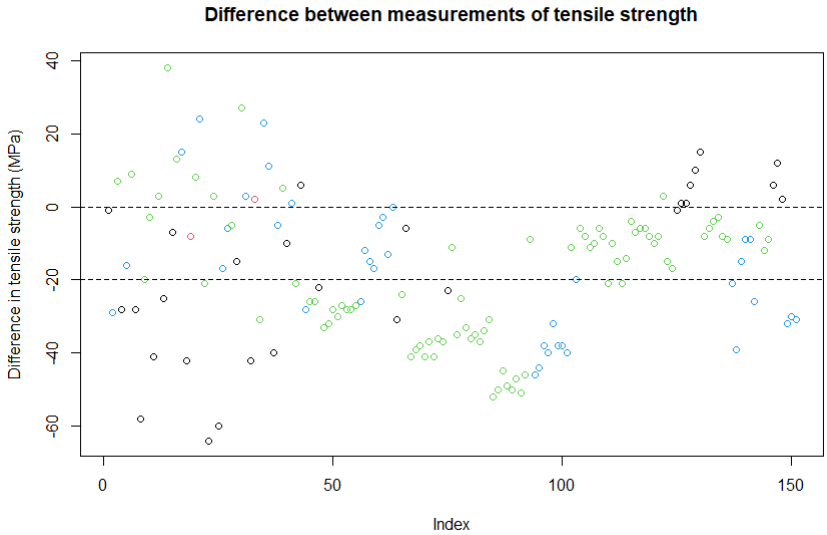


Figure 19: Difference between the supplier's and Hercules' measurements of tensile strength of Reinforcement steel. Colors represent different dimensions of the steel. Due to the processes at Hercules, the tensile strength can be expected to drop with up to 20 MPa. The dotted lines show this region.

The unaccessible data consists of hundreds of pdf files containing information on the chemical composition of a batch of reinforcement steel. One line of such data is presented in table 1.

C	Mn	Si	P	S	N	C_{eq}	Re(MPa)	Rm(MPa)	...
0,17	1,10	0,19	0,022	0,028	0,009	0,41	552	639	...

Table 1: An example data point of the chemical compositions that are available in pdf files for each batch of reinforcement steel.

4.4.6 Production data

The obtained data relating to production was from 2016-2020 and comprised monthly information on:

- Worked hours (h)
- Sick leave (%)
- Delivered meters (m)

- Produced meters (m)
- Billing (sek)
- Delivered meters within Hercules' organisation (m)
- Billing within Hercules' organisation (sek)
- Delivered meters to external customers (m)

4.5 Data analysis

This section describes the findings from analysis of the data sets described in section 4.4.

4.5.1 First round of data

The relation between f_c and W/C was examined with a simple linear regression, see fig. 20. $R^2 = 0.022$ and visual inspection of the plotted data indicates that this relation is not useful in itself. The low variation of W/C data from 2020 is concluded to be the cause of the low R^2 value, see fig. 21. 80% of the W/C data is distributed in $\{0.39, 0.40\}$.

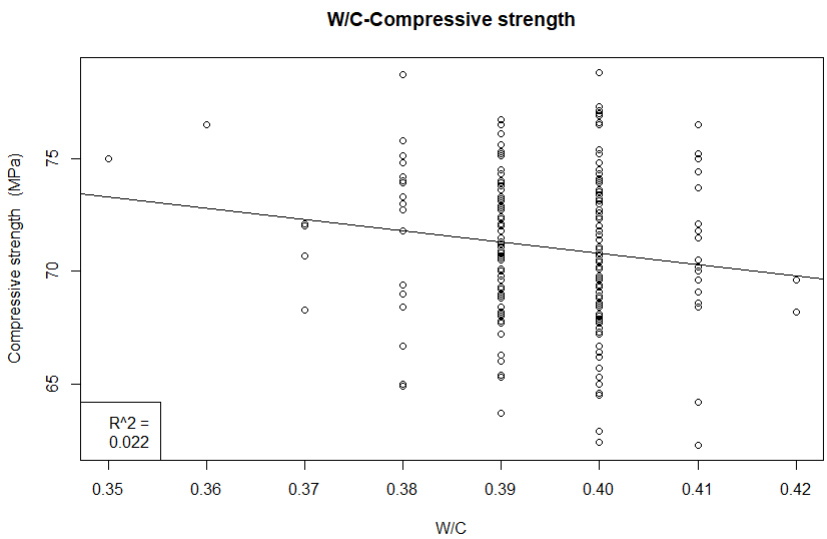


Figure 20: Linear regression between f_c and W/C.

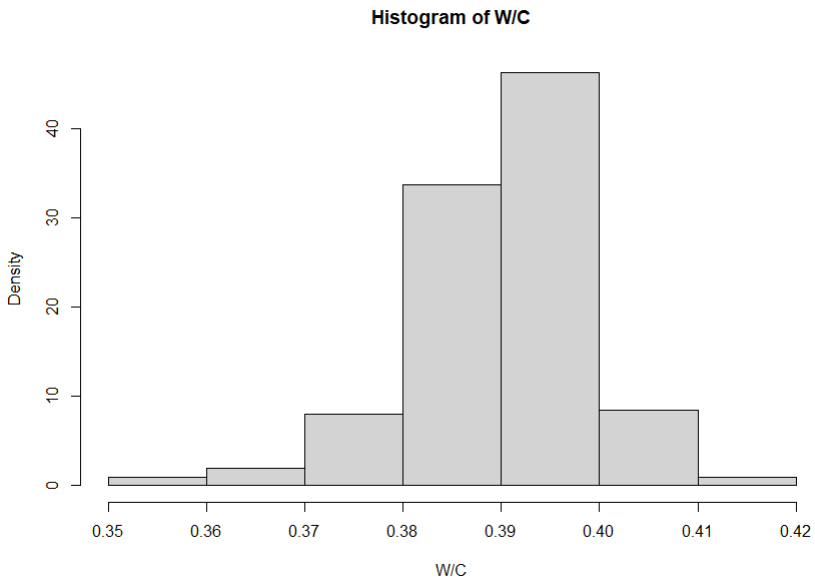


Figure 21: Histogram of W/C values from 2020, the supposedly most influential parameter on f_c. Note that 80% of the data points are distributed on 0.39 and 0.40.

Compressive strength over time during 2020 is presented in fig 22. There is a noteworthy drop in compressive strength during the summer.

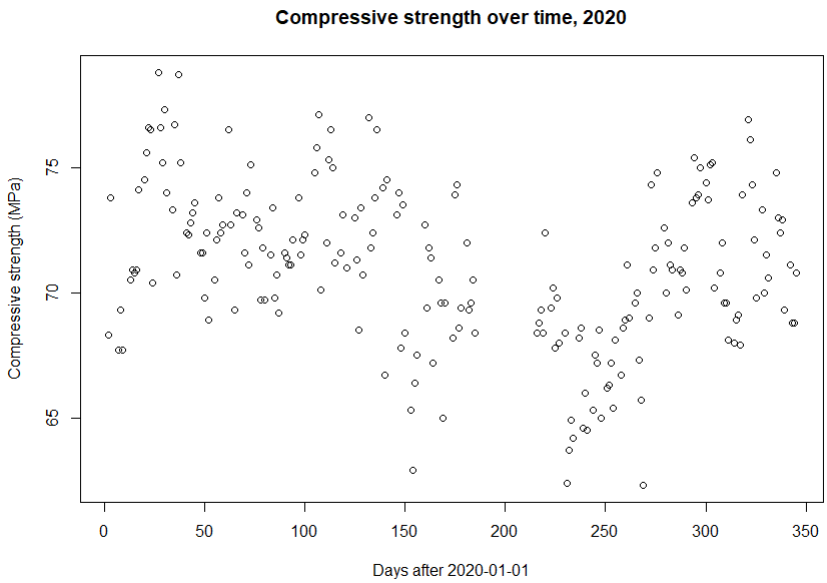


Figure 22: Compressive strength over time for 2020. Note the drop during the summer.

This relation was examined with three approaches: Fitting a sinusoidal model to the data, evaluating the data with mean values for two periods; when water for concrete mixtures is heated or not heated, evaluating the same two periods with a quadratic model when the water is not heated and the mean when water is heated. The R^2-values of 0.27-0.38 could suggest a useful relation, but the residuals do not seem evenly distributed for any of the models, see Figures 23, 24 and 25.

Figure 23: Sinusoidal model fitted to compressive strength over time for 2020.

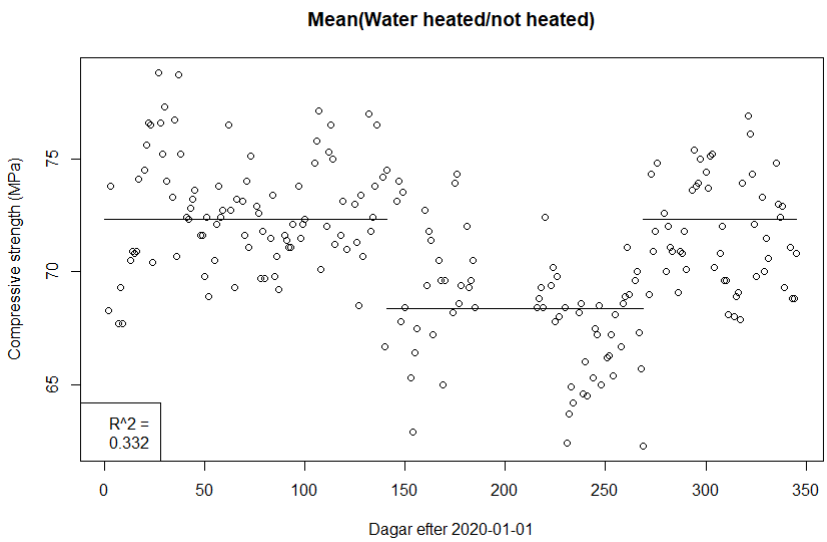


Figure 24: Model of means of two periods fitted to compressive strength over time for 2020.

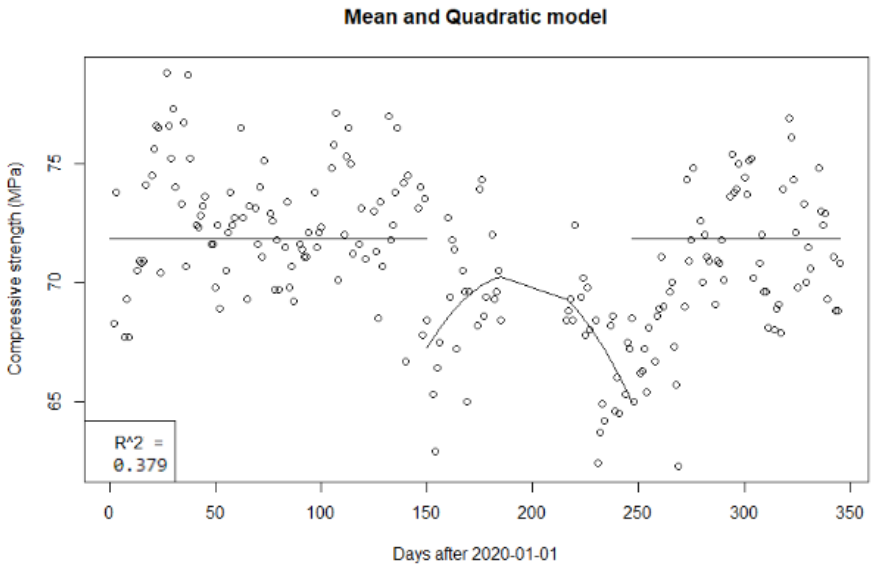


Figure 25: Model as mean of compressive strength when water is heated and a quadratic model when water is not heated. Fitted to data from 2020.

4.5.2 Second round of data

Simple linear regression models were fitted to the variables $T_{Evenås}$, $T_{concrete}$, *Weight, Humidity, W/C*, see figures 26 27, 28 and 29. The models were concluded by visual inspection to be fit enough to disregard the usefulness of the variables based on their low R^2-values.

A multilinear model was fitted on all available parameters, and gave $R^2 = 0.21$.

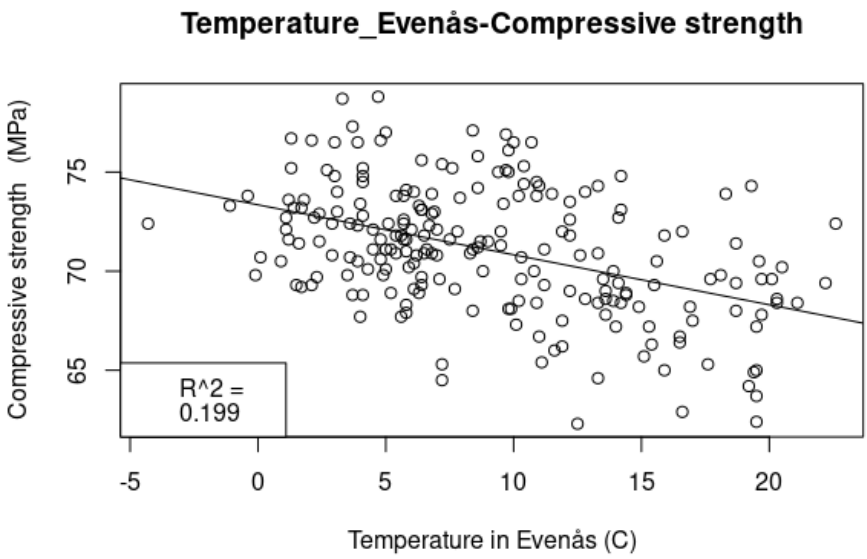


Figure 26

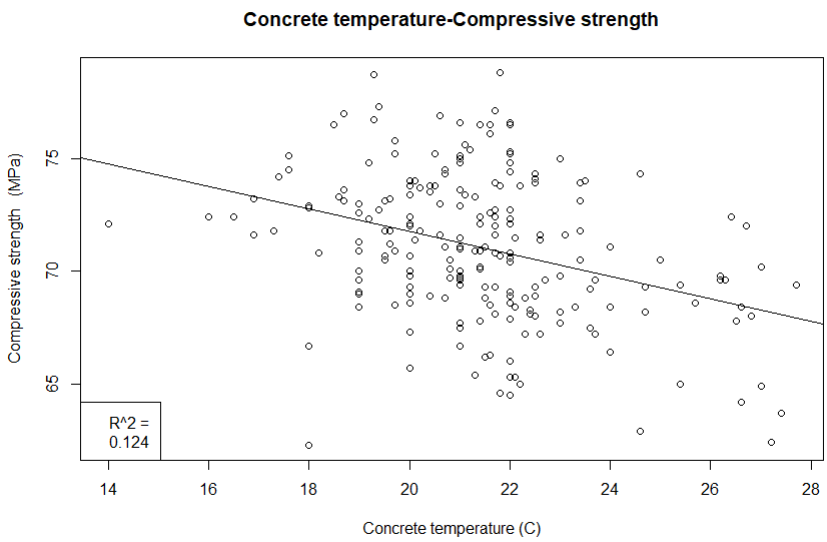


Figure 27

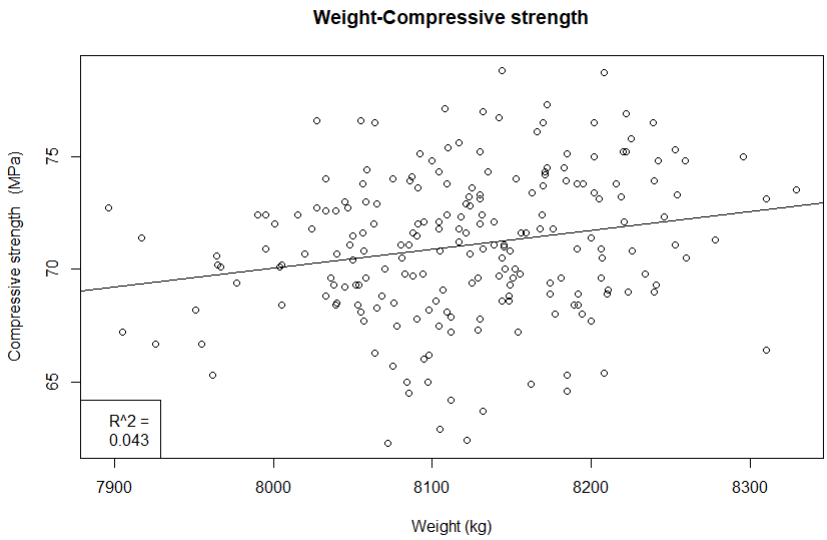


Figure 28

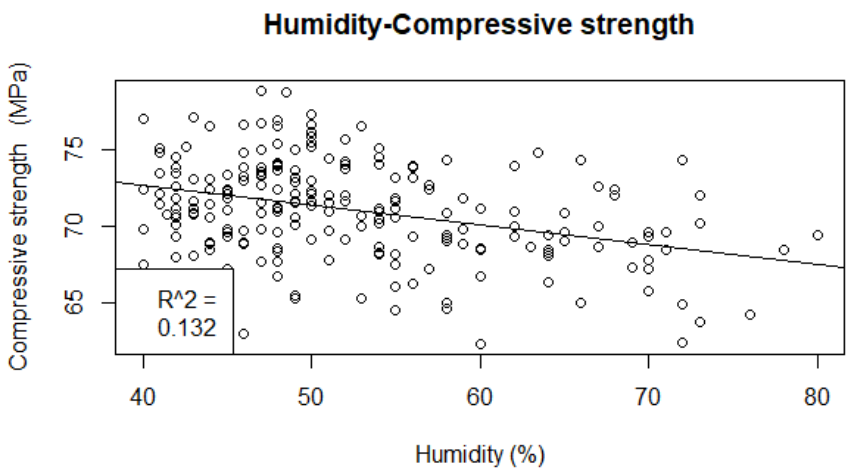


Figure 29

To get a final sense of the predictive ability of the full 2020 data set, a gradient boosted tree was fitted to the complete data set with f_c as output and $T_{Evenås}, T_{concrete}, Weight, Humidity, W/C, Days$ as input variables. Table 2 shows R^2-values and cross-validated R^2_{cv}-values over 4 folds, and for shrinkage $\nu = 0.1, 0.05, 0.01, 0.001$. Fig. 30 shows the relative influence of the input variables for the model with best predictive performance, which is when $\nu = 0.05$. This model was fitted with the gbm-package in R.

ν	R^2	R^2_{cv}
0.1	0.65	0.33
0.05	0.60	**0.35**
0.01	0.56	0.31
0.001	0.58	0.31

Table 2: R^2-values and cross-validated R^2_{cv}-values for a gradient boosted tree over 4 folds and for shrinkage values $\nu = (0.1, 0.05, 0.01, 0.001)$.

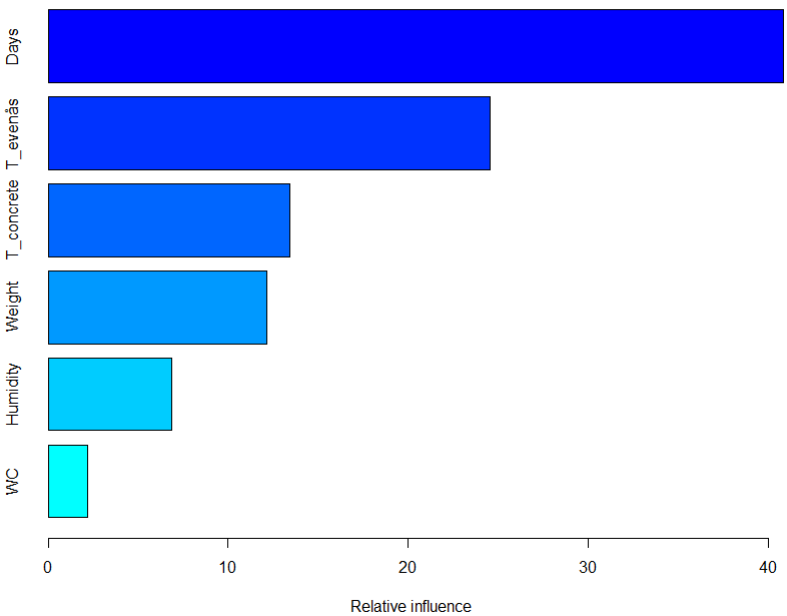


Figure 30: Relative variable importance of input parameters on a gradient boosted tree trained on the full 2020 data set with shrinkage tuned to $\nu = 0.05$.

4.5.3 Compressive strength - unaccessible data

No further analysis than what is presented in section 4.4.3 has been made.

4.5.4 I-Cheng Yeh's data

Table 3 compares the results of an untuned CatBoost model to the results of [16] where XGBoost is tuned to predict compressive strength of concrete based on the same data set.

Boosting Library	R^2	R^2 basis
CatBoost$_{untuned}$	0.93	test set
XGBoost$_{tuned}$	0.97	cross-validation

Table 3: Comparison of an untuned CatBoost machine with a tuned XGBoost [16] machine.

Variable importance and SHAP plots are shown in figures 31 and 32. These plots are used to interpret the model, they visualize how much the different input parameters contribute to the created ML model.

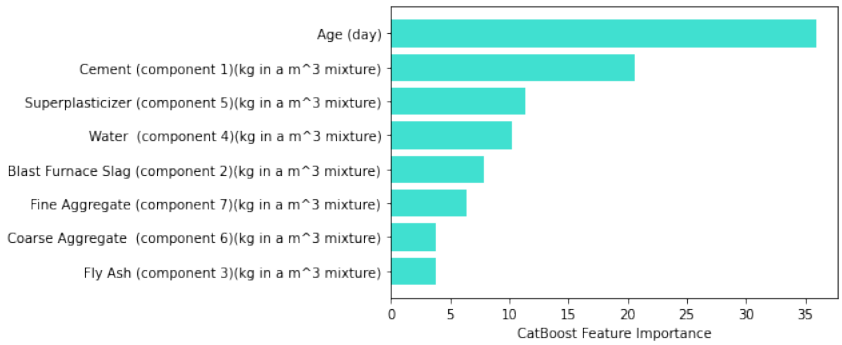


Figure 31: Feature importance of a CatBoost model created with the application.

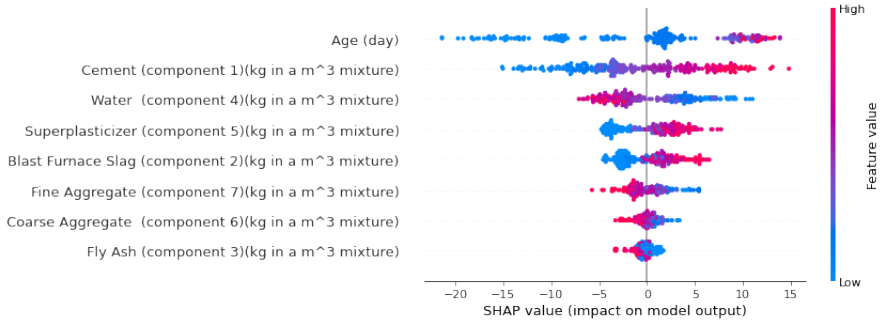


Figure 32: SHAP values and distributions of features of a CatBoost model created with the application. The data point for which *Age* contributes most is found at the top right. The *Age*-value of this data point contributes with a prediction which is 15 MPa higher than the mean of compressive strength. The *Age*-value of this data point is close to the mean as its color is purple.

4.5.5 Reinforcement data

Differences in strength measurements are shown in figures 33-38. Some data points origin from the same batch of steel. The subsets of data

for each dimension correspond to 20 unique batches for dimension 12b, 14 unique batches for dimension 16b and 30 unique batches for dimension 16c.

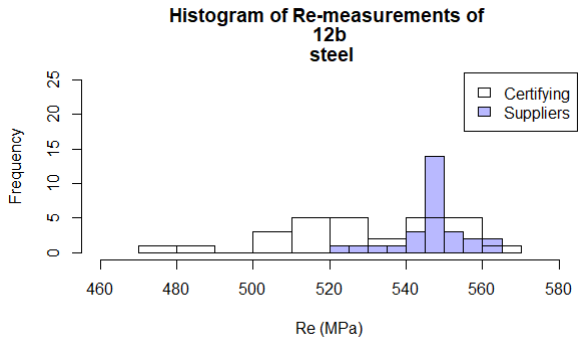


Figure 33: Histogram of suppliers' measurements (blue) and certifying measurements (white) of yield strength for steel with dimension 12b.

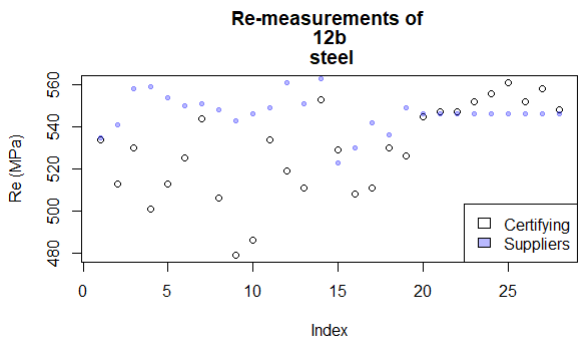


Figure 34: Plot of suppliers' measurements (blue) and certifying measurements (white) of yield strength for steel with dimension 12b.

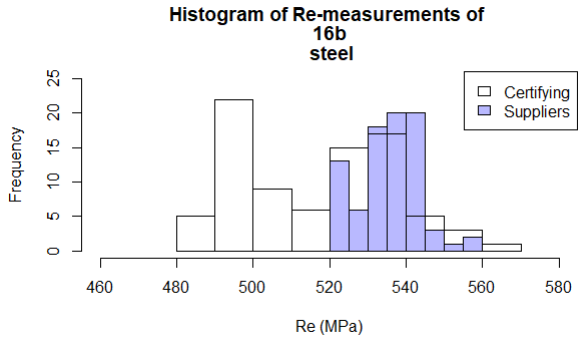


Figure 35: Histogram of suppliers' measurements (blue) and certifying measurements (white) of yield strength for steel with dimension 16b.

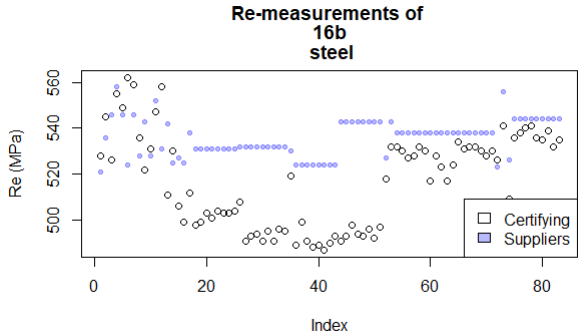


Figure 36: Plot of suppliers' measurements (blue) and certifying measurements (white) of yield strength for steel with dimension 16b.

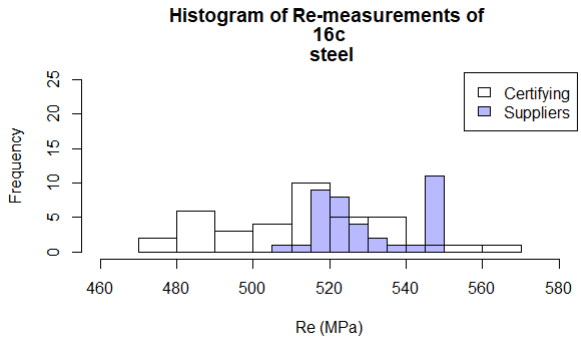


Figure 37: Histogram of suppliers' measurements (blue) and certifying measurements (white) of yield strength for steel with dimension 16c.

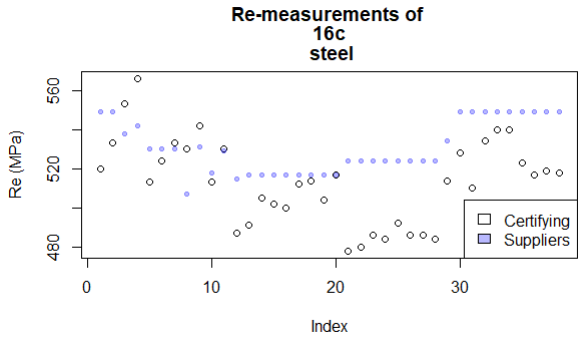


Figure 38: Plot of suppliers' measurements (blue) and certifying measurements (white) of yield strength for steel with dimension 16c.

Unaccessible data of reinforcement steel is yet to be analyzed. It would be interesting to examine the relation between chemical compounds and strength measures with the gradient boosted trees algorithm.

4.5.6 Production data

The production data reveals that pile meters per month, for delivery, production and storage, during the years 2016-2020, vary between 7 600-34 000, 10 600-27 800 and 8200-28 600. Here, storage values are based on the assumption that starting storage as of January 2016 is 20 000 pile meters. Fig. 39 shows these measures developed in time and reveals a seasonal trend in demand. Demand is high during the summer except for 2018 when the lowest demand during the whole 6-year period occurs during the summer. An increase in demand can be seen in the beginning of each year as well, a trend that is consistent throughout the 6-year period.

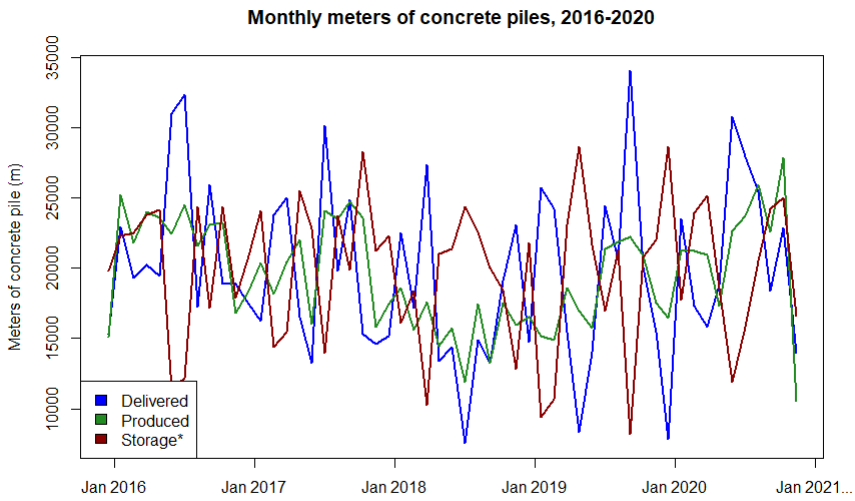


Figure 39: Monthly meters of production (green), delivery (blue) and simulated storage (red) of concrete piles, 2016-2020. Storage values are calculated on the assumption that starting storage is equal to 20 000 pile meters.

5 Discussion

This chapter begins by discussing the results in their relation to the specified purpose and continues to propose how further work towards the purpose can be carried out.

5.1 Discussion of the specfied purpose

The following section discusses theory and results in relation to the specified purpose of section 1.4.

What measures are relevant for concrete pile quality?

Concrete pile quality is mainly determined through the compressive strength, f_c, of the concrete. Concrete strength of concrete piles produced at Hercules Ucklum needs to meet 3 conditions ; $f_c > 25$ MPa at the time of removal of piles from the formwork, $f_c > 50$ MPa at the time of delivery of piles to customer, $f_c > 56$ MPa and the mean of the 35 latest strength tests must fulfill $\bar{f}_{c35} > f_c + 1,48\hat{\sigma}_{f_c35}$ at 28 days after casting the concrete. Formwork removal and delivery to customer are currently set to one and seven days after casting. Although there is a possibility to extend the three points of time by restructuring production and certification processes, 25, 50 and 56 and $60 + 1,48\hat{\sigma}_{f_c35}$ MPa must be achieved to allow for formwork removal, delivery to customer and final certification of the piles respectively.

Table 4 reports measures required for well-founded predictions of f_c and the type of information each measure constitutes. Information containing time and chemical properties have been emphasized with their own columns as these types of information stand out. Time series can result in challenges when using ML models and needs to be considered with extra care. Chemical properties of concrete ingredients has been used as predictive variables in other works considering compressive strength of concrete as the outcome, however these may not need to be considered.

The general purpose was to enable control of the piles' carbon footprint and costs. Therefore, proportional CO_2 footprint and the cost of each of the ingredients in a concrete mixture needs to be assessed as these would enable evaluation of reduction of carbon footprint and cost. See [12] for a more comprehensive account of the theory of concrete ingredients.

Concrete piles also need to be certified through the yield strength of reinforcement steel. Measures for basis of predicting steel strength needs further investigation.

What is the present state of data collection in Hercules' processes?

In the current state of Hercules' processes, most data is collected for the requirements of various certification processes. See section 4.4 for a complete

Table 4: The physical quantities and their related data types for assessing compressive strength of concrete. The grey column describes how the data of each measure is reported in Ucklum. The red column marks that for a reduction of CO_2-emissions, data on a variety of recipes must be collected and the distribution of the data must be planned carefully. The two green columns marks that CO_2-emissions and costs per unit of each quantity must be obtained in order to enable optimization of CO_2-emissions and costs. Yellow cells may be omitted in the data collection, if they are expected to be superfluous.

	Numerical	**Categorical**	**Time**	**Chemical properties**	**In Ucklum**	**Proposed distribution of data**	**CO2/unit**	**Cost/unit**
Curing time		x			Constant			
Fineness modulus of Ballast	x				Unavailable			
Cement	x	x			Unavailable			
Water	x			x	Unavailable			
Fillers	x	x		x	Unavailable			
Plasticizer	x	x		x	Unavailable			
Temperature (core)			x		-			
Temperature (ambient)			x		Insufficiently			
Humidity			x		Insufficiently			
Formwork material		x			Constant			
Curing time	x				Constant			
...								
Compressive strength	x				Decently			

description of data found during the project and fig. 5 for an overview of where the various data is produced in relation to the Hercules' production system. Column 6 in table 4 accounts for the relevant measures of compressive strength of concrete in relation to Hercules' collection of information. Some interesting data is unavailable for analysis due to software issues. The mixing reports, which hold information from multiple parameters in concrete mixtures, are produced in a local computer with a software that does not allow a compilation of multiple mixing reports. One mixing report may be rendered at a time, so to analyze the data in the mixing reports two challenges occur: Generating the mixing report which describes the concrete mixture from which the concrete cube is produced, and translating the mixing report-file into a database suitable for analysis in statistical software.

The reliability of the data is affected mainly by the fact that concrete cubes are produced with different procedures, see [12]. Cubes may vary in both number of layers of packed concrete and in size. Otherwise most data is collected automatically. The reliability of automatically collected data has not been considered.

Which ML-algorithms are appropriate to predict relevant quality measures in Hercules plant in Ucklum?

According to literature studies, efforts with similar aims to this project conclude that compressive strength is best predicted through the use of Boosted trees and Neural networks. Gradient boosted trees, especially through the CatBoost library, best meets the requirement of easy implementation. Catboost does not demand parameter tuning, although parameter tuning is likely to improve predictive performance to some extent. CatBoost automatically deals with categorical and numerical data as well as with overfitting. CatBoost also has a fast learning rate, which is advantageous as the data required for prediciting compressive strength is expensive.

A CatBoost model was tested on Yeh's data [24] and gave an R^2 of 0.93 for an untuned model evaluated on a test data set which indicates that the CatBoost is suitable for easy access ML on compressive strength of concrete.

How does the present state of data collection in Hercules' processes allow for ML?

This question has been approached by trying to predict compressive strength of concrete with variables collected prior to the strength tests. Studied ML models have been linear models and gradient boosted trees. Through this data analysis, the date when a concrete is mixed stands out as the most useful variable for ML. However, no R^2 value of any model indicate that the model would be useful to base decisions on. The date of concrete

mixing stands out with higher R^2-values, but this parameter in itself is not a physical parameter supposed to influence the compressive strength of concrete. Data on other relevant physical quantities produced too low R^2 values with linear models to evaluate the models further. When analysing the data through gradient boosted trees, the conclusion was made again that no measured physical quantity supposed to influence the compressive strength of concrete could predict the compressive strength of concrete. However, the notable drop in compressive strength during the summer should be noted and investigated further and gives means to improve the production of the cubes [12]. In addition, access to the mixing reports could add important data which would present a better case for ML in Hercules' processes.

Fig. 40 gives an illustration of the conclusions about collected data sets allowance for ML.

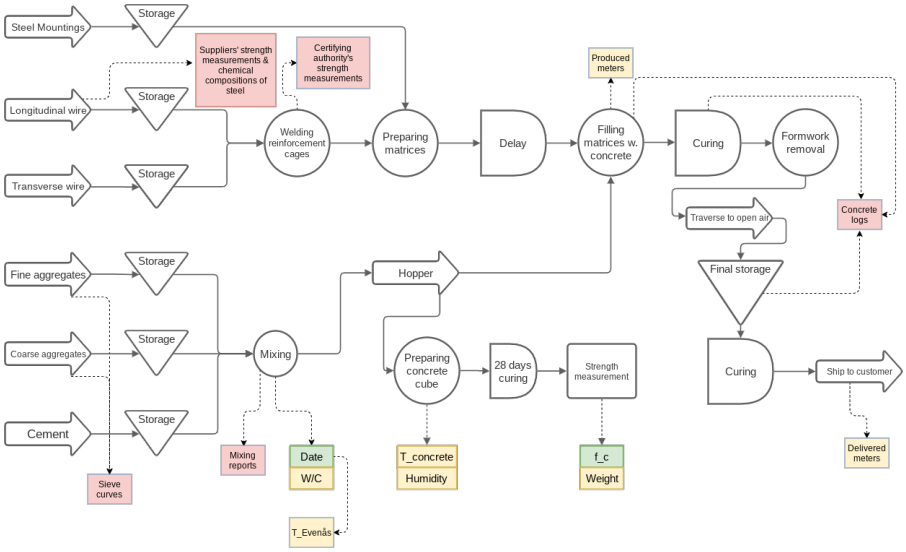


Figure 40: Overview of where in Hercules' processes the data sets found during this project are collected. Green represents data that helped achieve the aims of this project. Yellow represents data that was available for ML application but served no use during this project. Red represents data that was generated in Hercules' processes but not in an accessible format. An enlarged version of this figure is found in appendix 7.1.

In hindsight, the poor case for ML could have been anticipated. There is already a system for controlling the compressive strength of the cubes through the certification requirements. The available data is collected only to monitor that the most important parameters are within certain ranges which aims to ensure that compressive strength is controlled. This lies in contrast to collect data with the aim to explain an output variable. The project was started with the assumption that some data is collected in Her-

cules' processes which could enable a reduction of CO_2-emissions through reduced cement usage in the recipe. However, Hercules' has produced the same mixture of concrete for the all years represented in the data, 2015-2020. None of the parameters that are supposed to affect the compressive strength of concrete have purposely been made to vary and are tightly controlled through the certification system. It has emerged during the project that changes to the recipe, even minor, are too risky since a batch of concrete costs too much. Furthermore, Hercules manages their processes well enough to never need to scrap mixed concrete. In essence, there are no major errors in the production, which ML may learn from or attempt to prevent and the information from just one recipe is not sufficient to propose a new recipe. In this context, one recipe is just one data point and any deviation from this recipe would involve extrapolation. However, both the work summarized in section 4.3 and the untuned CatBoost machines performance on Yeh's data set [24] suggests that ML is a viable tool for estimating compressive strength of theoretical concrete mixtures.

The situation with the reinforcement steel data demands further attention. Data shows that there are troublesome deviances between the supplier's and the certifying organ's measurements. If these deviances are due to problems in Hercules' processes or shortfalls in any measuring process is unclear. But the reliability of these measurements needs to be understood.

As a final comment, the data analysis has revealed some potential for reduction of compressive strength of the cubes. Maybe this can translate to a reduction of CO_2-emissions and costs. The data on compressive strength reveals that Hercules' measurements from the last 6 years overperforms with 4 MPa on average compared to the recommended mean compressive strength of 67 MPa, see fig. 42. This overperformance gives room for some reduction of cement in the concrete mixtures through an increase in W/C ratio. Such an increase requires consideration of the technical properties of concrete, but it appears feasible to increase the W/C ratio with about 0.04 on average, see [12] for a technical motivation. An increase of 0.04 in the average W/C ratio in the concrete recipe used in Hercules Ucklum implies a 0.1% decrease of cement useage. However, a 0.1% decrease in cement usage, which approximates to a decrease in CO_2-emissions of 0.07% and 6000 sek per year. It can be concluded that efforts to decrease the CO_2-emissions or costs should not aim at increasing the W/C ratio or other small alterations of the proportion of cement in the recipe.

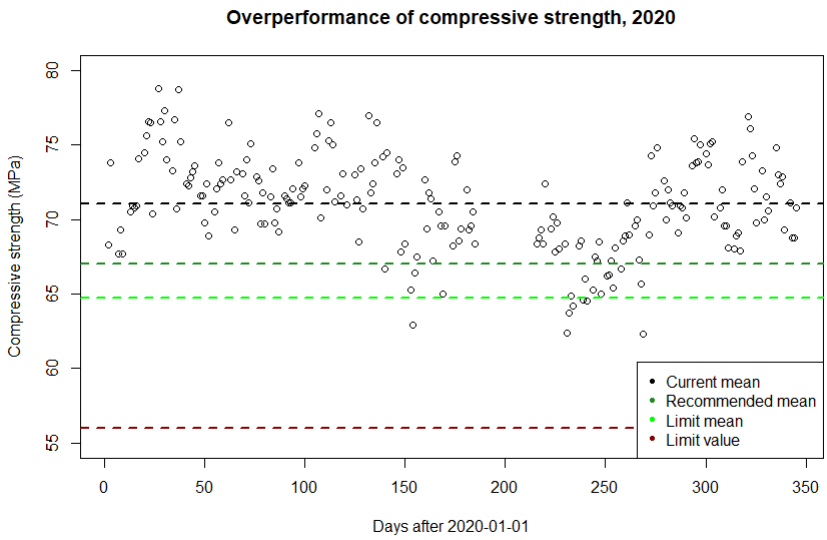


Figure 41: The current mean value of $f_c = 71$ MPa is overperforming. Recommended mean is 67 MPa, certification allows for 65 MPa with retained standard deviation. All values must fulfill $f_c > 56$ MPa.

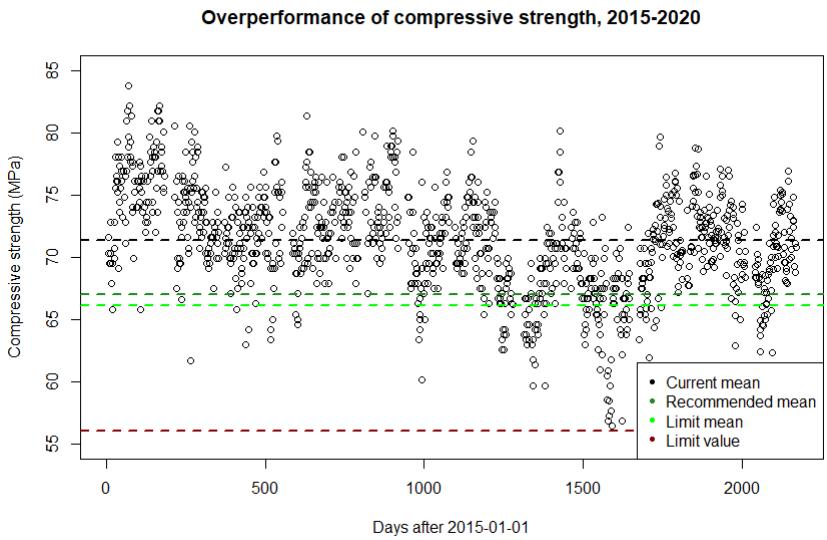


Figure 42: The analysis from fig. 41 on data from 2015-2020. The conclusion is almost the same, but note that a few data points touch the limit $f_c > 56$ MPa and that standard deviation is increased for this data.

5.2 Proposed further work

Further work with regard to quality improvements is suggested to take on two subjects depending on what the purpose is for any continued research. Three major subjects have been in focus in the communication with NCC; firstly, reduce CO_2 footprint. Secondly, reduce costs. Thirdly, comprise an example which demonstrates the benefit of efficient data collection through the use of ML.

Reducing CO_2 is recommended to aim for by investigating how fly ash or other ingredients with smaller carbon footprint can substitute cement. Future work should then comprise careful planning of how data outlined in table 4 should be collected in order to produce trustworthy predictions of concrete strength through a CatBoost model. If such data is acquired, SHAP and variable importance plots can provide insights of how the model interprets the data. Range of data in all input parameters must be considered, but also the precision with which ingredients can be distributed in Hercules' processes. If a recipe is implemented which works well in precise laboratory mixtures, it might be costly to implement in Hercules. If a batch of concrete contains more fillers in reality than what has been planned in theory, concrete compressive strength might fall below the certifying limit.

Reducing costs and demonstrating the use of data collection, is recommended to aim for with an investigation of the logistical data system. It was mentioned by two independent respondents that Hercules would benefit from developing the logistical system. It was also revealed that order times are shorter than production times on a regular basis, which causes stress within the organisation. Produced pile meter per month varies greatly, which is not in line with lean production. Aiming to improve this situation is likely to improve Hercules' economy and well being of workers as well as make time available in Hercules for engaging in other quality work activities. In addition, ML models can be usable in this setting since forecasting usually builds around non-complex ML models.

A good start to attending the logistical situation would be to provide a financial incentive for making early orders. I propose to increase the price on late orders or decrease the price on early orders by a reasonable amount.

A suggestion for NCC is also to implement a standard requiring that any data collecting system should reach or surpass a baseline level. Such a baseline level should be that data collection results in Excel files available through the cloud, or equivalently.

6 Conclusions

The main conclusions of this project are:

- This project has found that data available in Hercules Ucklum's processes provides no basis for reduction of CO_2.
- The CatBoost library provides easy access to ML through the gradient boosted trees algorithm. The SHAP package can be used to interpret created models.
- For further work, two pathways are proposed. A reduction of CO_2-emissions can be achieved by obtaining data on a variety of concrete recipes and find an optimized recipe with a CatBoost model.
- Reducing costs and demonstrating the value of data analysis is recommended to aim for via an investigation of the logistical situation in Hercules Ucklum.
- Two suggestions for facilitating Hercules' and NCC's activities are made: Quantify the cost of short-timed orders and introduce a lower limit standard for data collection systems, preferably that all data is available in Excel files.

References

[1] Hercules.se. Accessed: 2021-06-01.

[2] temperatur.nu/evenas.html. accessed: 2021-03-01.

[3] Amged O. Abdelatif, Amir M.Y. Shaddad, Mohammed B. Fathallah, Mohammed S. Ibrahim, and Mohammed H. Twfeeq. Concrete mix design and aggregate tests data between 2009 and 2017 in sudan. *Data in Brief*, 21:146–149, 2018.

[4] International Energy Agency. Cement technology roadmap plots path to cutting co2 emissions 24% by 2050 - news, iea.org/news/cement-technology-roadmap-plots-path-to-cutting-co2-emissions-24-by-2050. Accessed: 2021-03-01.

[5] Bo Bergman and Bengt Klefsjö. *Kvalitet från behov till användning.* Studentlitteratur AB, Lund, 2020.

[6] Thomas Betong. thomasconcretegroup.com/se/hallbarhet/betong-och-koldioxid. Accessed: 2021-06-01.

[7] Palika Chopra, Rajendra Sharma, and Maneek Kumar. Predicting compressive strength of concrete for varying workability using regression models. *International Journal Of Engineering Applied Sciences*, 6:10–10, 12 2014.

[8] Furqan Farooq, Wisal Ahmed, Arslan Akbar, Fahid Aslam, and Rayed Alyousef. Predictive modeling for sustainable high-performance concrete from industrial wastes: A comparison and optimization of models using ensemble learners. *Journal of Cleaner Production*, 292:126032, 2021.

[9] De-Cheng Feng, Zhen-Tao Liu, Xiao-Dan Wang, Yin Chen, Jia-Qi Chang, Dong-Fang Wei, and Zhong-Ming Jiang. Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach. *Construction and Building Materials*, 230:117000, 2020.

[10] The Norwegian EPD Foundation. Environmental product declaration, norcem standardsement fa justert, brevik - cem ii/ b-m 42,5 r. 2020. https://www.epd-norge.no/, accessed 2021-03-01.

[11] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning : data mining, inference, and prediction.* Springer series in statistics. Springer, 2009.

[12] Deco Josephson. Quality control of concrete piles based on historical data. Master's thesis, Karlstad University, 2021.

[13] Jeffrey K. Liker. *The Toyota Way: 14 management principles.* McGraw-Hill, 2004.

[14] Zain M.F.M, Suhad Abd, Roszilah Hamid, and Md Jamil. Potential for utilising concrete mix properties to predict strength at different ages. *Journal of Applied Sciences*, 10, 12 2010.

[15] Christoph Molnar. *Interpretable Machine Learning A Guide for Making Black Box Models Explainable.* leanpub.com, 2021.

[16] Hoang Nguyen, Thanh Vu, Thuc P. Vo, and Huu-Tai Thai. Efficient machine learning models for prediction of concrete strengths. *Construction and Building Materials*, 266:120950, 2021.

[17] Khoa Tan Nguyen, Quang Dang Nguyen, Tuan Anh Le, Jiuk Shin, and Kihak Lee. Analyzing the compressive strength of green fly ash based geopolymer concrete using experiment and machine learning approaches. *Construction and Building Materials*, 247:118581, 2020.

[18] Boya Ouyang, Yuhai Li, Yu Song, Feishu Wu, Huizi Yu, Yongzhe Wang, Mathieu Bauchy, and Gaurav Sant. Learning from sparse datasets: Predicting concrete's strength by machine learning, 2020.

[19] Umur Korkut Sevim, Hasan Huseyin Bilgic, Omer Faruk Cansiz, Murat Ozturk, and Cengiz Duran Atis. Compressive strength prediction models for cementitious composites with fly ash using machine learning techniques. *Construction and Building Materials*, 271:121584, 2021.

[20] Concrete – Specification, performance, production and conformity. Standard, Swedish Institute for Standards (SIS), Stockholm, May 2018.

[21] The International EPD® System. Environmental product declaration for hercules concrete piles. 2020.

[22] Katarina Wärmark. Sveriges territoriella utsläpp 2020 (preliminära).

[23] Yandex. catboost.ai. Accessed: 2021-06-01.

[24] I-Cheng Yeh. Modeling of strength of high-performance concrete using artificial neural networks." cement and concrete research, 28(12), 1797-1808. *Cement and Concrete Research*, 28:1797–1808, 12 1998.

[25] Patryk Ziolkowski and Maciej Niedostatkiewicz. Machine learning techniques in concrete mix design. *Materials*, 12(8), 2019.

7 Appendix

7.1 Enlarged figures 10, 18, 45

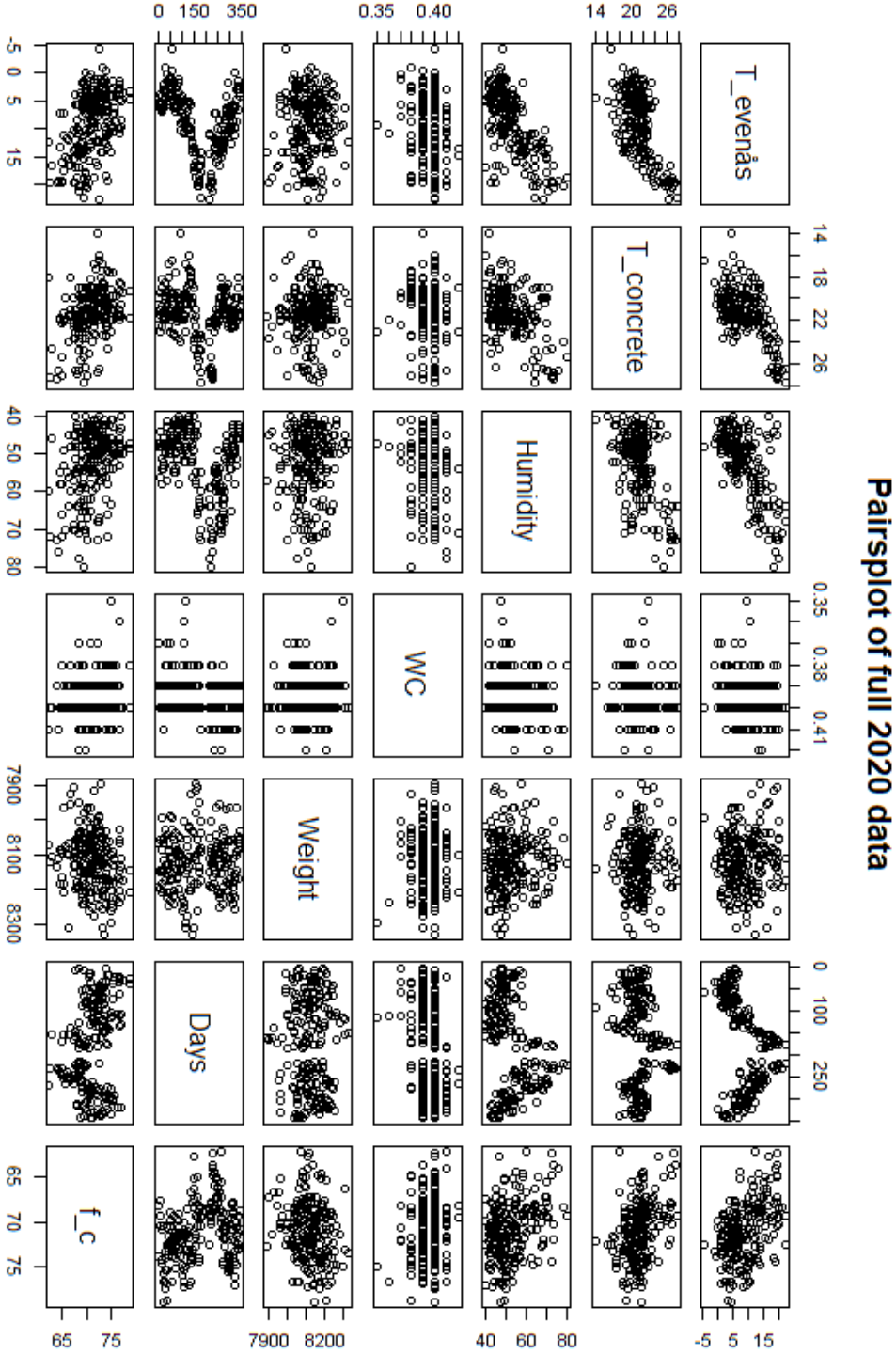


Figure 43: Enlarged pairsplot of the full 2020 data.

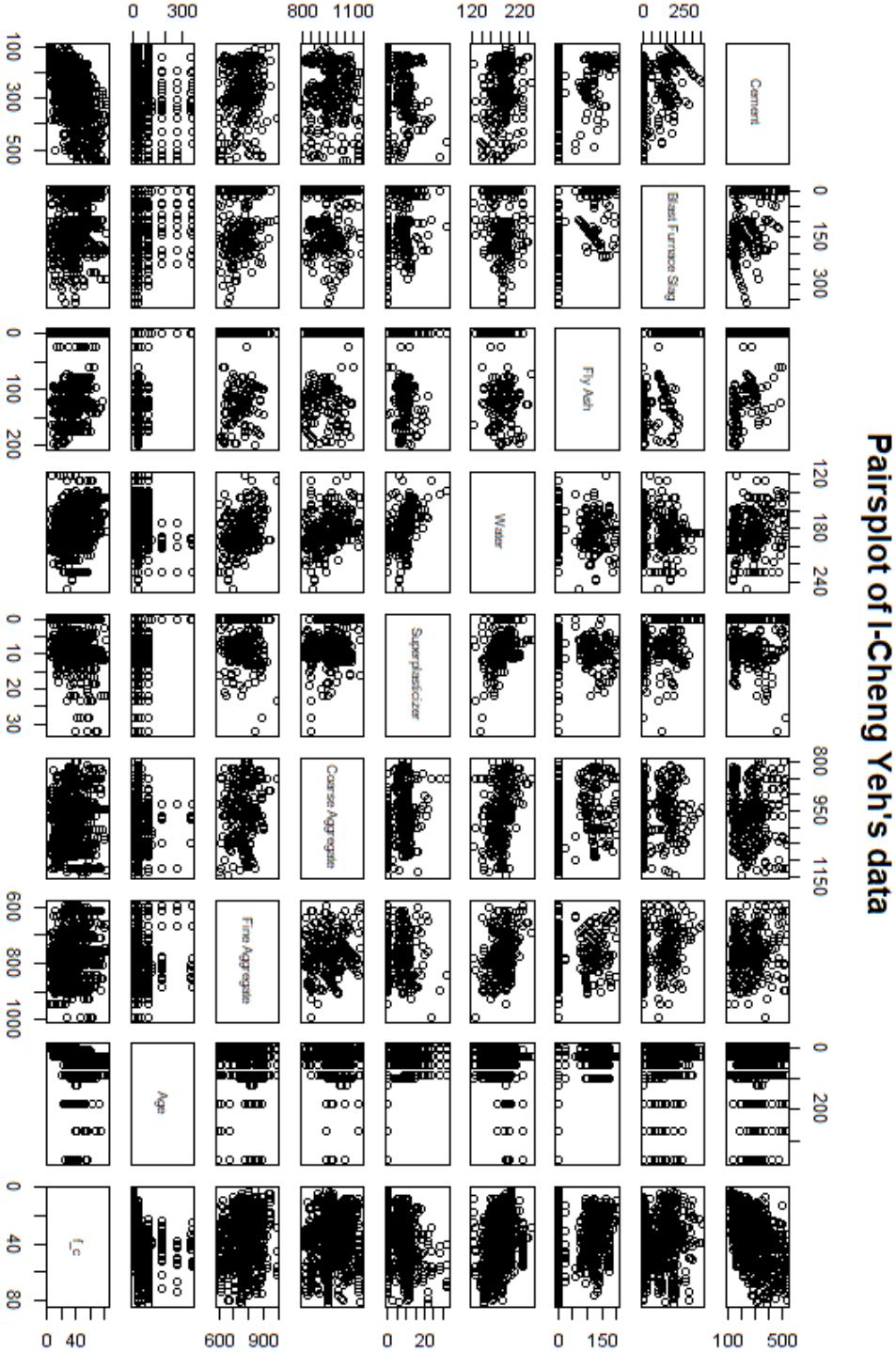


Figure 44: Enlarged pairsplot of I-Cheng Yeh's data.

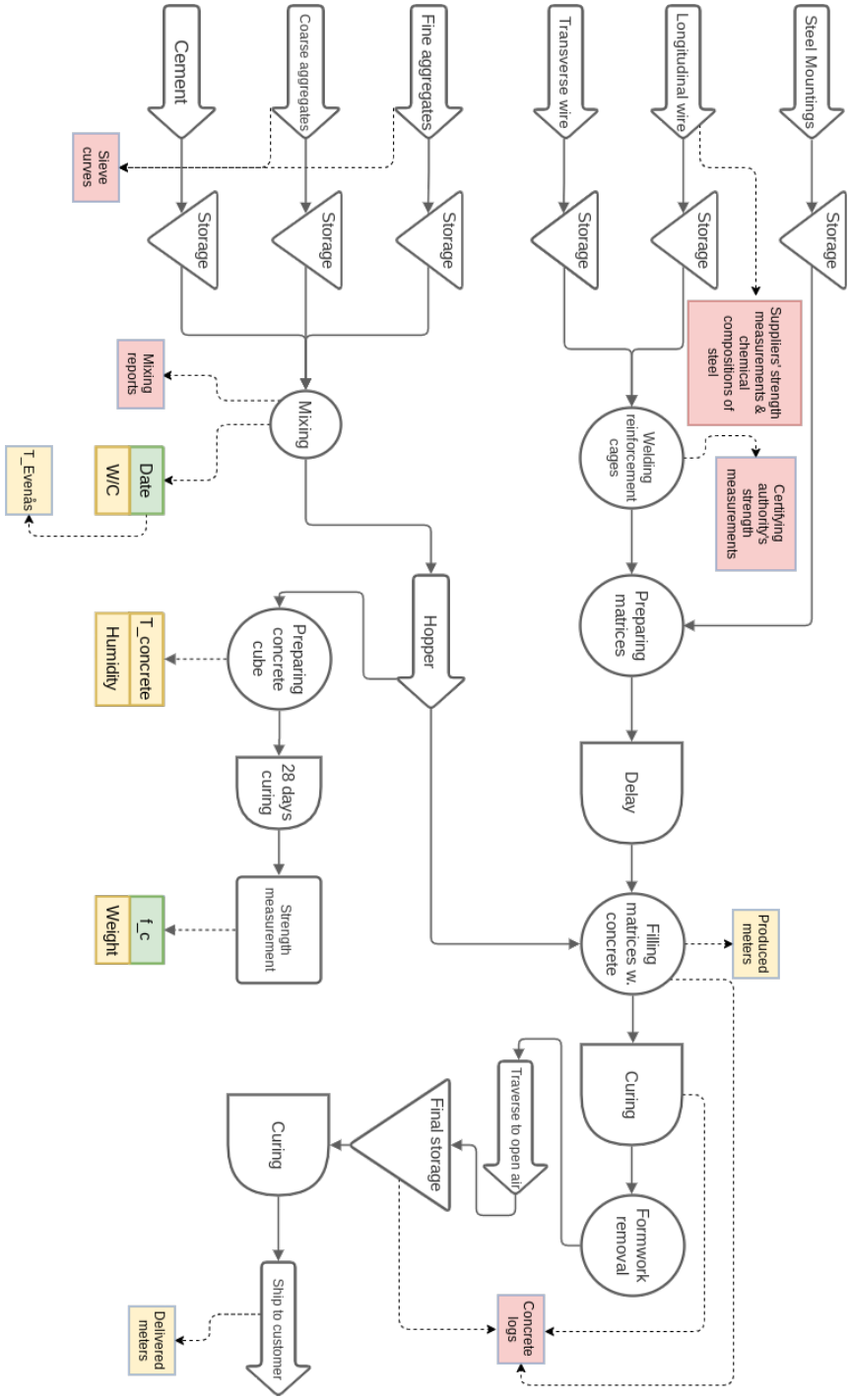


Figure 45: Enlarged overview of where in Hercules' processes the data sets found during this project are collected. Green represents data that helped achieve the aims of this project. Yellow represents data that was available for ML application but served no use during this project. Red represents data that was generated in Hercules' processes but not in an accessible format.