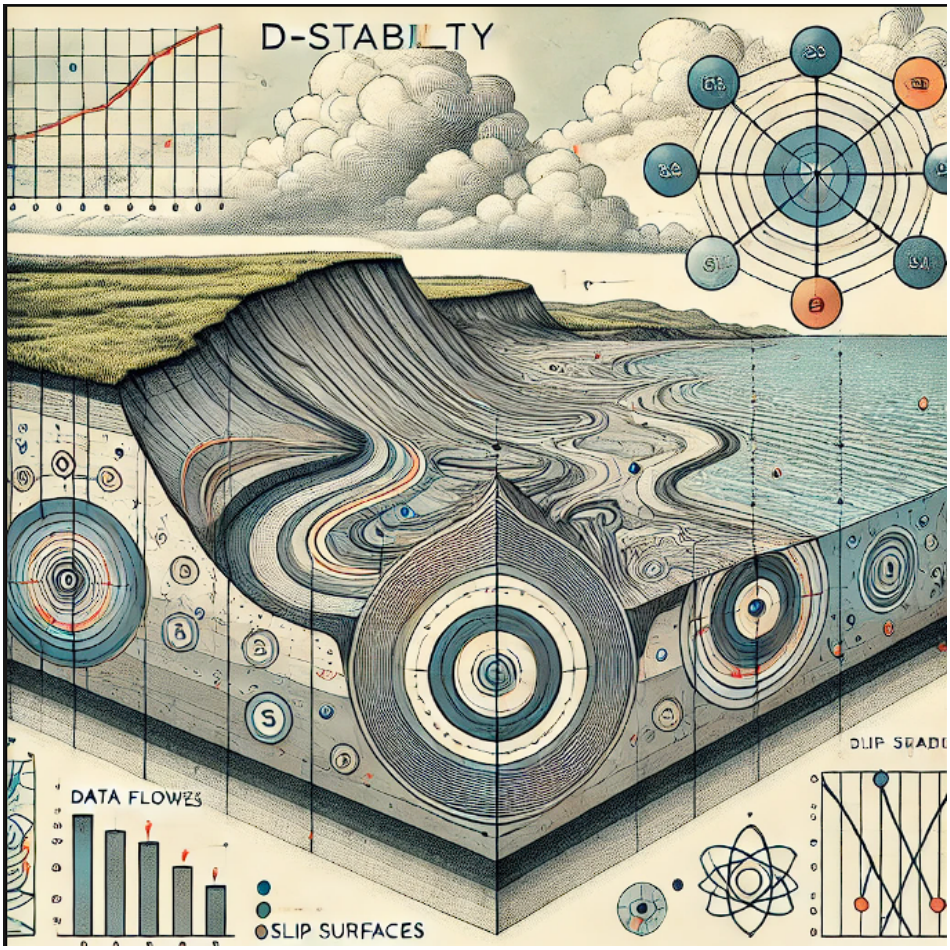
Investigating the Potential of Machine Learning Methods to Predict Soil Variables for Dike's Macro Stability Analysis

Mostafa Yaghi, S2180014
Internal supervisor: Hongyang Cheng
External supervisor: Werner Halter

October 2024



FUGRO UNIVERSITY OF TWENTE.

Contents

List of Figures

List of Tables

Acknowledgement:

This research marks the end of my academic journey, and I want to extend my most profound appreciation to those who have supported me. First and foremost, I am profoundly thankful to my parents, who guided me to finish this journey with their unwavering encouragement, guidance, and belief in me throughout this endeavor.

I am also immensely grateful to Werner Halter and Hongyang Cheng, who generously provided knowledge and expertise to complete this research. Their mentorship provided me with the direction and insight needed to succeed.

Furthermore, I would like to thank my colleagues at Fugro for their ongoing support and readiness to assist whenever needed. Special recognition goes to Ben Rijneveld & Anne Backer, whose assistance and willingness to answer my countless questions made a significant difference in my work.

To all who contributed to this journey, thank you.

Definition of key terms

The following variables are defined and explained in the book of Budhu, 2011.

Table 1: Symbols and Descriptions of Soil variables

Symbol	Description	Unit
$\boldsymbol{LL}$	**Liquid limit**: The water content at which a soil changes from a plastic state to a liquid state.	%
$\boldsymbol{PL}$	**Plastic limit**: The water content at which soil changes from semisolid to plastic.	%
$\boldsymbol{PI}$	**Plasticity index**: The range of water contents over which the soil deforms plastically.	%
$\boldsymbol{q}$	**Deviatoric stress**: The shear or distortional stress or stress difference on a body.	kPa
$\boldsymbol{W_c}$	**Water content**: The ratio of the weight of water to the weight of solids.	%
$\boldsymbol{W_{opt}}$	**Optimum water content**: The water content required to allow a soil to attain its maximum dry unit weight following a specified means of compaction.	%
γ	**Bulk unit weight**: The weight density, that is, the weight of the soil per unit volume.	kN/m^3
$\boldsymbol{\gamma_d}$	**Dry unit weight**: The weight of the dry soil per unit volume.	kN/m^3
$\boldsymbol{\gamma_{sat}}$	**Saturated unit weight**: The weight of the saturated soil per unit volume.	kN/m^3
$\boldsymbol{\sigma}$	**Total stress**: The stress of the soil particles and the liquid and gases in the voids.	kPa
$\boldsymbol{\sigma'}$	**Effective stress**: The stress carried by the soil particles.	kPa
$\boldsymbol{\phi'}$	**Effective friction angle**: A measure of the shear strength of soils due to friction.	–
$\boldsymbol{V_s}$	**Volume** of the proctor sample $= 944cm^3$	m^3
$\boldsymbol{g}$	**Acceleration** due to gravity	m/s^2
$\boldsymbol{CC}$	**Clay content**: Clay fraction in a soil type	%
$\boldsymbol{SiC}$	**Silt content**: Silt fraction in a soil type	%
$\boldsymbol{SC}$	**Sand content**: Sand fraction in a soil type	%
$\boldsymbol{OC}$	**Organic content**: Organic fraction in a soil type	%
$\boldsymbol{SaC}$	**Salt content**: Salt fraction in a soil moisture	%
$\boldsymbol{M}_{\text{loss}}$	**Mass loss**: Mass loss due to the use of "chemical"	%
$content63$	Mix between the two data groups $>$ and $<$ than 63	%
G_s	**Specific gravity:** is the ratio of the weight of the soil solids to the weight of water of equal to volume, assumed as 2.7	[-]
ρ_d	**Dry density** The mass of solid soil particles per unit volume of soil	[kg/m^3]
ρ_{dp}	**Proctor dry density** The mass of solid soil particles per unit volume of soil inside of the proctor test	[kg/m^3]

Abstract

This report examines the idea of predicting Dike's stability analysis variables using machine learning and linear regression models. This is done to address the limitations of the current practice, which involves obtaining the variables through three tests: clay inspection, clay compaction, and triaxial tests. While the clay inspection is straightforward, the other two are expensive, time-consuming, and not available in the design phase of the project. Furthermore, those tests are unavailable in the design phase, forcing the designers to estimate the needed variables. These limitations underscore the need for a more efficient and accurate method for predicting these variables.

On that point, before performing these tests, the current practice estimates the clay variables by knowing the clay and sand ratio of the clay. Then, it predicts the unit weight and friction angle by assuming weak, moderate, and strong consistency indices. Therefore, the thesis hypothesis explores predicting the dry unit weight and friction angle before performing the compaction and triaxial tests and explores the idea of having a more precise estimate of the variables following the same order as the current estimation process.

This study involved gathering Dutch data from the Fugro database, which was then input into two machine learning models: the Neural Networks and Random Forest. These models efficiently processed the data frames and predicted the variables using a set of scenarios. Each scenario represents what an engineer would know about the clay sample in the design phase. Scenario one explored the idea of having the Atterberg limits; scenario two explored knowing the contents of the clay (clay, sand, and silt), while scenarios three and four explored knowing the Triangular classification (NEN 5014) of the clay and the plasticity diagram categories, respectively.

Then, the research showed that using the Random Forest led to better predictions, with scenario one having the best prediction towards the other variables, especially the dry unit weight. This was then combined with estimated water content variables to calculate the unit weight. These water content variables were based on different consistency index measures (0.60, 0.75, and 0.85). Therefore, the study shows that it is possible to predict the unit weight with low uncertainty by knowing the plasticity diagram classifications and Atterberg limits, which formed a combined scenario to mimic the order of the current estimation method. Furthermore, the research could not use the triaxial data because of the poor quality and small amount of data, isolating the friction angle from being predicted as the unit weight.

In conclusion, the research underscores the potential for significant improvement in the current practices. This is particularly evident in the difference in the unit weight values between the different clay classes, a factor not accounted for in the current estimated values. The research also suggests that the unit weight could be predicted accurately for the design phase, eliminating the need for the compaction and triaxial tests for more accurate estimation.

Keywords: Neural Networks, Random Forests, Auto-encoders, Clay contents, Atterberg limits, Dry unit weight, Friction angle, Clay inspection, Clay compaction, Triaxial tests, stability analysis.

1 Introduction

Numerous dike reinforcement and construction projects are undertaken in the Netherlands to protect against failure mechanisms. Testing a dike design against specific failure mechanisms requires inputting different strength variables of the clay into different testing models, like D-stability (developed by Deltares). These models use shear strength models like SHANSEP and MohrCoulomb to test the design's stability, Van der Meij, 2020. On that note, the soil testing phase is one of the most crucial phases of a geotechnical project, as it ensures the suitability of the soil and its bearing capacity to the project, Aimil Ltd., 2024.

Therefore, the design and construction of dikes rely heavily on accurate soil variables data, influencing many decisions throughout the project's lifespan. However, obtaining these soil variables is a labor-intensive and time-consuming process. This necessity has driven dike reinforcement designers to seek quicker and more reliable assessment techniques (of the variables) to perform these tests with reliable variables' values according to the required norms.

Current practices and challenges:

The current practice, as noted by PovDGG, 2022, involves a progression from coarse to accurate estimation of varaibles. It starts with determining global suitability based on area knowledge, the expertise of designers, and borrowed information from other projects. The soil is then examined, and its suitability is determined visually and with increasing precision. To achieve an accurate estimation level, specific tests are required to comply with established norms: the clay inspection, compaction, and triaxial tests. While the inspection test is straightforward, the compaction and triaxial tests are more complex, requiring substantial resources and time.

A significant limitation of the current practice is that the predictions are not verified after construction, and it is impossible to perform tests on new dike embankment soil in the design phase, as it does not yet exist (on the dike location, as soils need to be moved to the area in lumps). Furthermore, according to O'Sullivan, 1997, disadvantages of soil analyses include the difficulty of obtaining methods suited to varied soil types and problems in sampling due to soil variation across a field. As a result, designers rely on borrowed data.

Importance of predictive models:

Considering these challenges, based on the results of clay tests and soil identification, a predictive model for dike materials would be beneficial if solid correlations are found to predict the needed variables for stability analysis, namely, unit weight and Friction angle.

This thesis investigates the potential of establishing such correlations using traditional and Machine learning methods, focusing on the aforementioned three primary clay tests. These tests cover most of the variables needed for constructing a dike. Therefore, the correlations between these variables are investigated with a greater focus on the unit weight and Friction angle, as they are the main variables used for stability calculations, Van der Meij, 2020. In other words, it would be extremely beneficial to predict mechanical variables (compaction and triaxial) using the physical variables (clay inspection) of the clay sample. Table 2 shows the variables of these tests, with the order of the test from most accessible (on the left) to the hardest (to the right).

Table 2: Summary of Soil varaibles from Different Tests

Clay Inspection Test	**Clay Compaction Test**	**Triaxial Test**
Clay content	Dry density	Saturated unit weight
Sand content	Proctor dry density	Dry unit weight
Silt content	Water content	Friction angle
Organic content	Dry unit weight	Class of clay (descriptive)
Liquid limit		Deviatoric stress
Plastic limit		Water content
Plasticity index		Horizontal effective stress
Class of clay		Vertical effective stress

1.1 Hypothesis:

Given the available information at the beginning of a project and the easily accessible tests, namely, borrowed information from other projects and inspection tests, it is hypothesized that complex correlations could be established using probabilistic and machine-learning algorithms. These correlations could predict variables needed for stability analysis, potentially eliminating the need for expensive and time-consuming clay tests such as compaction and triaxial.

1.2 Research questions:

Main Question:

How can the dike embankment soil variables for dike stability calculations be better predicted based on borrowed source information available in the design phase (before construction)?

Sub-reserach Questions:

1. **What is the state of the data?**
   (a) How is the data collected?
   (b) Which cleaning and filling methodologies are used to reach the final data state?
   (c) What are the statistics for the data after cleaning?
2. **How are the data correlated internally?**
   (a) Which varaibles have noticeable correlations?
   (b) How do these correlations vary across different clay classes or soil types?
3. **To what extent can Machine Learning models predict specific varaibles?**
   (a) What are the optimal hyperparameters for these models?
   (b) How does model performance compare across different input variable combinations?
4. **What is the level of uncertainty of the found correlations?**
   (a) To what extent do machine learning models give a better prediction than the current experience-based method?
   (b) What are the recommendations for new tests on dike embankment material to make them more useful as input for machine learning in the future?

1.3 Report layout:

Figure 1 shows the report's design. First, the report explains the background information of the thesis; it touches upon topics like the role of clay in dikes, how the dikes are transported, what are the collection steps of the data variables, and explains the testing procedures to get them for the clay inspection, compaction, and triaxial tests Section 2. Then, the report shows vital information on the clay's behavior, like Atterberg limits, erosion categories, and macro stability calculations, Section 3. After that, the report mentions the found correlations in the literature between the variables; it then shows what machine learning models other researchers in the field used, Section 4

From that point onwards, the report starts answering the thesis questions, showing the procedures for collecting the data, preprocessing it, and eventually displaying it along with its statistics, Section 6. Then, the thesis begins exploring the methods for applying the selected machine learning methods, optimizing the models, and using data manipulation techniques to train them, Section 7.7. From there, the results of the models are shown and compared with each other, which marks the end of question 3, Section 8. Lastly, the best model's results are assessed on its uncertainty. Then, it shows how these predictions affect the current practices and explains if the results could be used instead of them, Section 9.

Lastly, the results are discussed to give the reader an idea of what has affected the results, and then recommendations are provided to improve any similar studies in the future.

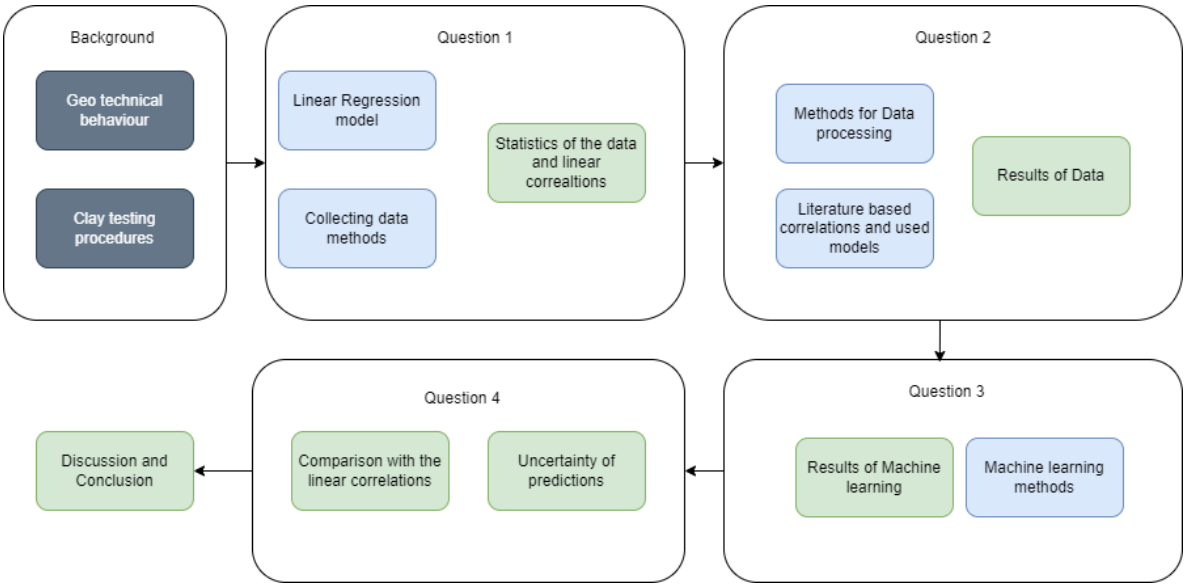


Figure 1: Report layout

Reader Guide: The research explores many aspects that may not be important to all readers. If you are a reader who is interested primarily in the predictions and statistics of the data, then I recommend that you skip Sections 7 and 8. On the other hand, if you are interested in Machine learning, then the previous sections may be of importance as they explore the architecture of the ML models.

2 State of the art: Clay testing procedure:

Soil testing is one of the most critical aspects of geotechnical projects. According to Aimil Ltd., 2024, this procedure examines whether the soil has the required quality and bearing capacity to withstand the expected buildings, dams, or bridges. This phase is crucial to determining the soil's characteristics, assessing the site's suitability, designing the foundation, ensuring safety, and saving costs. The following section explores the importance of clay and why it gets transported to the dike location. It also explores the current practice of estimating clay variables from coarse estimations using defined categories to fine estimations using different clay tests.

2.1 Role of clay in dike construction:

Clay is a versatile material strategically placed in different sections of a dike, including the top, core, and layer beneath the dike. The top layer covers the inner and outer slopes as well as the horizontal surface of the dike. A clay top layer is specifically applied near the surface to enhance the dike's resistance against various failure mechanisms, which are more likely to occur if this layer is not sufficiently strong. To reinforce this, the slopes of the dike's core are often covered with grassland or stones, further improving resistance.

The primary function of clay is to limit water permeability to the dike's core Delft, 1996. The core, which forms the structural heart of the dike, provides most of its strength, while the buried clay layer helps reduce water flow to the underlying sandy layers.

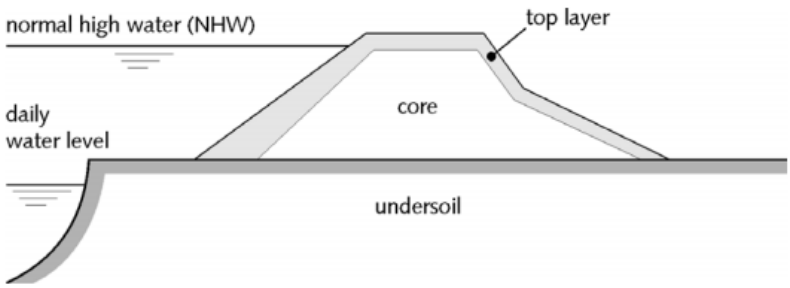


Figure 2: Illustration of Dike's top layer & core, Delft, 1996

2.2 Compaction of dike materials:

Building Dutch dikes uses transported clay, which consists of loose pieces and lumps. Therefore, the compaction of the dike's materials is necessary to work these lumps together so that the large pores and holes are pressed together. Thus, the clay must be applied in no more than 0.40m thick layers, each compacted separately using a bulldozer, Delft, 1996 & Halter et al., 2018. After compaction, the dry density must be at least 97% of the proctor density (compaction test; one-point Proctor test). This increases stability, reduces permeability, improves load-bearing capacity, controls settlement, and prevents structural failure at the in-situ water content.

Figure 3: Compaction of dikes Halter et al., 2018

Halter et al., 2018 also mentioned that maximum compaction is required for maximum stability. In other words, the stability of a constructed foundation layer with a compaction degree of 98% will generally be less than 50% of the stability at 100%. It is necessary to control the compaction by calculating the compaction degree per foundation layer to achieve maximum compaction. The control compaction measurements are carried out according to the applicable standard RAW specifications, and the laboratory is accredited by the Dutch Accreditation Council (accreditation number L 034).

2.3 Variables collection steps

When starting a new dike project, PovDGG, 2022 argues that the designer needs to know the relationship between functional requirements and the requirements set in standard specifications. On that note, among many functional requirements that the author pays attention to, minimal strength and stability are the essential functions that a dike design needs to satisfy, which interests this research.

PovDGG, 2022 mentioned a process that usually works towards accurate estimation while beginning with a coarse assessment based on regional knowledge. Then, the soil is examined, and its suitability is determined more precisely through visual inspection and testing. The author then distinguishes 5 levels of assessments:

1. Desk study based on regional knowledge.
2. Visual assessment of soil samples.
3. Classification test.
4. Standard tests for dike materials
5. Custom research/ special tests.

Assessment levels (AL) one and two are usually conducted to determine the general suitability and compare multiple soil deposits. Furthermore, their required time and costs are limited as no tests are performed. AL three and four are done to assess promising soil batches regarding how the soil variables relate to threshold values. The author also mentioned that it is economically advantageous first to perform a sufficient number of classification tests to account for any possible variations and possibly perform the standard tests on possible suitable batches. The last design level is only undertaken for special design circumstances.

On that note, the author also mentioned that the **Desk Study** usually consists of using multiple resources like Dinoloket, an archive of many soil testing agencies that are publicly available, a map of clay content, another map with organic matter content in soil layers, and archived soil testing results from dike reinforcement projects. In other words, those sources try to provide a reasonable amount of clay variables that allow the designer to start predicting the needed ones for the design. Those estimated values could be based on a national level or some report or project. Such estimations are like a value between 14-21 [kN/m3] for saturated unit weight for different types of soils, but 17 or 18 for various types of clay, 17 for an unspecified type of clay, and 18 for a sandy one, Table 4

Furthermore, Table NEN 6740 is also used, Table 3, which starts by defining the type of the clay, followed by a class of the consistency index. These two categorizations estimate the clay's unit weight, friction angle, and other variables.

Table 3: Current practice values (portion of Table NEN 6740)

Admixture	**Consistency**	**unit weight**
Clean	Weak	14
	Modeate	17
	Fixed/solid	19 or 20
weak sandy	Weak	15
	Moderate	18
	Fixed/solid	20 or 21
strong sandy	-	18 or 20

Table 4: Estimated values taken from regional knowledge

Area	**unit weight [kN/m3]**	**Friction angle [degree]**
Landelijk (klei)	17.0	17.5 [characteristic value]
Landelijk (klei, sterk zandig)	18.0	22.5 [characteristic value]
Landelijk (klei, zwak zandig)	18.0	27.5 [characteristic value]
Landelijk	14-21	32 [average]
Dijkversterking Neder-Betuwe	17.0	missing
Dijkversterking Wolferen - Sprok	17.5	27 [avg.characteristic]
Piekberging Haarlemmermeer (afdekklei)	17.0	16.9 [2%]
Piekberging Haarlemmermeer (kernklei)	17.7	32.4 [2%]
Markermeerdijk Hoorn - Amsterdam	16.5	missing
Eiland van Dordrecht	17.0	16.87 [2%]

The author then mentions that the clay testing phase begins with a **visual inspection**, where the clay is classified using two norms: NEN-5104 and NEN-EN-ISO 14688-1. The NEN-EN-ISO 14688-1 provides information about the clay's erosion category, which gives insight into its applicability for the dike design. However, this research is more interested in the NEN-5104 classification.

2.4 Clay inspection test: Physical variables

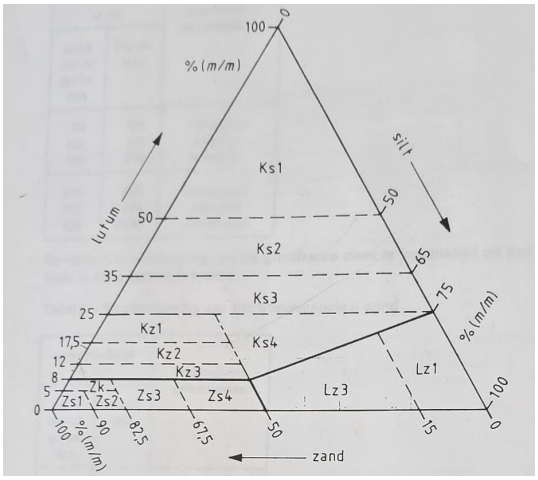
A clay inspection test is performed for the visual assessment and classification test of soil (levels 2 and 3). According to Delft, 1996, the **visual inspection** is essential to determine the following aspects of the material: extreme discoloration; strong, deviating smell; homogeneity; impurities; sand content; chalk content; consistency/hardness. Figure 4 shows two examples of this visual inspection.

Figure 4: Snips of soil visual inspection

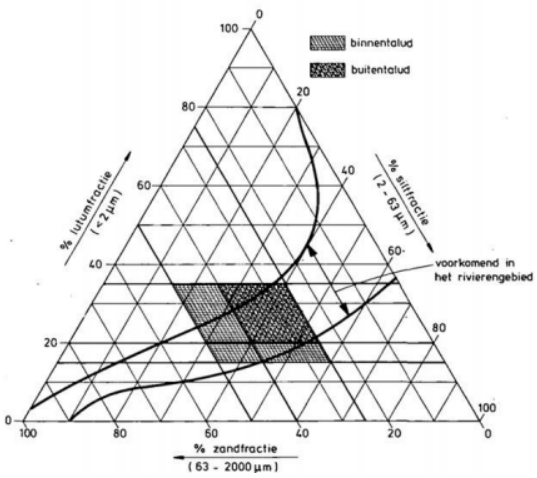
Furthermore, soil can be classified according to the Dutch Standard NEN geotechnics 5104, for which the clay, silt, and sand contents and the amount of organic matter of a sufficient number of samples are estimated. NEN 5104 uses the triangular relationships between clay, silt, and sand for classification, shown in Figure 5a. On this note, this test is essential for the site investigation phase, where the results influence the decisions further in the project process. Furthermore, it is crucial to note that this triangle assumes that only the clay, silt, and sand content form 100 %, which neglects the organic, salt, and chalk content. This research interests the clay-dominated classes shown in Table 5. Other classes in the figure below are more silt- and sand-dominated. On that note, PovDGG, 2022 shows the frequent clay classes river dikes in the Netherlands, mainly belonging to the classification area as shown in Figure 5b.

Furthermore, PovDGG, 2022 mentioned that the Raw standard clay inspection tests determine the following varaibles. Appendix D shows an example of the clay inspection report.

1. Atterberg limits (RAW test 14)
2. sand content as well as the mass of the grains $> 63\ \mu m$ (RAW test 2)
3. Salt content (NaCl) (Raw test 36)



(a) Triangular relationships clay-silt-sand and organic
Delft, 1996



(b) River clay classes in NL
PovDGG, 2022

Figure 5: NEN 5014 triangular relationship for clay classification caption

Table 5: Clay classifications

Class	Description
Ks1	Clay, slightly silty
Ks2	Clay, moderately silty
Ks3	Clay, strongly silty
Ks4	Clay, extremely silty
Kz1	Clay, weakly sandy
Kz2	Clay, moderately sandy
Kz3	Clay, strongly sandy

4. HCI mass loss (Raw test 37)
5. Clay content or grain size $< 2\ \mu m$ (RAW test 29)
6. Select good quality clay (not too organic).
7. Determine the erosion category of the clay sample.
8. check the consistency index during construction.
9. One-point proctor density immediately after installation with existing moisture content, which is the clay compaction test, Section 2.5. (RAW test 9)
10. Water content at a particular consistency index and control for inconsistencies (which is not an aspect that this thesis is interested in).

2.5 Compaction tests: Mechanical variables

As clay needs to be compacted when building a dike, it is essential to know the relationship between the maximum dry density and the optimum moisture content. Therefore, The proctored test is a laboratory test that determines the following variables: degree of compaction, dry density in situ, proctor dry density, and maximum water content. Furthermore, it assesses soil's ability to be compacted to a specified density under controlled conditions. The test is developed to deliver a standard amount of mechanical energy to the soil sample using the apparatus shown in Figure 6, Budhu, 2011. Snips of the actual test are provided in Figure 8. Appendix D shows an example of the clay compaction report.

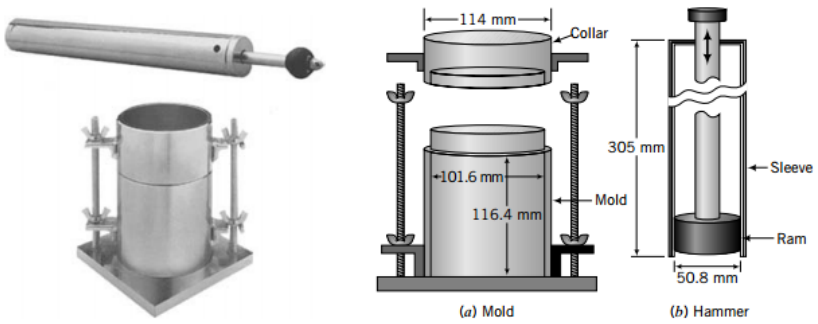


Figure 6: Compaction apparatus Budhu, 2011

2.5.1 Proctor Curves:

The results of the Proctor curve are plotted in a figure where dry unit weight is plotted against water content. Here, results tend to take a bell-shaped curve, especially for clay (compared to other materials like sand). This curve's optimum water content is the water percentage registered at the maximum dry unit weight. Below it, air is expelled, and water facilitates the rearrangement of soil grains into a denser configuration. On the other hand, when water content is just above the optimum, the compaction effort cannot expel more air, and additional water displaces soil grains, thus decreasing the number of soil grains per unit weight. Consequently, the dry unit weight decreases. Figure 7 shows an example of this proctor curve. This test is essential to control that dry density at in-situ moisture content is higher than 97% of proctor density, which is a Dutch norm. More on this is explained in Section 2.2

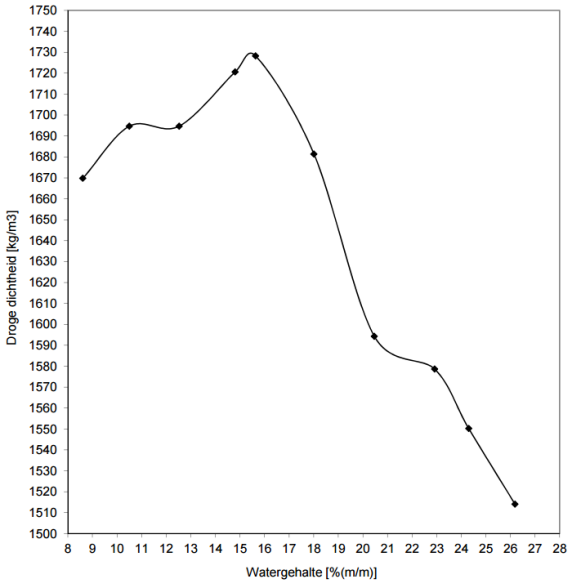


Figure 7: Proctor curve, (Taken from a Fugro report)

2.6 Triaxial tests: Mechanical variables

A triaxial test is a standard laboratory procedure used to determine the mechanical properties of deformable solids. In a standard triaxial test, a cylindrical soil sample is subjected to a confining hydrostatic pressure applied equally in all horizontal directions. This confining pressure simulates the in-situ stress (stress the soil experiences in its natural environment) acting on the soil particles.

An an-isotropic test applies unequal horizontal stresses to the sample, meaning the confining pressure can differ in different horizontal directions. In this case, the research looked only at undrained an-

Figure 8: Snips of soil compaction test, taken from Fugro website

isotropic triaxial tests describing the dike's behavior against water. Therefore, they are prescribed for dike stability variables behavior.

On that note, this thesis is interested in the following variables out of the triaxial tests: (1) unit weight, (2) moisture content, (3) friction angle, (4) deviatoric stress, and, lastly, a visual description of the sample. Furthermore, the Triaxial test uses the Misnomer apparatus, shown in Figure 9. Appendix D shows an example of the clay Triaxial report.

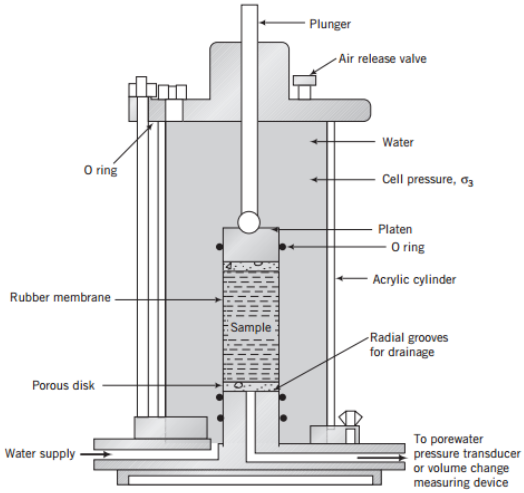


Figure 9: Schematic of a Triaxial cell, Budhu, 2011

Furthermore, testing the stability of the dike design for macro stability is usually done using D stability software developed by Deltares. According to Van der Meij, 2020, shear strength models, the SHANSEP and MohrCoulomb models, are used for the stability calculations, which depend on the friction angle, unit weight of the soil, and cohesion. However, cohesion plays no role in the macro stability calculation according to WBI, 2021.

3 Geo technical behavior:

The following section explores some of the essential aspects of geotechnical behavior in the context of dikes, enriching the reader's understanding of the procedures used when building a dike. It starts by defining the Atterberg limits and then describing the clay structure. This section also explores the consistency index and describes the erosion categories used for a plasticity diagram classification system. Eventually, a summary of how macro stability is calculated in D-stability is provided.

3.1 Physical states and index properties

According to Budhu, 2011, there are 4 physical states of fine-grained soils: solid, semi-solid, plastic, and liquid. When plotting a volume diagram versus water content, the four states could be represented in Figure10.

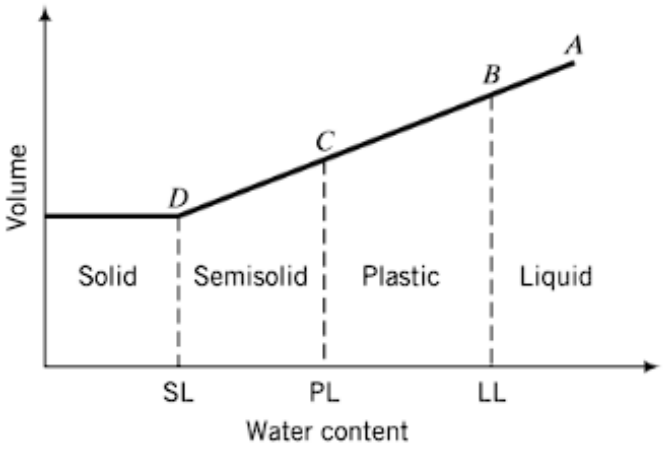


Figure 10: soil states Budhu, 2011

This figure shows that the liquid state is at point A. At point B, the soil becomes so stiff that it can no longer become liquid. The boundary between B and A is then called the **Liquid Limit (LL)**, which is the water content at which soil changes from a plastic state to a liquid state. As the soil continues to dry, there is a range of water content at which the soil could be molded into different shapes without rupture. This is where the soil has plastic behavior—the ability to deform without rupture. In this framework, if the soil continues to dry, it reaches the semi-soil state, at which point the soil cannot be molded without cracks. The water content at which the soil changes from a plastic to a semi-solid is known as the **Plastic limit (PL)**. On that note, the range of water at which the soil deforms plastically is known as the **Plastic index (PI)** .

$$PI = LL - PL \tag{1}$$

Budhu, 2011 mentioned the typical Atterberg limit for different soils. In the case of clay, typical ranges are mentioned in Table 6.

3.2 Clay Structure:

According to Knappett and Craig, 2012, the basic structure of clay minerals influences their properties and potential geotechnical implications.

Table 6: Typical ranges for Atterberg limits

Atterberg limit	typical ranges
LL	30-150 [%]
PL	25-50 [%
PI	15-100 [%]

- Kaolinite: Strong bonds between silica and gibbsite sheets limit substitution, leading to minimal swelling and stable behavior. This is primarily found in clay fractions due to their fine particles and is typically associated with less plastic clay.
- Illite: Weak bonds due to potassium ions between sheets allow some substitution and moderate swelling potential.
- Montmorillonite: Fragile bonds due to water and exchangeable cations facilitate significant substitution and high swelling potential, posing geotechnical challenges. It is also found in clay fractions due to its fine particles, like Kaolinite. However, it is associated wit highly plastic & expansive clays.

Understanding the mineralogy of soil provides insights into its expected behavior, aiding in anticipating and mitigating potential geotechnical issues related to swelling and stability.

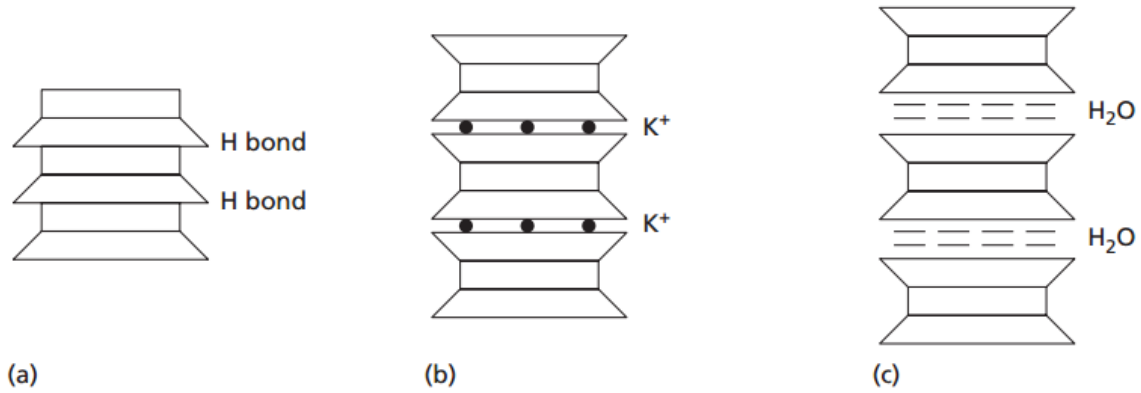


Figure 11: Clay mineral Knappett and Craig, 2012

3.3 Consistency-index:

Delft, 1996 mentioned that the consistency index indicates how the water content of clay is related to the flow and plastic limit. This relation estimates the maximum amount of water that clay can retain with more significant suction pressure.

$$I_c = \frac{LL - WC}{I_\mathrm{p}} \tag{2}$$

Therefore, clay can easily be deformed with a relatively high water content and a low consistent index. If the available water content lies close to the plastic limit, then the clay will not easily deform plastically. The Author then mentioned that a consistency index of 0.75 is an acceptable approximation of the water content at the suction pressure of 10m head of water (outer body of the dike). On that note, the clay content in the core of the dike above the water table is calculated using a clay content with a consistency index of 0.6.

3.4 Erosion Category:

Clay is one of the most used materials in dikes. When impermeable, it can perform well against erosion. However, before the clay is suitable for work into a dike, it must comply with the requirements for water content (consistency index). Complying with these requirements and good compaction can achieve the optimum results for the various construction sections, including permeability, shape retention, and erosion resistance.

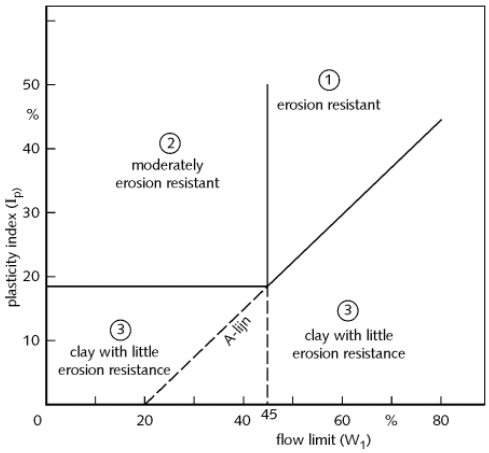
Therefore, clay has multiple erosion categories that it is classified upon in geoengineering, which are:

- Category 1: Erosion resistant.
- Category 2: Moderately erosion resistant.
- Category 3: Little erosion resistance.

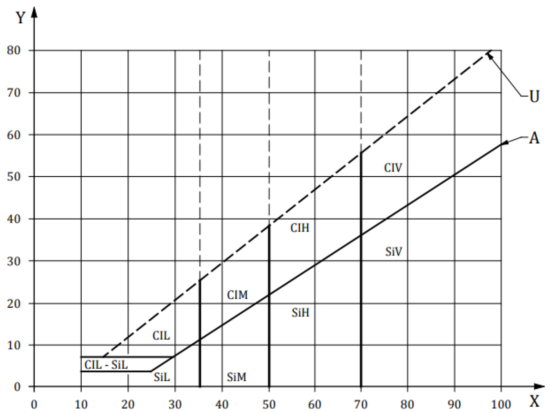
These categories are based on the Atterberg limits and the sand content, Table 7 & Figure 12a

Table 7: Classification of Clay Based on Erosion Resistance

Type of Clay	**Variables**	**Condition**
1. Erosion-resistant clay	LL	> 45
		and
	I_p	$> 0.73 * (LL - 20)$
		and
	sand content	< 40
2. Moderately erosion-resistant clay	LL	< 45
		and
	I_p	> 18
		and
	sand content	< 40
3. Clay with little erosion resistance	LL	$< 0.73 * (LL - 20)$
		and/or
	I_p	< 18
		and/or
	sand content	> 40



(a) Erosion resistance as a plasticity diagram; the A-line is $I_\mathbf{p} = 0.73 \cdot (LL - 20)$ Delft, 1996



(b) Plasticity diagram; the A-line is $I_\mathbf{p} = 0.73 \cdot (LL - 20)$ PovDGG, 2022

Figure 12: Comparison of erosion resistance and plasticity diagrams.

This plasticity diagram also offers a new way to classify the clay according to its liquid limit and plasticity index. This categorization method classifies the clay into the following categories: Table

8, which shows how to read the names of the clay classes. CI refers to clay, while SI refers to silt; furthermore, the L, M, H, and V refer to plasticity levels from light to very high, respectively, while O refers to organic. Therefore, class CIL would read as clay with low plasticity.

These categories are to be compared with the old ones to see what differences could occur between the two. One immediately apparent difference is that this classification is based on clay and silt, while the other (old one) is based on clay, silt, and sand. On that point, it is essential to mention the **U-line** (upper line in the graph), an empirical upper limit for natural soils calculated as $0.9 \cdot (LL - 8)$. Therefore, soil samples above that line are scarce, making this line serve as a quality check for the test results. On the other hand, the **A-line** helps distinguish between clay and silt.

Table 8: Plasticity diagram classifications, PovDGG, 2022

Ground type		Plasticity	
CI	Clay	L	Light
SI	Silt	M	Medium
		H	High
		V	Very high
		O	Organic

3.5 Macro Stability calculations:

There are many ways to perform a stability calculation for macro stability. However, D-stability is the most used software for this failure mechanism. While the manual of Van der Meij, 2020 has an extensive explanation of how this is done. This subsection focuses on the aspects that the research directly influences.

To have macro stability calculations, the geometry of the dike (a dike design) needs to be inputted accurately. Then, the different layers of the dike, such as clay and sand, are inputted. At that point, the soil characteristics must be inputted to help define the dike profile. These characteristics are the leading player in determining whether the dike is safe or not, next to the geometry, of course, as parts like a dike's berm also directly influence the design's safety.

According to WBI, 2021, three main variables for clay need to be inserted in the model: the dry unit weight, friction angle, and cohesion. However, it is also mentioned that cohesion is always considered equal to 0, and therefore, it is unimportant.

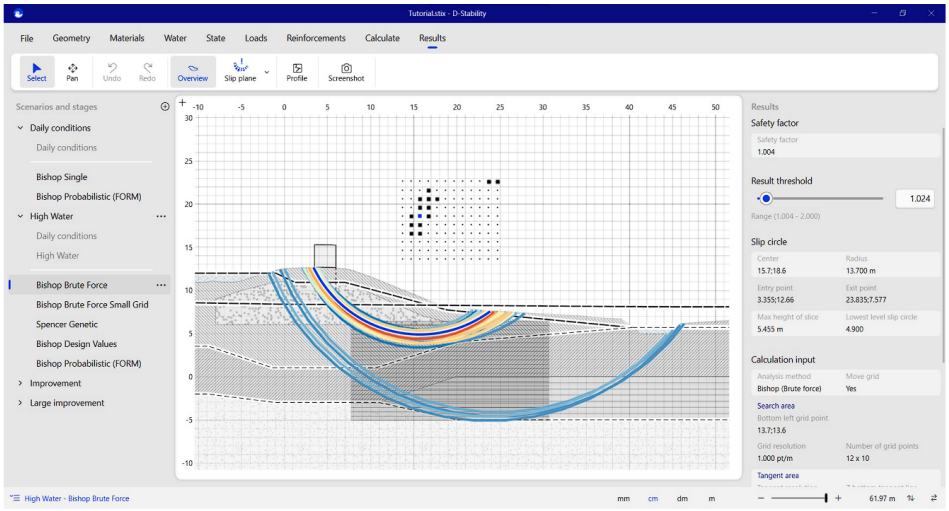


Figure 13: Stability calculation in D stability Van der Meij, 2020

Then, the stability of the dike can be calculated using defined circle areas (which is determined using multiple options in D-stability; for more information, check the manual Van der Meij, 2020). This

calculation can be done either deterministically or probabilistically, which will eventually provide a number indicating whether the design is safe, Figure 13.

The calculations use the unit weight and friction angle, so the research is interested in predicting these two variables.

3.6 Summary: Geotechnical behavior

- Atterberg limits define the physical states of soil: liquid limit, plastic limit, and plasticity index.
- Clay has three different minerals: Kaolinite, Illite, and Montmorillonite.
- Consistency index estimates the maximum amount of water the clay can retain, calculated using Atterberg limits and Water content.
- Clays are categorized into three erosion categories, plotted in a plasticity diagram, offering a new classification system in the so-called plasticity diagram.
- macro stability is calculated using D-stability software's dry unit weight and friction angle.

4 State of the art: literature-based information

The following section explores the state of the art of literature-based correlation using regression and what correlations other researchers found using traditional regression methods. It then shows which Machine learning methods researchers used and which models outperformed others, which was essential to selecting the machine learning models for this research.

4.1 Regression: Literature-based correlations

As the research is interested in dike stability, the friction angle and the unit weight are the most critical variables that require expensive, time-consuming testing. On that note, according to PovDGG, 2022, a visual inspection is one of the first tests performed on projects and mentions that it is one of the most accessible and straightforward tests to perform. In that sense, the following section explores the current correlations found in the literature between the Atterberg limits, clay contents, friction angle, and unit weight.

High Liquid Limit (LL) and Plasticity: According to Lubking, 2000, a high LL and plasticity indicate a relatively high percentage of clay content (CC). Furthermore, the author suggests a positive correlation between the dry density and the water content. Moreover, the author suggests that it is possible to have the max proctor density and the optimum water content to be predicted using the liquid limit.

Plastic properties and Moisture content: Highway Research Board, 1962 suggest that an increase in the plastic properties of the soils is accompanied by an increase in the optimum moisture content and a decrease in the maximum dry density.
Furthermore, they show a positive correlation between the LL and the optimum moisture content (W_{opt}). However, this correlation depends on the Plasticity index (PI) value, where different PI intervals lead to different linear correlations; a higher PI value leads to a higher value for both the LL and the (W_{opt}).
The author also indicated a good linear fit between the (W_{opt}) and the LL and the plasticity limit (PL). However, the correlation is weaker with PI. On that note, the maximum dry density shows a good curved correlation with the PL but a weaker curved correlation with both LL and the PI.

Atterberg limits correlations : The author also mentioned some R-squared correlation values between the Atterberg limits. The PL-LL = 0.84, PI-LL = 0.95, and PI-PL = 0.62 were similar on the logarithmic scale.

Friction angle and Plasticity index: Budhu, 2011 mentions an empirical equation applicable for remodeled clays relevant to the dike stability context. This equation is:

$$\phi'_{\text{cs}} = \sin^{-1}(0.35 - 0.1 * \ln \frac{PI}{100}) \tag{3}$$

Shimobe and Spagnoli, 2021 also mentioned a differently found correlation between the friction angle and the plasticity index. It mentioned a general negative correlation between the two.
Furthermore, the paper noted that the strength variables of clay depend on the **type of clay, water content**, disturbance factor, porosity, and stress history. In this case, only the kind of clay and water content are relevant.
Furthermore, the paper mentioned using PI to predict the effective friction angle and LL to predict the undrained shear strength. In the case of the PI-ϕ correlation, the correlation is negatively curved, based on clays from different parts of the world, Ahmed, 2018. This reduction relationship was also reported by Ameratunga et al., 2016 between the PI and peak effective friction angle (ϕ'_{peak}). This author also reported that the best (mean) equation to estimate the friction angle could be taken as:

$$\phi'_{\text{peak}} = 43 - 10 \log PI \tag{4}$$

The author also argues that there is a positive trend line derived between $\sin \phi' - \frac{PI}{LL}$. This relationship was proposed by Mayne in 1980. Furthermore, a reasonably good positive trend is mentioned between the $\frac{S_u}{\sigma'_v} - \phi'$, where S_u is the shear strength.

$$\sin\phi' = 0.247 + 0.409 * \frac{PI}{LL} \tag{5}$$

Akayuli et al., 2013 mentioned some correlations based on sandy clay in Ghana; however, it is worth noting that the study was based on 10 clay samples. The author argues that the plasticity index of the soil increases linearly with the amount of clay occurring in the soil. Furthermore, the author found a correlation between the friction angle and clay content with an R-squared value of 0.90. This linear correlation is:

$$\phi = -0.743 \cdot CC + 36 \tag{6}$$

The author also mentions the following relationship between the friction angle and the plasticity index:

$$\sin\phi'_{\text{cv}} = 0.8 - 0.094\ln(PI) \tag{7}$$

Atterberg limits and Clay contents: Even though it was challenging to find studies that show sufficient or insufficient correlations between the clay composition and the Atterberg limits. Two studies could discuss some of the results that reflected upon this correlation. Author Polidorli, 2007 surveyed inorganic soils and found a linear correlation between liquid limit and clay content, which is expressed in the following equation:

$$LL = k_1 \cdot CC \tag{8}$$

However, the k_1 did have a wide range of values that depended on several factors that this research does not consider, such as (type of minerals and degree of crystallinity. etc.) The combinations of such factors provided a range between 0.67 and 4.86, as shown in Figure 14.

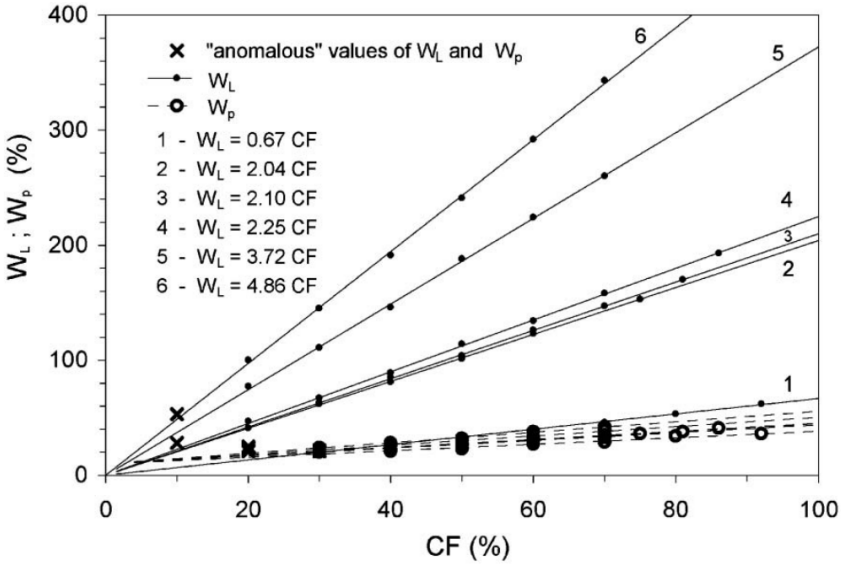


Figure 14: LL and CC correlation Polidorli, 2007, where CF is CC, W_L is LL, and $W - P$ is PL

Furthermore, the author also mentioned a similar correlation between the plastic limit and clay content, as in the previous figure. However, the plastic limit shows less linearity and relatively less variation [0.28, 0.45], which is expressed in the following equation:

$$PL = k_2 \cdot CC + 10 \tag{9}$$

On that same note, the author argues that for inorganic clays, if the values of the liquid limit and the clay fraction are known and are linearly proportional, then the plasticity index could be calculated using the following equation:

$$I_p = 0.96 \cdot LL - (0.26 \cdot CC + 10) \tag{10}$$

Hence, the author concludes that for a given inorganic material with a known liquid limit and clay content, then the values of LL, and I_p for any other Clay fraction could be estimated using Equation 8 and Equation 10 respectively. On the other hand, if the values of the Atterberg limits are known, then it is possible to calculate the clay content using one of the following equations:

$$CC = [(0.96 \cdot LL - IP) - 10]/0.26 \tag{11}$$

$$CC = [(PL - 0.04 \cdot LL) - 10]/0.26 \tag{12}$$

Maximum dry density, Plastic limit, Optimum water content, and dry unit weight: According to Gurtug, 2015, there is a correlation between the minimum dry density and optimum water content for different soils collected from the literature with three different energy levels, which are applied through various versions of proctor test. The author shows that there is a correlation between the Plastic limit and optimum water content according to the following equation:

$$OMC = 0.94 \cdot PL, \qquad R^2 = 0.98 \tag{13}$$

Furthermore, the author also shows a correlation between the maximum dry density and optimum water content, according to the following equation:

$$\rho_{\mathrm{D_{max}}} = 33.85 \cdot \log(\mathrm{OMC}), \qquad R^2 = 0.99 \tag{14}$$

This allows the dry unit weight to be calculated using the following equation:

$$\gamma_{\mathbf{D_{max}}} = \frac{\rho_{\mathrm{D_{max}}} \cdot g}{1000} \tag{15}$$

Therefore, the author concludes that the previous two equations can be used to predict the maximum dry density and optimum water content. This can calculate the max dry unit weight according to Equation 27.

Furthermore, author Nagaraj et al., 2015 also mentioned a correction between OMC and PI, which was applied to 42 different natural soil types.

$$OMC = 0.76 \cdot PI \tag{16}$$

$$\gamma_{\mathrm{d\_max}} = 20.62 - 0.19 \cdot \mathrm{PI} \tag{17}$$

The author also mentioned that the plasticity index was modified using the following equation to overcome the effects of coarse fraction in > 425 macrometers. Two other studies also mentioned these modifications, Srinivasa Murthy et al. (1987) and Pandian et al. (1997).

$$PI_{\mathrm{m}} = PI \cdot (1 - \frac{CC}{100}) \tag{18}$$

$$OMC = 0.82 \cdot PI_{\mathrm{m}} \tag{19}$$

This also leads to similar modifications to the OMC according to the following equation:

$$OMC_{\mathrm{m}} = OMC \cdot (1 - \frac{CC}{100}) \tag{20}$$

These modifications lead to the following correlations with the max dry unit weight:

$$\gamma_{\mathrm{d\_max}} = 20.35 - 0.17 \cdot PI_{\mathrm{m}} \tag{21}$$

$$\gamma_{\mathrm{d\_max}} = 20.7 - 0.22 * OMC \tag{22}$$

On the other hand, Ali et al., 2019, who looked into multiple literatures, argues that many papers tried to establish a correlation between the liquid and plastic limits with the optimum water content and maximum dry density. However, the author concludes that based on 27 samples from Koya City, neither shows an adequate correlation between OMC and maximum dry density. The author also adds that the previously established correlation between the soil indices and compaction characteristics asserts that the plastic limit provides a better relationship than the liquid limit. However, his research did not provide enough evidence for that. It is essential to mention that this author did not cite Gurtug, 2015 in this study, even though it is an older study with a significantly different outcome.

Liquid limit and dry unit weight In the soil mechanic book of Budhu, 2011, an empirical correlation is mentioned between the liquid limit and the void ratio, which can be used to calculate the dry unit weight.

$$e = LL \cdot G_s \tag{23}$$

$$\gamma_d = \frac{G_s \cdot \gamma_w}{1 + e} \tag{24}$$

Where:

Table 9: Dry unit weight symbol description

Symbol	Description
γ_d	Dry unit weight [kN/m3]
ρ_{dp}	Proctor dry density [kg/m3]
G_s	Specific soil gravity for the soil particles = 2.7
g	9.81 [m/s2]

correction empirical equation: The empirical equation is used for different types of clay. Therefore, since the inspection data is predominantly of inorganic clays and Dutch clay samples, a check has been done to see if this empirical equation predicts the void ratios well. On that point, some reports were found with the same clay sample's compaction and inspection test. Therefore, it was possible to calculate the void ratio using the in situ dry density and the liquid limit using the following equation:

$$e = G_{\mathrm{s}} \cdot \frac{\gamma_{\mathrm{w}}}{\gamma_{\mathrm{d}}} - 1 \tag{25}$$

This equation showed that the empirical equation did overestimate the void ratio with an average value of 0.15. On that point, it was decided to add -0.15 to the equation as a correction, which will be used in equation 24. The difference is also shown in Figure 15. The resulting dry unit weight histogram had a similar interval as the calculated dry unit weight calculated from the compaction data, showing that this corrected empirical equation is acceptable for creating a shared variable between the inspection and the compaction test data. Therefore, Equation 26 represents the last equation used for the void ratio calculation.

$$e = LL \cdot G_{\mathbf{s}} - 0.15 \tag{26}$$

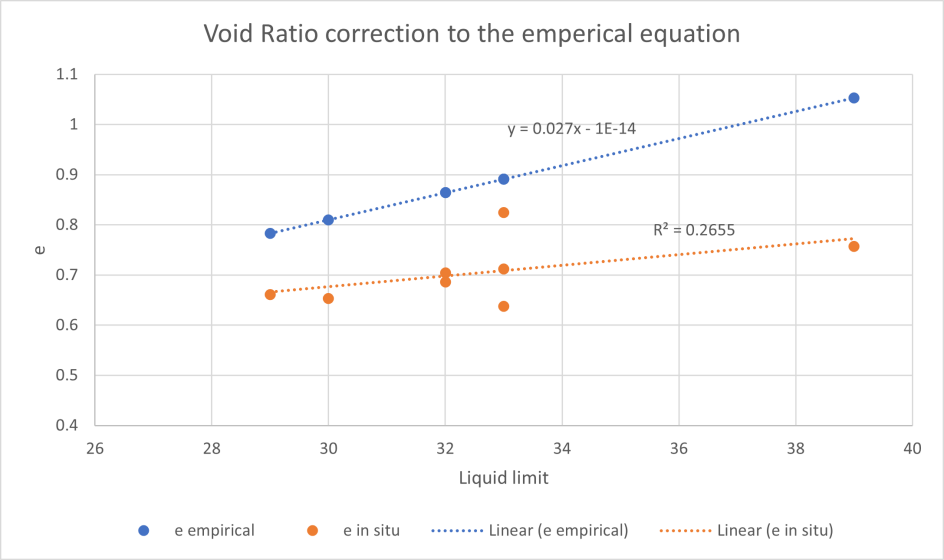


Figure 15: Correction void ratio

Influence of Clay minerals: according to Raad, 2021 montmorillonite could significantly reduce friction angle and increase liquid & plastic limit. In this case, this could be related to soil with high clay fraction and high activity. This influence was based on expansive soil in the Bhavnagar region of India.

4.2 Machine learning models in Geo engineering

Pentoś et al., 2022 claims that neural network models (ANN) outperformed support vector machines and various linear regression techniques. Neural networks and support vector machines produced more accurate results than multiple linear regression predicting soil compaction and shear stress using electrical variables. Also, Jasim et al., 2019 highlighted the success of using ANN to predict a correlation between the liquid limit and soil plasticity index.

Chala and Ray, 2023 stated that their investigation revealed that most ML models correctly classified soils. The models with the highest accuracy were the Support Vector Machine (SVM) and Artificial Neural Networks (ANN), which both achieved slightly lower accuracy. High accuracy was also attained by the decision tree (DT) and random forest (RF) models. They also conclude that software applications can incorporate the SVM and RF algorithms for quick and precise soil classification.

Arama et al., 2021 used artificial neural networks (ANN) and regression analysis to extract the plasticity index from the liquid limit; ANN proved more effective than conventional techniques. Integrating the soil's fine content and depth into the network produced similar outcomes. This study found that while all models demonstrate a relevant linear relationship, using artificial neural network techniques yields more accurate approximations than regression models.

Katuwal et al., 2020 discovered that there is a positive correlation between soil bulk density (BD) and soil organic carbon (SOC). They developed and compared several models based on SOC and other soil variables to predict BD. The study's findings demonstrated that SOC was both one of the variables in the multi-variable models and a significant single predictor for predicting soil BD. To predict soil BD, the study used three machine learning techniques: regression rules (RR), random forest (RF), and multiple linear regression (MLR). The study's findings demonstrated that out of the machine learning models put to the test, RF offered the most accurate prediction of soil bulk density (BD).

4.3 Summery: literature-based correlations:

- Positive correlations exist between dry density and water content, as well as liquid limit and optimum moisture content (depending on PI).
- Negative correlations exist between friction angle and plasticity index.
- Empirical equations have been proposed for predicting friction angle based on plasticity index and other factors.
- A linear correlation between friction angle and clay content has been observed in sandy clays.
- The Atterberg limits are linearly proportional to clay content. The values of two of the three variables (LL, LP, CC) that are measurable using standard tests suffice to obtain the values of the other three variables of the test soil.
- Both optimum water content and maximum dry density might be calculated knowing the plasticity index of the soil. However, other research disagrees.
- A correlation to obtain dry unit weight was suggested by estimating the void ratio based on the liquid limit. This equation was then corrected to match the Dutch clays.
- A correlation between liquid limit and dry unit weight was found and had to be corrected by subtracting 0.15.
- negative correlation was mentioned between dry unit weight and plasticity index, which is also modified to account for higher sand friction in some clay samples.
- positive correlation was reported between optimum water content and plastic limit, which could be used to estimate the maximum dry density. This could then be used to calculate the max dry unit weight.

5 Data pre-prosessing

Rasearch Question 1: What is the state of the data?
Data preparation is essential to statistical studies, also known as data pre-processing. This step transforms raw data into a format that can be run through machine learning algorithms, enabling them to make predictions. This is essential to ensure that datasets can provide meaningful insight when inputted into ML algorithms.

5.1 Data collection:

Clay variables were collected from different file formats. This was done by extracting the data from Excel and other PDF files using Python libraries like PyPDF2, Camelot, and PDFQuery. The codes used to collect the data are added to GitHub.

The research collected the contents of the clay (clay, sand, and silt), Atterberg limits (liquid limit, plastic limit, and plasticity index), and the classification class of the clay from the inspection test. On the other hand, water content and dry density were taken from the clay compaction test. Lastly, the triaxial tests data gathered the dry unit weight, class of the clay, deviatoric stress, and water content. More on the quantity of the data and quality is mentioned in Section 6.1

5.2 Missing data:

Missing data could appear in any data set; therefore, checking if it greatly affects the data set is important. Generally, it is either imputation or removing the data.
the imputation method substitutes reasonable guesses for missing data. This method is mostly useful when the percentage of missing data is low. Removing data, on the other hand, is used when dealing with missing or random data; in that case, the entire data point could be removed to help reduce bias, DataRobot, 2024. Furthermore, an average value is used when the percentage of the missing data is low.

5.2.1 Removing outliers:

An outlier is a data point at an abnormal distance from other data points in the dataset. Therefore, they influence the descriptive statistics of the data group. Not removing an outlier would make the statistics less reliable in this framework.

According to Dhadse, 2024, removing the outliers could be done using 2 different methods:(1) Using Interquartile Ranges (IQR) and (2) Using standard deviation.

According to Dhadse, 2024, the IQR method is best suited for the skewed dataset. Otherwise, using standard deviation to detect outliers is better when dealing with normally distributed data.

5.3 Calculated Data variables:

After cleaning the data, some variables were derived using the original dataset, which is essential to the research. These variables are mentioned below.

5.3.1 Dry unit weight - Clay compaction test:

Dry unit weight is one of the most important variables estimated and used in many probabilistic analysis applications. The provided clay compaction variables made calculating the dry unit weight possible. the following equation was taken from Budhu, 2011

$$\gamma_d = \frac{\rho_{dp} * g}{1000} \tag{27}$$

5.3.2 Triangular classification of clay (NEN 5014) - Clay inspection data

The triangular classification was done by first plotting the triangle from 5a in Python and then plotting the x and y points according to the following equations: Table 10, which were taken from a Fugro excel sheet that was used to determine the classes of clay samples.

These equations depend on clay and sand content, which differs from the reported clay, sand, and silt contents in the clay inspection tests. The main difference is that the reported percentages in the clay inspection tests include organic, mass loss, and salt contents. Therefore, these ratios had to be recalculated for only clay, sand, and silt content, making these 3 variables form 100% of the clay sample. Appendix G shows how these variables were calculated.

Table 10: Equations for the clay classes coordinates in the Triangular classification system (NEN 5014)

	point coordinates
x	$(100 - \text{sand content}) - \text{clay content} * \cos(\frac{1}{3} * \pi)$
y	$\sin(\frac{1}{3} * \pi)$

Once the points were plotted, they were distributed within the triangular diagram and divided into distinct polygons. Each polygon represents a different clay class within the triangular classification system. A code was developed to assign each plotted point to the appropriate polygon, determining each instance's clay class; these polygons are shown in Figure 20, and the code is available on GitHub.

5.3.3 Friction angle (after peak): Triaxial test

As horizontal and effective stress are provided at a 25% strain, the peak friction angle was calculated using Equation 28. In this context, the (') sign denotes stresses in their effective state, which is the difference between the total stress and the pore water pressure. It represents the stress carried by the soil's solid particles, which determines its strength and deformation behavior.

$$\phi = \arcsin\left(\frac{\sigma_1' - \sigma_3'}{\sigma_3' + \sigma_1'}\right) \tag{28}$$

where:

Table 11: friction angle's equation symbol description

Symbol	Description
σ_1'	vertical effective stress [kPa]
σ_3'	horizontal effective stress [kPa]

5.3.4 Water content - Clay inspection data:

To apply Machine learning, a common variable is needed to connect the different data frames. Furthermore, as mentioned in Section 3, a consistency index of 0.75 is used in the body of the dike, and an index of 0.6 is used in the core of the dike.

Hence, three water content variables were determined based on three different consistency indices using Equation 2. The first two indices were 0.6 and 0.75, while the third index was 0.85. The third variable was assumed to correspond to the water content of the clay compaction data because the other two indices resulted in higher water content values than those reported in the clay compaction data. The statistics of these variables are presented in Table 14 and Table 15. These three measures were calculated using the following equations:

$$WC_{0.75} = LL - 0.75 \cdot I_\text{p} \tag{29}$$

$$WC_{0.60} = LL - 0.60 \cdot I_{\mathrm{p}} \tag{30}$$

$$WC_{0.85} = LL - 0.85 \cdot I_{\mathrm{p}} \tag{31}$$

5.4 Regression model & Results' metrics:

As explained before, this research explored the differences between regression and Machine learning models. While the used ML models are described in Section 7, this section explains the linear regression model, followed by three metrics usually used to assess machine learning results: the R-squared, the Mean Squared Error (MSE), and the Mean Absolute Error (MAE).

5.4.1 Linear regression model:

According to Kanade, 2023, Linear regression is a model that provides a linear relationship between two independent and dependent variables. The dependent is used to predict the other. This model searches for the best fit between the data points using a linear regression line, as shown in Figure **??**.

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{32}$$

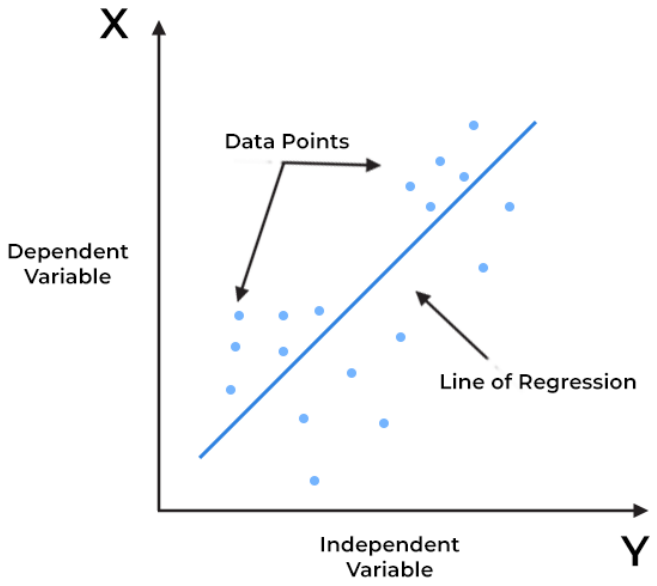


Figure 16: Regression model Kanade, 2023

5.4.2 R-squared method:

According to G., 2024, R-squared is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures the strength of the prediction of the model in comparison with the original dataset. The values typically range between 0-1, with 0 indicating a weak performance of the model and 1 indicating a strong performance. R-squared is calculated using the following equation:

$$R^2 = 1 - \frac{SS_{\mathbf{res}}}{SS_{\mathbf{tot}}} \tag{33}$$

$SS_{\mathbf{res}}$ is the residual sum of squares, and $SS_{\mathbf{tot}}$ is the total sum of squares.The total sum of squares is calculated by summing the squares of the perpendicular distances between the data points and the mean line. The residual sum of squares is calculated by summing the squares of the perpendicular distances between the data points and the best-fitting line, G., 2024.

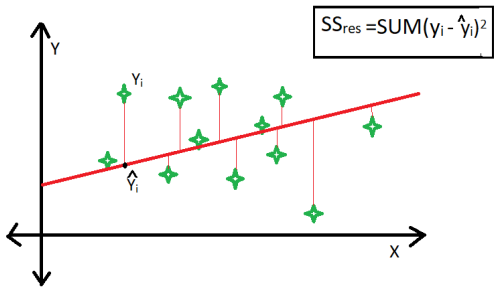


Figure 17: SS_{res} G., 2024

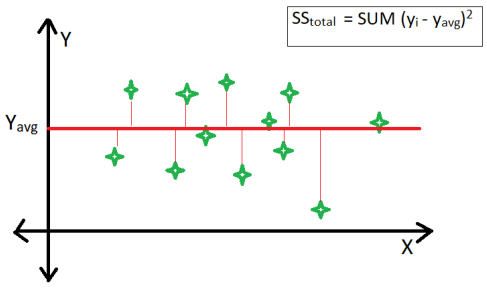


Figure 18: SS_{total} G., 2024

5.4.3 Mean Square Error ¨MSE¨

According to Gupta, 2024, the Mean Squared Error (MSE) shows how well a regression line fits a set of data points. Think of it as a way to measure the "average mistake" the line makes when trying to match the data. The MSE is calculated by taking the errors (the difference between the predicted values and the actual data points), squaring them (to eliminate negative signs and emphasize more significant mistakes), and then finding the average of those squared errors. In simple terms, MSE helps us see how far off our predictions are, on average, from what they should be. Figure **??** shows the error around a regression line.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{34}$$

A higher MSE means the data points are scattered far from the mean, showing that the predictions are often off by a large margin. On the other hand, a lower MSE suggests that the data points are clustered closely around the mean, meaning the predictions are more accurate. We aim for a lower MSE because it shows that the predictions are generally close to the actual data.

By squaring the errors, MSE penalizes more significant errors more heavily than smaller ones. The squaring process can make MSE less intuitive to interpret because the errors are no longer in the data's original scale. Therefore, MSE is more sensitive to outliers since the squaring amplifies more significant errors disproportionately.

5.4.4 Mean Absolute Error "MAE"

According to DeepChecks, 2024, Mean Absolute Error (MAE) measures the average size of errors in a set of predictions, regardless of whether the predictions were too high or too low. It calculates this by taking the average of the absolute differences between the predicted and actual values, and it's commonly used to evaluate how well a regression model performs.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{35}$$

MAE treats all errors equally, meaning it gives a straightforward measure of how far off, on average, the predictions are from the actual values. Since it uses absolute values, MAE is more straightforward to interpret and directly relates to the scale of the data.

5.5 Summery Assessment of results:

- Linear regression model is compared with the machine learning models.
- R-squared: Measures the goodness-of-fit for linear regression models by indicating the percentage of variance in the dependent variable explained by the independent variables, with values ranging from 0 (weak) to 1 (strong).

- Mean Squared Error (MSE): Assesses the average squared differences between predicted and actual values, penalizing more significant errors more heavily and is sensitive to outliers, with lower values indicating better model performance.
- Mean Absolute Error (MAE): Calculates the average absolute difference between predicted and actual values, providing a straightforward measure of prediction accuracy, with all errors treated equally and directly relating to the data's scale.

6 Results: Data variables, classifications & statistics

The following section shows the found data variables and their classifications according to the triangular and plasticity diagram classifications. Then, it individually represents the internal linear correlations of the data frames, along with the statistics of the data variables. This section resulted from multiple data collection steps explained in Appendix G

6.1 Data variables:

Figure 19 shows an overview of the data sets with the different variables. The figure illustrates that water content is a shared variable between the data frames (clay inspection, compaction, and triaxial tests). Furthermore, it shows that the contents of the clay in the inspection tests are derived variables (clay, sand, and silt), which is a result of the triangular classification system, as explained in Section 5.3.2, Appendix G. The figure then also shows three measures of water content, showing three different features in the data frame; each feature is derived using a different consistency index (0.60, 0.75,0.80), Section 5.3.4.

In contrast, the clay compaction data frame includes dry unit weight as a derived variable, while the triaxial test data frame features the friction angle as a derived variable. Lastly, the class of clay is a shared variable between the inspection and triaxial tests. The clay inspection tests yield two classifications: the triangular classification and the plasticity diagram classes.

Furthermore, Table 12 presents the data count for each data frame. The clay inspection data frame contains the largest dataset, with 4,406 rows, followed by the clay compaction data frame with 2,463 rows. In contrast, the triaxial test data frame comprises only 130 rows. Each row represents a complete set of variables, as Figure 19 shows.

Table 12: Data Samples Count

Test	Data Count
Clay inspection	4406
Clay compaction	2463
Triaxial Data	130

Classifications of clay: according to the x and y variables calculated in Table 10. The clay was classified according to the name of the polygon that contained the point. Figure 20 shows the position of the points in the triangle. On that note, the points mostly exist in the same interval as Figure 5b shows, which indicates the common clay classes in the Netherlands. Furthermore, the red points were removed as they don't represent clay classes. Lastly, Table 13 shows the count of the classes, which indicates that most of the points are classified as Ks3 and Ks4. On the other hand, the count of Kz2, Ks1, Kz1, and Kz3 is below 200 points. Furthermore, the clay Triaxial test (CT) also had similar classifications; the same table shows several of these tables. It is noticeable here that classifications Ks4 and Kz2 are missing from the dataset.

Table 13: clay classes count

Clay class	Count per class in CI	Count per class in CT
Ks3	1626	59
Ks4	1515	0
Ks2	734	35
Kz2	175	0
Ks1	157	17
Kz1	130	11
Kz3	51	8

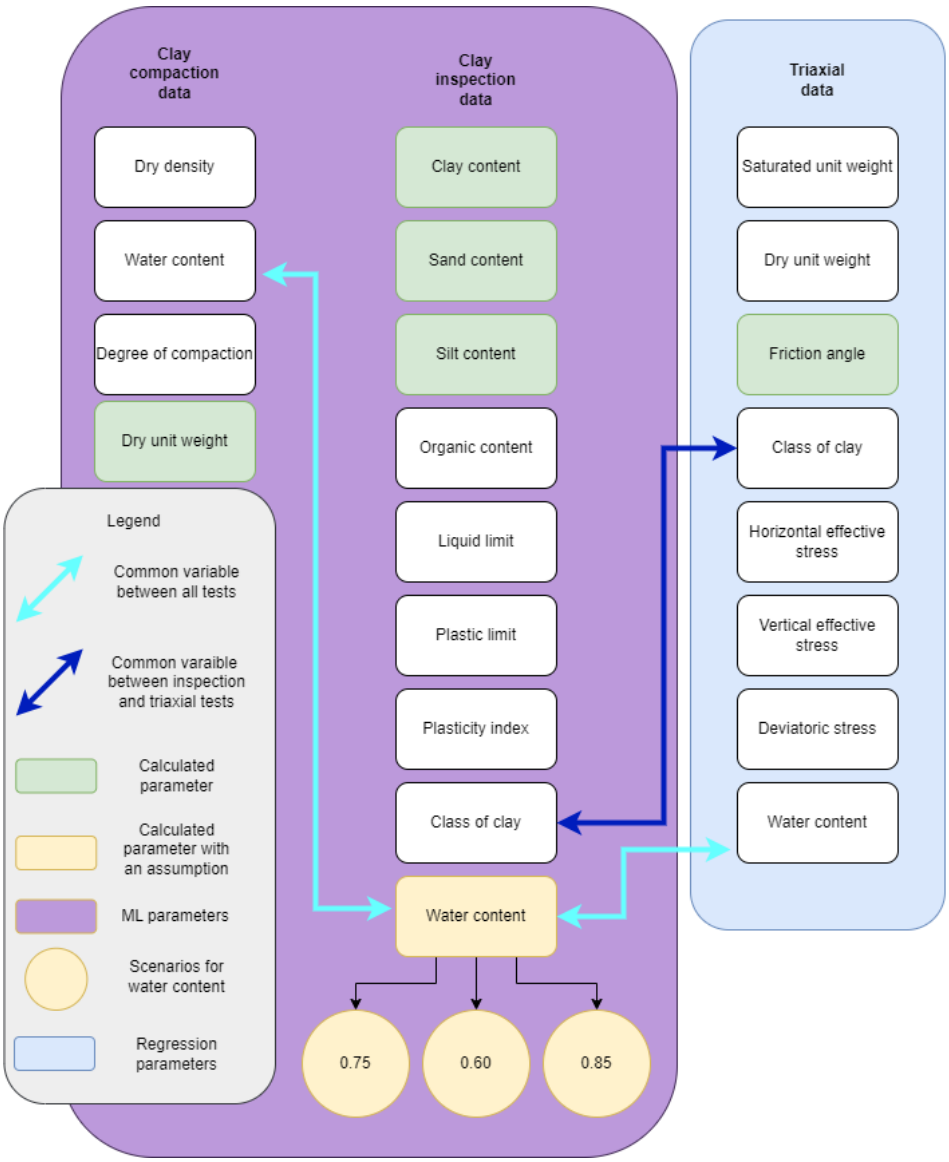


Figure 19: test variables

6.2 Comparision to the plasticity diagram classification system:

Figure 21 shows a difference between the old and the new classification systems. While the old one (triangular) depends on the clay contents, the new one (plasticity graph) depends on the Atterberg limits and categorizes the soil into clay and silt with different plasticity levels, making the two classifications based on different physics.

The main idea is that the classes of old classifications (triangular) spread across various classes in the plasticity diagram classes. Appendix E, shows the individual points of the classes, where it is clear that class Ks4 is spread across CIH, CIM, and CIL classes. Furthermore, classes Ks3 and Ks2 spread between CIV, CIH, and CIM, while class Ks1 spreads across CIV and CIH. The sandy classes Kz3, Kz2, and Kz1 are spread between CIM and CIL. Furthermore, all of the clay sample points exist under the U-line, indicating no rare samples in the used database.

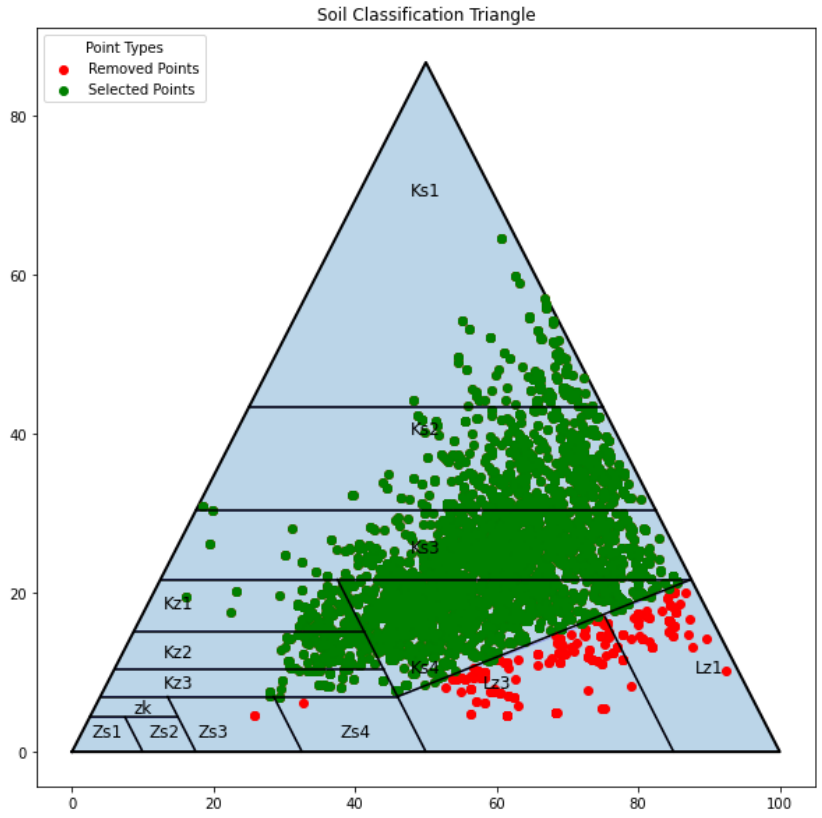


Figure 20: Clay classifications (follows the common range of NL classes as shown in Figure 5b)

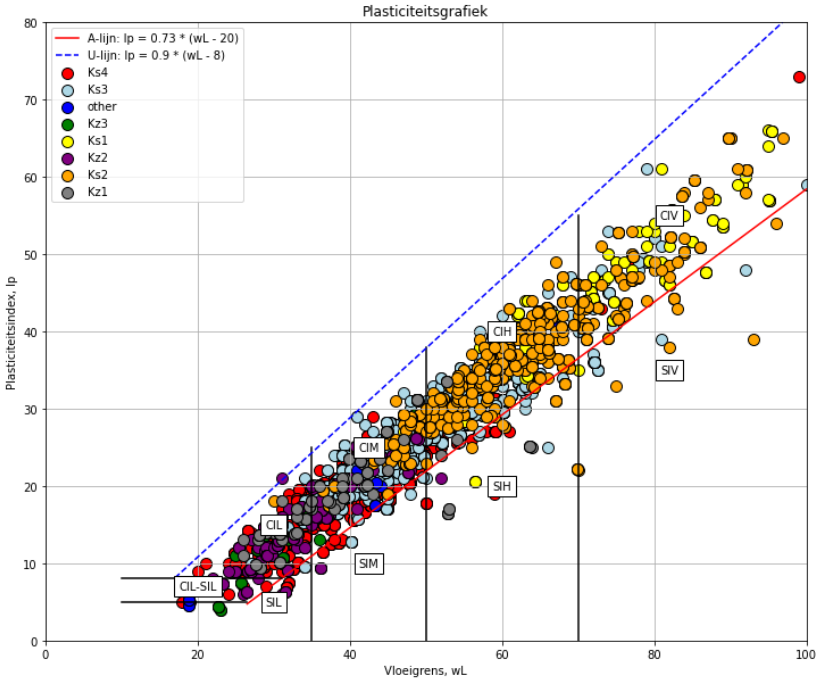


Figure 21: Plasticity diagram classifications

6.3 Descriptive statistics:

This section presents the fitted distribution, mean, median, max, min, and standard deviation values for all the variables per test. Furthermore, the linear correlations between the variables per test and per clay class (for the inspection and triaxial test data) are shown. Appendix B shows the histograms and the heatmaps for the mentioned statistics.

Clay inspection test: Table 14 shows the descriptive statistics of the clay inspection variables. These tests were only done on clay that was suitable for dikes based on the visual inspection. Therefore, the clay samples do not have a high organic content (the requirement is lower than 5%). This is also apparent in the mean organic contents value. Furthermore, there has been a mass loss of the samples. However, the clay, sand, and silt contents are calculated independently. Therefore, the portions represented in the mass loss and organic content are not a part of the represented clay contents. Lastly, the water content (0.6) is slightly more significant than the water content (0.75) values, which is logical as the 0.6 consistency index is used in the core of the dike.

Table 14: Statistics of clay inspection data

Variable	Distribution	Mean	Median	Max	Min	Std
Clay content [%]	Gamma	27.93	26.67	74.43	8.01	9.91
Sand content [%]	Gamma	27.38	26.02	72.65	0.43	15.08
Silt content [%]	Gamma	44.69	44.30	73.28	0.66	10.66
Organic content [%]	Gamma	2.68	2.52	6.28	0.01	1.38
Mass loss [%]	Gamma	8.78	32.70	32.70	0.09	4.17
Liquid limit [%]	Log normal	47.84	45.61	100.00	18.00	13.07
Plastic limit [%]	Log normal	22.29	21.72	34.09	10.00	4.42
Plasticity index [%]	Log normal	25.47	24.00	73.00	4.00	9.89
Water content (0.75) "calculated" [%]	Log normal	28.74	28.00	63.75	13.25	6.39
Water content (0.60) "calculated" [%]	Log normal	32.56	31.58	69.60	14.60	7.62
Water content (0.85) "calculated" [%]	Log normal	26.19	25.53	59.85	12.35	5.65

Clay compaction test: For the clay compaction test (CC), the tests represent clay that was carefully compacted in dikes. Therefore, as the table shows, all the samples have high compaction. Furthermore, (Figure 68 in Appendix B, shows the compaction curve with the air contents of the compaction test data. This figure shows that most data points lie between 0 and 16% air content. On that point, only a couple of samples had compaction of lower than 97, which were deleted from the data frame.

Table 15: Statistics of clay compaction data

Variable	Distribution	Mean	Median	Max	Min	Std
Dry density [kg/m3]	Gamma	1479.95	1457.00	1909.00	1111.00	123.24
Proctor dry density [kg/m3]	Gamma	1859.80	1847.62	2164.62	1548.81	90.57
Water content [%]	Gamma	26.00	27.00	55.00	7.00	5.55
Degree of compaction [-]	Gamma	99.00	100.00	109.00	97.00	2.48
Dry unit weight "calculated" [kN/m3]	Gamma	14.53	14.30	18.73	10.90	1.21
Unit weight "calculated" [kN/m3]	Gamma	18.24	18.13	21.24	15.19	0.89

Triaxial test: Table 16 shows the statistics of the data. However, in Appendix B, it is noticeable that the data quality is bad as the histograms indicate a quite huge spread of the values. This data set is not suitable for either linear or machine-learning applications. This is also apparent in the linear correlations mentioned in Figure 24

Table 16: Statistics of Triaxial data

Variable	Distribution	Mean	Median	Max	Min	Std
Unit weight [kN/m3]	Gamma	16.93	17.50	20.5	12.4	2.21
Dry unit weight [kN/m3]	Gamma	11.59	12.35	17.60	5.00	3.45
Water content [%]	Log normal	54.60	41.05	146.10	8.50	32.46
Friction angle (After-peak) [degrees]	Gamma	36.71	34.71	71.33	9.73	9.76
Deviatoric stress [kPa]	Log normal	106.66	97.90	409.30	10.00	64.93

6.4 Linear correlations:

Research question 2: How are the data correlated Internally

Clay inspection data

Figure 22 shows the heatmap between all the clay inspection variables. This map shows correlations between the contents, which is self-explanatory, as they were calculated using each other. On that note, Table 17 shows some interesting correlations between the Atterberg limits and the contents of the soil. Furthermore, the water content does show some correlation with clay, sand, and organic contents. On the note, Appendix C shows the heatmaps of the different classes separately; the correlations seem to follow the same pattern as the general one. However, the water content does not seem to have good correlations with the clay classes separately, except for the Ks4 class, which shows a correlation of around 0.54.

Table 17: Correlations of clay inspections data

Correlation	Variable 1	Variable 2
0.74	LL	CC
-0.61	LL	SC
0.6	PL	OC
0.78	PL	LL
0.76	PI	CC
0.95	PI	LL
0.63	WC	CC

Figure 22: Heatmap clay inspection data

Clay compaction data: The heatmap in Figure 23 shows that for the clay compaction data, there is a strong correlation between the water content and both the dry and saturated unit weight, with the correlation of the dry unit weight being slightly stronger. a correlation is also found between the liquid limit and the unit weight.

Table 18: Correlations of clay compaction data

Vorrelation	Variable 1	Variable 2
-0.93	WC	γ_d
-0.86	WC	γ

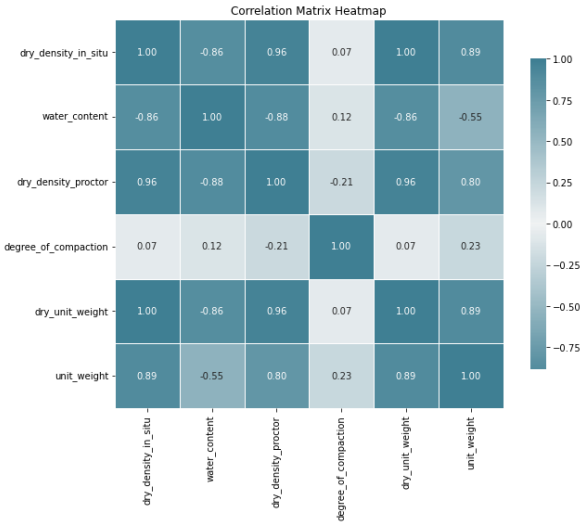


Figure 23: Heatmap Clay comapction

Triaxial tests: Figure 24 shows the heatmap of all the clay classes. This shows a correlation between the water content and dry unit weight. Also, it shows a weak correlation between the friction angle and sigma 3. Lastly, the deviatoric stress seems to have a weak correlation with a dry unit weight of 0.43. The correlations between q and the sigmas are irrelevant, as the deviatoric stress was calculated using these two variables. The other classes do not show any other significant correlations. The rest of the Heatmaps is shown in Appendix B.

Table 19: Correlations of triaxial data

Correlation	Variable 1	Variable 2
-0.93	WC	γ_d
-0.43	ϕ	σ_3
0.43	q	γ_d

Figure 24: Heatmap triaxial test

7 Machine learning:

Research Question 3: To what extent can Machine Learning models predict specific variables?

Unlike traditional methods, machine learning algorithms enhance their performance on specific tasks through experience with data, which is typically structured using conventional methods. This data can be labeled (supervised) or unlabeled (unsupervised), allowing the algorithms to uncover patterns and relationships, which are then used to make predictions or classifications on new unseen data.

After applying traditional regression models to the data, various machine learning models were tested to evaluate their ability to predict the required variables. These implementations were carried out using Python and several machine learning and data analysis libraries, including Numpy, Pandas, TensorFlow, Matplotlib, Scikit-learn, and Seaborn. The subsequent section will provide a detailed overview of the machine-learning models applied in this research.

7.1 Pre-processing the data:

Training machine learning models requires preprocessing the data so that the models can learn the patterns more efficiently. Variables will have different scales, which can influence the behavior and performance of ML models. This thesis used the Standard scalar method, as all the needed variables followed a Gaussian distribution or a similar one. According to Scikit-learn, this standard scaler standardizes the features/variables in a dataset by removing the mean and scaling to the unit variance. In other words, it gives the features a mean of 0 and a standard deviation of 1.

7.2 Tensors in Machine learning:

ML models use a special kind of data structure called Tensors. A tensor is a way of organizing data into a shape that the computer can efficiently work with. Think of them as a generalization of more familiar concepts like numbers, vectors (list of numbers), or matrices (gris of numbers). These tensors define what comes in and come out of an ML model. On that note, since the validation process is used (explained in the following Subsection). The input and output tensors are different. The tensor input can be expressed as [features, number of data points], while the output tensor would be the same with a difference in the number of predicted data points, explained in the following 2 subsections.

7.3 Data splitting for machine learning:

The validation set utilizes a portion of the training data to objectively assess a model's performance. Differing from training and test sets, the validation dataset serves as an intermediary stage aimed at selecting and refining the optimal model. This phase involves crucial activities such as hyperparameter tuning. Detecting and preventing overfitting is a primary objective during validation, as it helps mitigate the risk of errors in future predictions and observations resulting from an analysis overly tailored to a specific dataset, Alexander, 2023. Figure 25 shows where this validation set takes place in making a machine-learning model.

7.3.1 K-fold cross validation:

Because of that, the research uses the K-fold validation technique, a powerful technique for evaluating predictive models in data science, Pandian, 2024. It splits the data into K folds, in this case, five folds. Each set is then used as a validation set in turn, while the remaining k-1 folds (In this case, 4) are used for training. Therefore, the performance is repeated K times until it reaches the split with the best metric results. This process is visualized in Figure 26. This affected the produced predicted data points of the machine learning models, as they were 1/5 the volume of the original data frame. In other words, **this affected the output tensor of the Neural Network**, while the input would be [number of input features, 4406 data points], the output tensor is [number of input features, 882]. The numbers of the input features depend on the tested scenario, which is explained in Section 7.7.

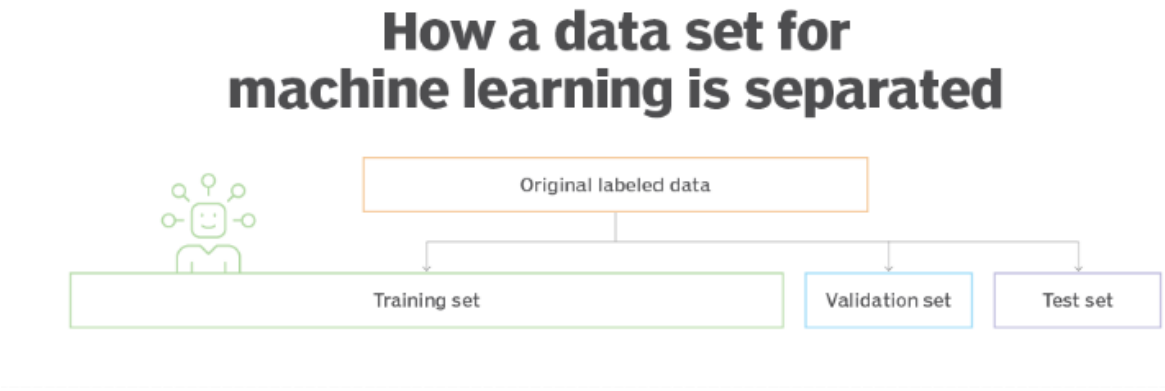


Figure 25: Validation set
Gillis, 2023

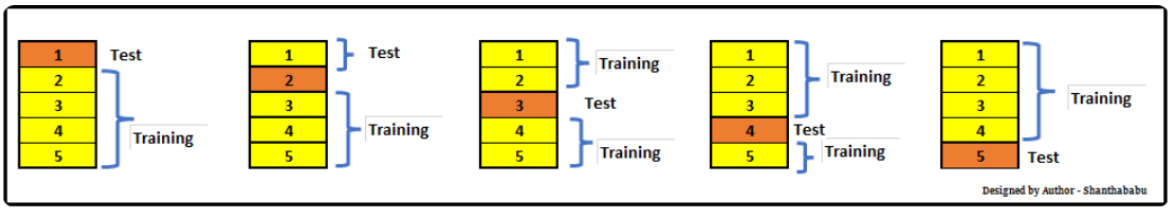


Figure 26: K fold validation
Pandian, 2024

7.4 Machine learning models:

Many machine learning models are available for predicting soil variables. However, literature research shows that Neural Networks, Random Forests, and Support vector machines are the most efficient. This research focuses more on NN and RF, as SVM performed poorly and was, therefore, removed from it.

7.4.1 Neural Networks:

Neural networks, known as artificial networks (ANNs), are a method of applying machine learning and are the heart of deep learning algorithms. These networks comprise a node layer containing an input layer, one or more hidden layers, and an output layer. Each node (an artificial neuron) connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network; otherwise, no data is passed along to the next layer of the network, IBM, 2023. This node works according to the Linear combination function, written as Equation 36, essentially a weighted sum of the inputs with an added bias term. More information on that specifics can be found in Medium: Neural Networks

$$y = \sum_{i=1}^{n} w_i x_i + b \tag{36}$$

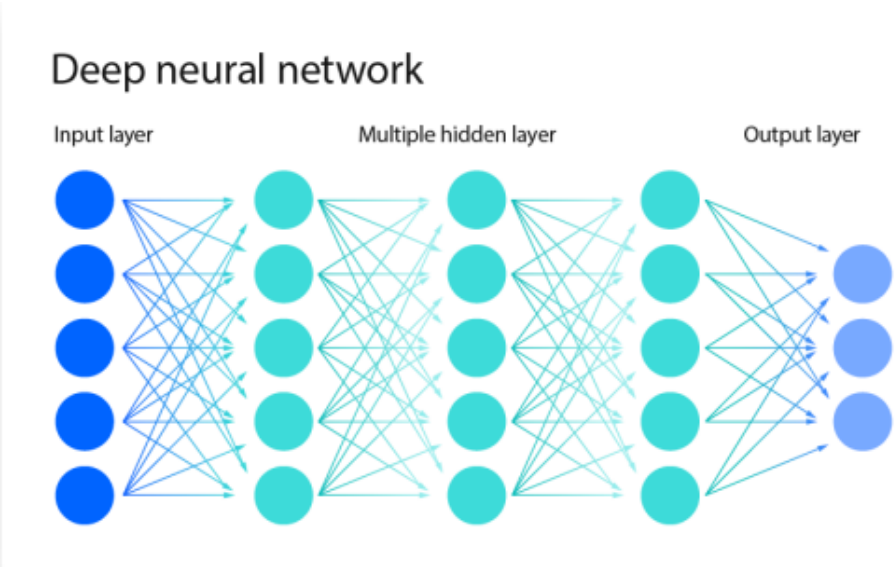


Figure 27: Neural network basic model
IBM, 2023

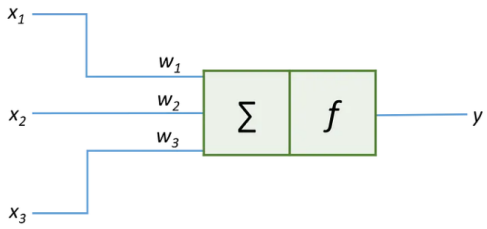


Figure 28: Neuron in a NN, Kumar, 2021

7.4.2 Random Forest:

A random forest combines the output of multiple decision trees to reach a single result, Figure 29. Its ease of use and flexibility have fueled its adoption, as it handles classification and regression problems. The random forest algorithm is an extension of the bagging method as it utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees. Feature randomness generates a random subset of features, which ensures low correlation among decision trees. This is a crucial difference between decision trees and random forests. While decision trees consider all the possible feature splits, random forests only select a subset of those features, IBM, 2024.

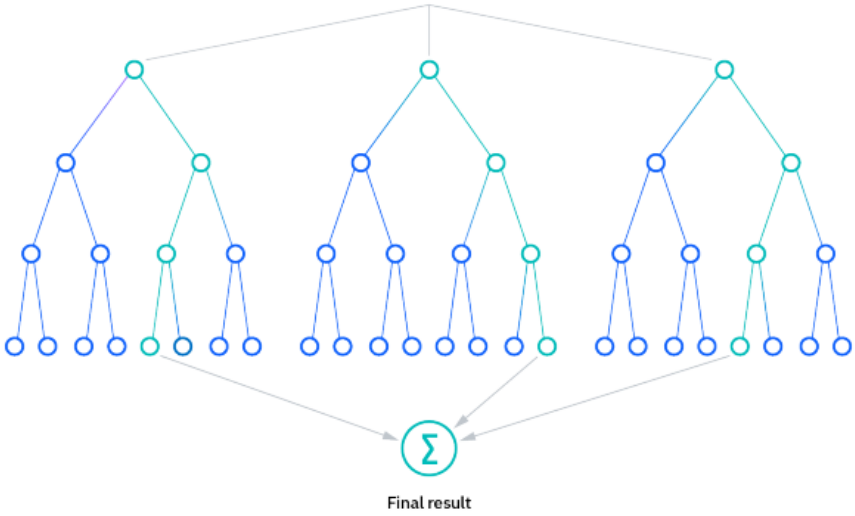


Figure 29: Random forest IBM, 2024

7.5 Hyper-parameter optimization:

Machine learning models are based on the usage of different hyper-parameters. They are critical for building robust and accurate machine learning models, where they help us find a balance between bias and variance, preventing the model from overfitting or underfitting, Yıldırım, 2021. This is done using an open-source package called Optuna. The software automates searching for the best parameters using intelligent sampling techniques, such as random search. More on this can be found in the Optuna official documentation Optuna on GitHub ,Team, 2022.

The main idea behind using this package is automating an extensive grid search for the best parameters and also being able to visualize the results of the hyperparameters so that the optimal intervals of these parameters are easily determined. Figure 30 shows an overview of the hyperparameters optimized in this research.

7.5.1 Neural Networks:

For ANN there are 2 levels of hyperparameters optimizations. The first level tunes the number of neurons, activation functions, optimizer, learning rate, batch size, and epoch. The second step will be tuning the number of layers, Rendyk, 2024.

The initial focus starts by tuning the **number of neurons** in each hidden layer. Generally, all layers could share the same number of neurons, but customization is possible. Tasks with higher complexity require a higher number of neurons. In this framework, the specified range of neurons spans from 10 to 100.

An **activation function** is a parameter in each layer that significantly ignites the hidden nodes to produce a more desirable output. The primary function of the activation function is to determine the neural network's output. Among many available activation functions, the Relu function is the most used one in Neural Networks, which stands for Rectified Linear Units. It introduces nonlinearity to neural networks, enabling them to learn complex patterns. It outputs the input directly if positive; otherwise, it outputs zero, helping to mitigate the vanishing gradient problem and improve training efficiency, Figure 31, Becker, 2018.

$$f(x) = \max(0, x) \tag{37}$$

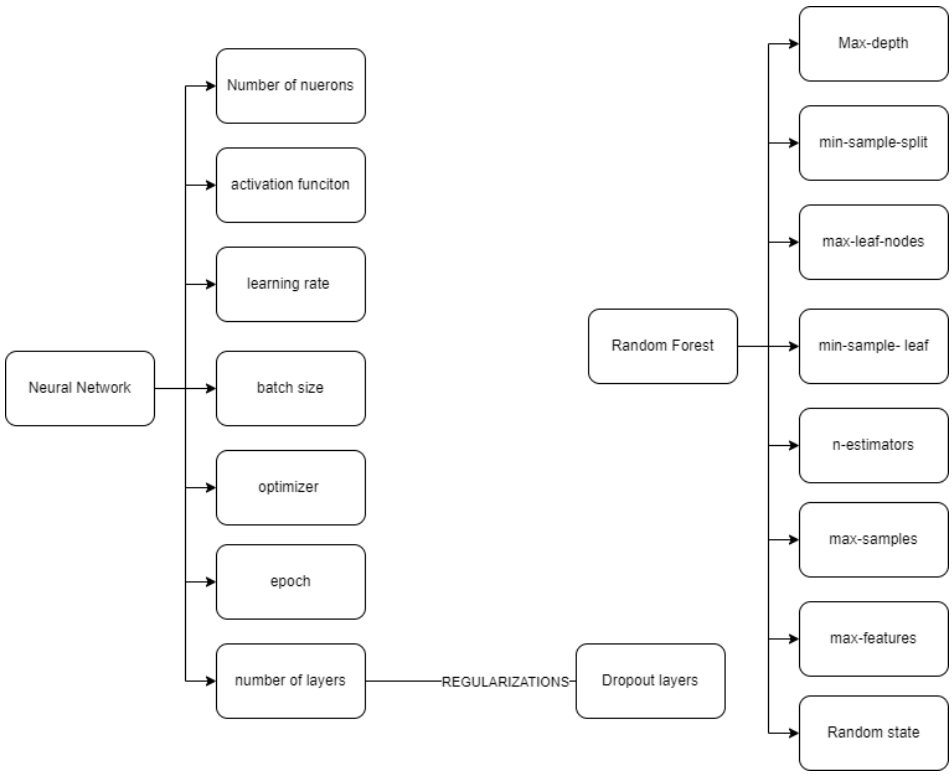


Figure 30: Overview hyperparameters

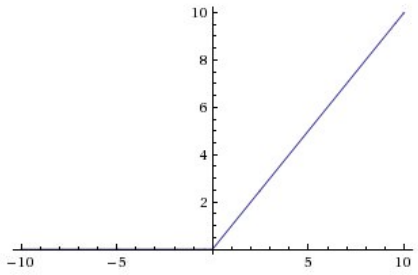


Figure 31: ReLU activation function, Becker, 2018

In this framework, an **optimizer** is responsible for changing the learning rate and weights of neurons in ANN to reach the minimum loss functions. They are critical to reach the highest accuracy or minimum loss. On that note, the **learning rate** is a hyperparameter of the optimizer itself. It controls the step size of a model to reach the minimum loss function. A higher learning rate makes the model learn faster, but it may miss the minimum loss function and only reach a point next to it. Thus, a lower learning rate gives a better chance of finding a minimum loss function.

Optimizing the **batch size** is helpful when dealing with large training datasets, as building models can be time-consuming. By optimizing the batch size, not all training data are fed to the model simultaneously. Therefore, a smaller batch size slows the learning rate but may increase variance in validation dataset accuracy. Conversely, a larger batch size slows learning while stabilizing validation dataset accuracy variance.

Following that, The term **"epoch"** in neural networks refers to the number of times a complete

dataset is fed forward and backward through the model during training. Insufficient epochs can cause underfitting, indicating inadequate learning, while excessive epochs may lead to overfitting, where the model memorizes existing data but performs poorly on new data. It's essential to find the right balance to optimize learning.

The **number of layers** in a neural network is a crucial hyperparameter that affects the model's predictive performance. Fewer layers suffice for more straightforward problems, while more layers may be required for complex issues.

Regularization technique: Since the neural network model tends to over-fit, regularization techniques prevent this by adding a penalty term to the loss function during training. This patently discourages the model from becoming complex or having large parameter values. The regularization techniques are L1 and L2 regularization, drop_out, early stopping, and more. Applying such techniques makes the model more robust and better at predicting unseen data, Jain, 2024.

As a part of regularization techniques **Dropout** randomly removes a certain percentage of neurons in a layer during training, preventing over-reliance on specific neurons. The dropout rate determines the proportion of neurons to drop.

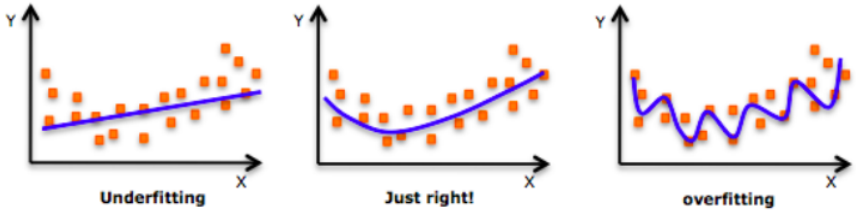


Figure 32: Difference underfitting and overfitting, Jain, 2024

Batch normalization normalizes the values within each batch, akin to standard scaling in traditional machine learning.

Early stopping: Early stopping t prevents the model from overfitting the data. It stops the training process before the model starts to over-fit. The idea is to stop training when the model's performance begins to degrade, which is indicated by the divergence between the training and validation losses, as Figure 33 indicates.

Lastly.,**ReduceLROPlateau:** Reduces the meaning rate when a metric has stopped improving. For more information about this topic, visit Analytics Vidhya - Neural Networks Tuning

7.5.2 Random Forest:

Random forest has seven different parameters that could optimized. Saxena, 2023 provides the following definitions.

max-depth: Determines the maximum depth of each decision tree in the forest. A deeper tree can capture more complex relationships in the data but may also lead to overfitting if not properly constrained.

m**in-samples-split**: Specifies the minimum number of samples required to split an internal node during tree-building. If the number of samples at a node is less than this value, the node will not be split further, effectively controlling tree growth.

max-leaf-nodes: Sets the maximum number of leaf nodes in each decision tree. Limiting the number of leaf nodes helps prevent overfitting by constraining the complexity of the trees.

min-samples-leaf: Specifies the minimum number of samples required at a leaf node. If a split results in a leaf node with fewer samples than this value, the split is not allowed, which helps control the size of the trees and prevents overfitting.

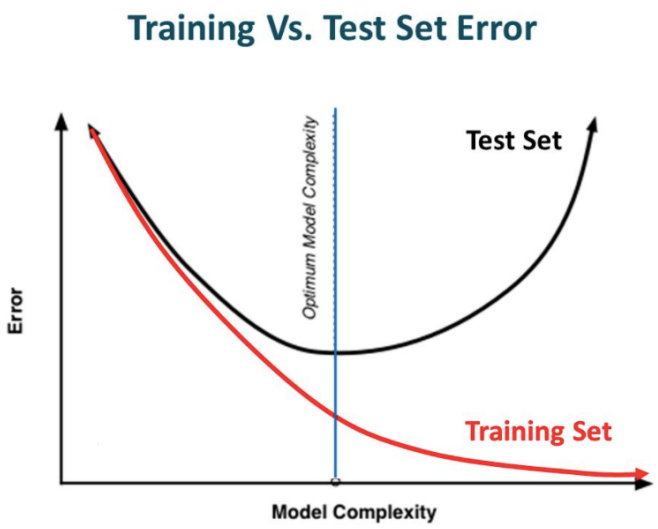


Figure 33: Validation and training losses, Jain, 2024

n-estimators: Represents the number of decision trees used in the Random Forest ensemble. Increasing the number of estimators can improve model performance but also increases computational cost.

max-samples (bootstrap sample): Indicates the maximum number of samples to be drawn from the training dataset for each decision tree in the ensemble. This parameter controls the bootstrap sample size used for training each tree.

max-features: Determines the maximum number of features to consider for the best split at each node. It helps introduce randomness into the Random Forest by limiting the number of features considered for each split, which can improve generalization and prevent overfitting.

The random state parameter is a way to "lock in" the algorithm's random processes so you can reproduce your results. More on this topic is documented in Analytics Vidhya - Random Forest Tuning

7.6 Data Manipulation - Auto-Encoders:

Dimensionality reduction reduces the number of features (or dimensions) in a dataset while retaining as much information as possible. This can be done for various reasons, such as to reduce the complexity of a model, improve the performance of a learning algorithm, or make it easier to visualize the data. Several dimensionality reduction techniques exist, including principal component analysis (PCA), Auto-encoders, etc. Each technique uses a different method to project the data onto a lower-dimensional space while preserving important information, Geeksforgeeks, 2023. However, in this research, only Auto-Encoders are included as the other methods did not perform well on my data.

Autoencoders are a type of Neural network designed to compress or encode the input data into its essential features and then represent (decode) the original input from this compressed representation, IBM, 2023. In simplified words, this encoder redefines the data in a less noisy way with stronger relationships to enhance the performance of machine learning models. This is similar to looking at a blurry photo next to a clear one. The blurry photo would be the original data, while the clearer one is the reconstructed data; check Figure 34. Auto Encoders are used when dealing with large, complex, unlabelled, or noisy data sets requiring non-linear dimensionality reduction, and latent features are more important than preserving the data's original structure.

Figure 34: Auto Encoder Explanation (AI generated photo)

7.7 Applying machine learning:

Training the machine learning path involves taking multiple steps before reaching the final results. Figure 36 shows an overview of the whole process. For that purpose, this research performed three different loops. First, the hyperparameters optimization loop. This is where the NN and RF hyperparameters were optimized using the Optuna package to assess the best hyperparameters using four different scenarios:

1. Atterberg limits [two input features]
2. Clay contents [three input features]
3. Triangular clay class [one input feature]
4. Plasticity index classes [one input feature].

These four scenarios were used as the input to predict the other variables for the clay inspection data frame. On the other hand, only the water content (a common variable between CI and CC data frames) was used to predict the clay compaction results.

After achieving good results, the second phase starts, where the NN and RF performances combine the data frames into one data frame, using either NN or RF's learning and precision capabilities. In this step, both models were trained on the CC data and then used to predict the dry unit weight and proctor density scenarios according to the calculated water content at the following consistency indices (0.60, 0.75, and 0.85). This resulted in a data frame as big as the CI data frame, which is twice the size of the CC data frame. Figure 35 shows a simple procedure representation.

In this case, the models were trained on the best fold (K-fold validation) and then predicted the variables for the 4600 rows of the data (which, in this case, are considered unseen new data). This also gave insight into the models' performance when dealing with unseen data and how well they are generalized. This is the only operation where the number of predicted data equals the number of input data, making the size of the input and out tensors the same.

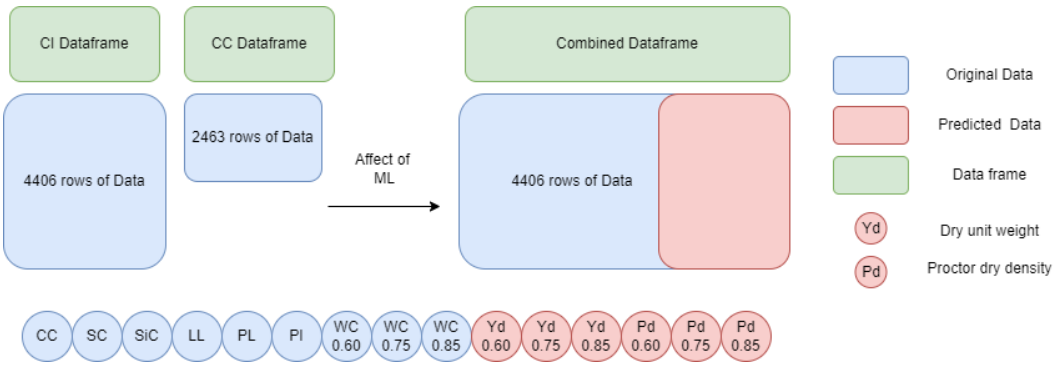


Figure 35: Combined data frames

This phase resulted in two combined data frames containing all the variables of the clay inspection data, including the three scenarios of water contents and six predicted variables (3 for dry unit weight and 3 for proctor density). Furthermore, an autoencoder was applied to both these data frames, which resulted in two reconstructed data frames. On this note, the research tested the model on **four** different data frames, which have the same features:

- COMNN: Combined Neural Network Dataframe
- COMANN: Combined reconstructed Neural Network Dataframe
- COMRF: Combined Random Forest Dataframe
- COMARF: Combined Auto encoded Random Forest Dataframe

These combined data have the following features, Table 20; all of these rows have 4406 rows of data. The performance of this combination procedure is shown in Section 8.3.3

Table 20: Features of the combined data frames

Feature
Liquid Limit
Plastic Limit
Plasticity Index
Clay Content
Sand Content
Silt Content
Class Triangular
Class Plasticity diagram
Water Content (0.75)
Water Content (0.60)
Water Content (0.85)
Predicted Dry Density Proctor (0.75)
Predicted Dry Unit Weight (0.75)
Predicted Dry Density Proctor (0.60)
Predicted Dry Unit Weight (0.60)
Predicted Dry Density Proctor (0.85)
Predicted Dry Unit Weight (0.85)

In phase 3, the performance of the machine learning models was tested on the data frames using the scenarios mentioned above. This resulted in a data frame with the best prediction performance. The performance of this data frame is compared against the actual data, linear correlation equations found from the original data frames, and equations from the literature. Eventually, the significance of the machine learning models will be compared with the current prediction methods with an assessment of how uncertain these predictions are.

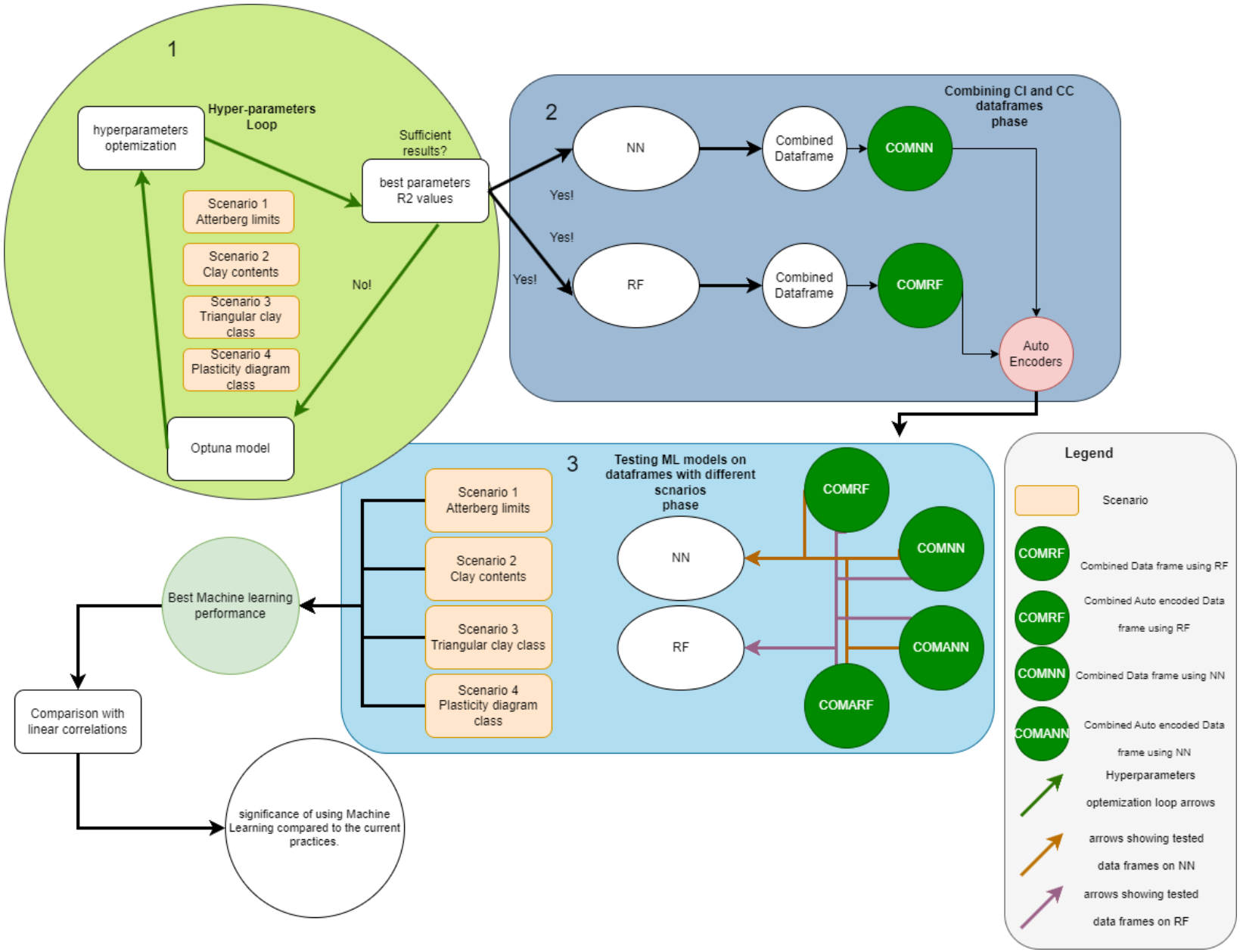


Figure 36: ML loops

7.8 Summery: Machine learning

A. Preprocessing the Data

- Standard scalar method is used for normalization due to the Gaussian distribution of variables.

B. Data Validation

- Employs a validation set for objective assessment of model performance.
- Helps in hyperparameter tuning and prevents overfitting.
- K fold validation has been used in this research; the number of folds is determined to be five fold for this research. This made the output tensors of the ML models 1/fifth the original volume, changing it from [number of features per scenario,4406 data points] to [number of features per scenario, 882 data points]

C. Machine Learning Models

- Focus on Neural Networks (NN) and Random Forests (RF) due to superior performance.
- Support Vector Machines (SVM) excluded due to poor results.

D. Hyperparameter Optimization

- Employs Optuna for automated hyperparameter tuning.
- Aims to find optimal parameters for enhanced model performance.
- Parameters optimized for both NN and RF models.

C. Data manipulations

- PCA and similar methods were not included in this research because of weak performance
- Auto-Encoder was eventually used due to the stability complexity of data.

D. Applying machine learning

- Phase 1 was used for the hyperparameters optimization using 4 scenarios
- Phase 2 combined the data frames using both NN and RF, producing two combined data frames, which were also reconstructed, ending with four different data frames.
- Phase 3 tested the scenarios on all the data frames and showed which ML model with the best predictions and best data frame allowed the best results.
- Predictions are then compared with the found equations from the literature and the linear correlations. This reflected on whether or not the performance of ML was able to improve the current estimation process.

8 Results: Machine learning performance & hyperparameters optimization

The following section presents the results of the machine learning models, Neural Networks, and Random Forests. However, the optimized hyperparameters are presented first, Section 8.1, followed by the hyperparameters for the Auto-encoders. After that, the auto-encoders' results are shown, and the reconstructed data is compared against the original data. The effect of the auto-encoders is then shown in the data classifications, both the triangular and plasticity diagram classifications, Section 8.2.

After that, the performance of the NN and RF is shown separately on the clay inspection and clay compaction data sets, both on the original data sets and the reconstructed ones, Section 8.3. The research then explores the performance of combining the two data sets, showing how the models performed when combining these two datasets and which one outperformed the other. Furthermore, both models' performances are compared using the data frame with the best performance (as 4 data frames were used to assess the performance as explained in the third phase of the previous figure), Section 8.4.

On that note, the ML performance is compared with literature-based correlations and the linear correlations from the combined data set, Section 8.5. Eventually, this part of the report ends with assessing the uncertainty of the found predictions, Section 9.

8.1 Hyperparameters optimization:

Using the Optuna model, the best hyperparameters were achieved after trying to predict the variables using the four mentioned scenarios. Furthermore, the ranges of the parameters achieved were also tested using the CC data frame. This optimization had the object of maximizing the R-squared value of the predictions, which are presented in Table 21.

Table 21: Best Hyperparameter Intervals, R-squared objective

NN Hyperparameter	**RF Hyperparameter**
Activation function: relu	Max features: sqrt
Batch size: [80-120]	Max samples: [0.6-1]
Dropout rate: [0.4-0.5]	Min sample leaf: [1-4]
Epochs: [900-1200]	Min sample split: [6-8]
Learning rate: [0.0001-0.001]	Number of estimators (n_estimators): [100-600]
Number of layers (num_layers): [10-12]	Max depth [10-100]
Units per layer: [150-250]	Random state: fixed at 42

On that note, Optuna also showed a hyperparameter importance figure (sensitivity analysis figure) that shows each hyperparameter's impact. The Neural Network was mainly affected by the learning rate of the model as it had 60% importance in the model; this was then followed by the drop_out rate, batch size, and unit per layer as they scored 14%,11%, and 10%, respectively Figure 37. Furthermore, epochs and the number of layers scored less than 5%. Lastly, many activation functions were tested at the beginning. However, the Relu activation function seemed to score the best, so only Relu was used in further Optuna trials to save computational time, as each NN optimization run took around 6 hours to finish. This is why Figure 37 shows 0% for the activation function hyperparameter.

On the other hand, the importance of the RF hyperparameters was more spread among them, where max_depth and max_samples scores 47% and 26%, respectively, as the most important hyperparameters of the RF model. Furthermore, the min_samples_leaf scored 14%, min_samples_split 7%, and number_of_estimators scored 6%. Similar to the activation function of NN., the max_feature was set to SQRT as it scored the best in the first Optuna trials so that it would save computational time. Lastly, the random state was fixed at 42 to ensure the predictions could be reproduced.

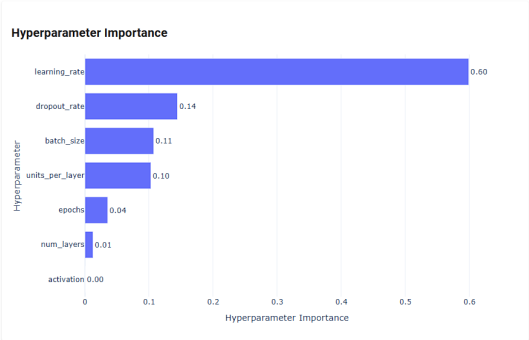


Figure 37: Neural network hyperparameters importance

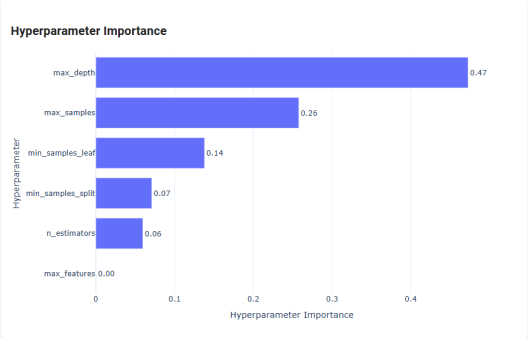


Figure 38: Random forest hyperparameters importance

8.1.1 Selection of hyperparameters:

In the case of the Neural Networks, the previous parameter optimization was directed toward maximizing the R-squared value; minimizing the validation and training losses is essential, as it helps generalize the model performance. Table 22 shows the selected parameters for both models.

Table 22: Selected Hyperparameter

NN Hyperparameter	RF Hyperparameter
Activation function: Relu	Max features: sqrt
Batch size: 120	Max samples: 0.85
Dropout rate: 0.55	Min samples leaf: 1
Epochs: 1200	Min samples split: 2
Learning rate: 0.0005	Number of estimators (n_estimators): 100
Number of layers (num_layers): 8	Max depth: 50
Units per layer: 250	Random state: 42

On that note, Neural Networks did score good R-squared values; however, there was a difference between the training and validation loss. The following new hyperparameters were introduced to the Neural Network model to improve its performance; minimizing this difference helped generalize the model's performance, as Section 7.5.1 explained, resulting in a good training and validation graph, Figure 39. Therefore, the following hyperparameters were added to the Neural Network model:

- L2 regularization (weight decay) with a coefficient of 0.01
- Batch normalization is applied after each dense and dropout layer to stabilize training.
- A ReduceLROnPlateau callback is used, which reduces the learning rate by half if the validation loss doesn't improve for 10 epochs.
- An EarlyStopping callback is used to stop training early if the validation loss doesn't improve for 50 epochs. This prevents overfitting.

Most of the following runs have a similar figure unless indicated otherwise to save space in the report.

8.1.2 Auto-encoders hyperparameters

Auto-encoders are used in this research for a different task than other machine learning models. As explained in Section 7.6, the auto-encoders are used to reconstruct the data, making it easier to be trained upon by the machine learning models. Therefore, as a variation of Neural Networks, it shares a structure similar to the Neural Network; the previous hyperparameters were also used for the auto-encoder, with fewer neurons and batch size. These two hyperparameters were reduced to simplify the model and minimize data loss. Therefore, the number of neurons was set to 3 and the batch size to 32.

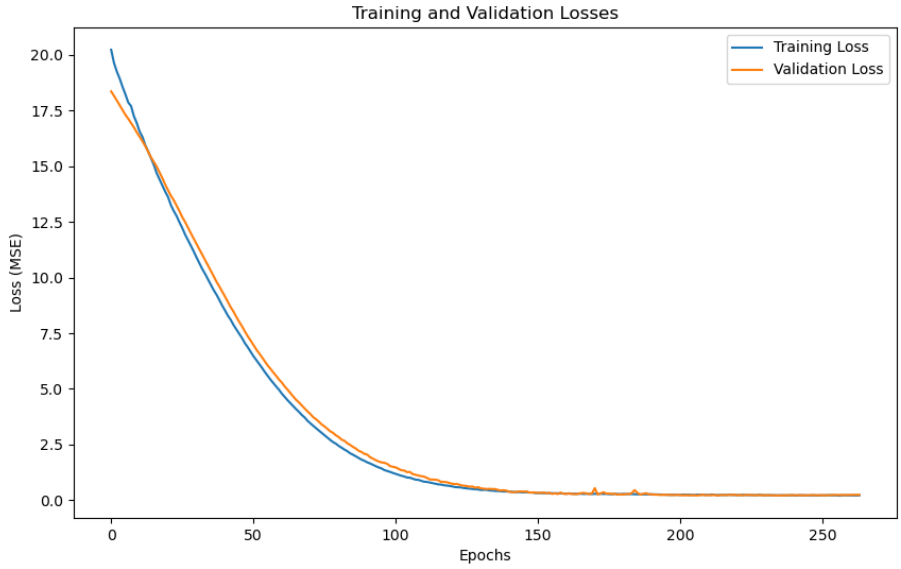


Figure 39: Validaiton and Training losses of the Neural Network

8.2 Auto-encoders performance

Both PCA and Auto-encoders are applied to the data. However, as the PCA did not perform well on the data, this method was excluded from this research as it could not capture the relationship of the data due to its non-linearity. The following shows the performance of the auto-encoder on both the clay inspection and compaction data sets.

8.2.1 Clay inspection data: auto encoded

As explained in Section 7.6, an auto-encoder is a model that reconstructs the data to improve the performance of machine learning models. These datasets are called the reconstructed data frames. Therefore, this subsection shows the encoders' results on the clay inspection and clay compaction data frames. The results show that the Atterberg limits have an average R-squared value of 0.90, Figure 40, Table 23. Furthermore, the contents got a more robust reconstruction R-squared value, around 0.95. However, only the clay content value got a weaker value of 0.86 Figure 41, Table 23 the MSE and MAE value scored good results, only the clay content and liquid limit MSE values are a bit higher, this is because of some outliers as the Figures indicate. Lastly, the three scenarios of the water contents were reconstructed with an excellent R-squared value, around 0.96, Figure 42, Table 23.

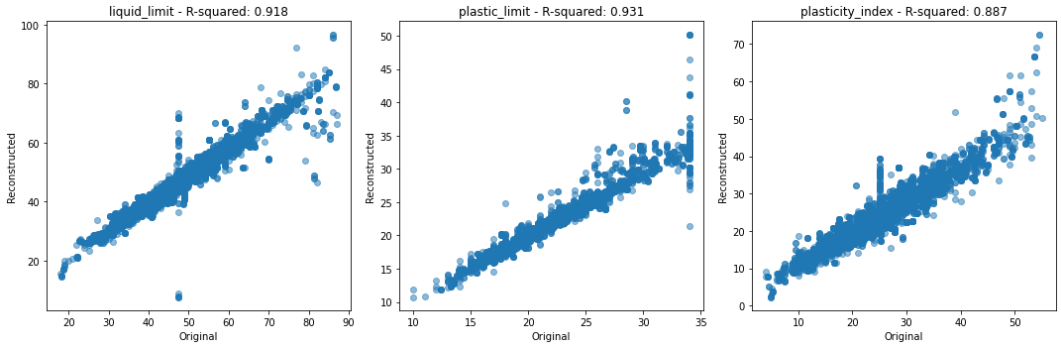


Figure 40: CI auto encoded variables: LL, PL, PI

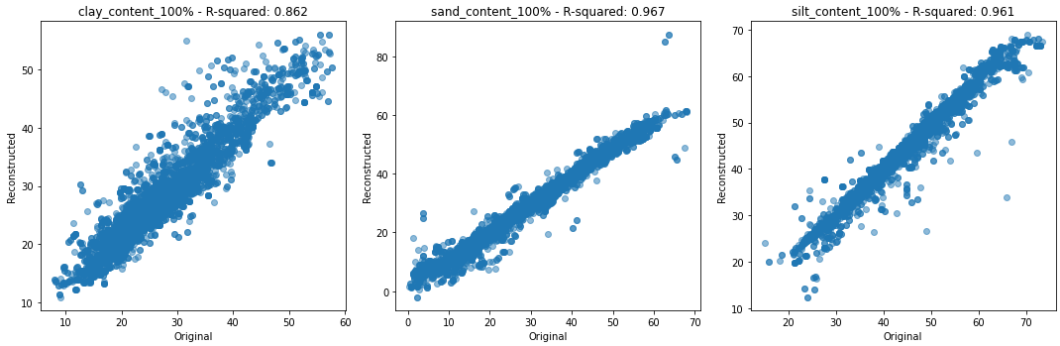


Figure 41: CI auto encoded variables: CC, SC, SiC

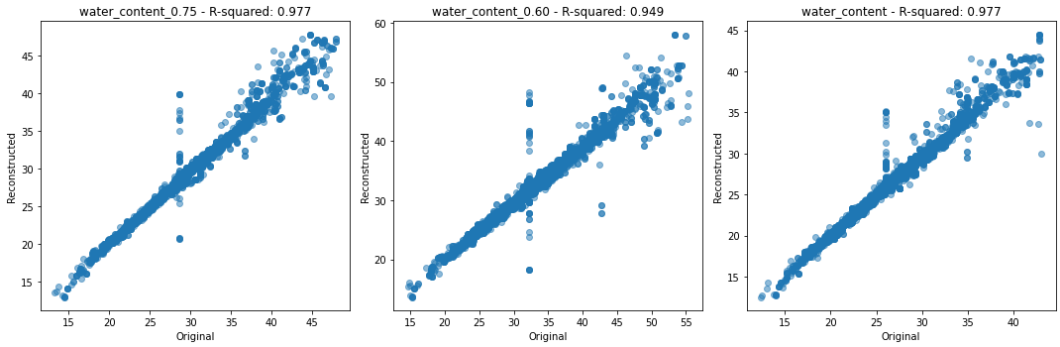


Figure 42: CI auto encoded variables: WC, WC:0.60, WC:0.85

Table 23: Auto-encoders Error Metrics and R-squared Values of Clay inspection variables

Variable	MSE	MAE	R-squared
liquid_limit	12.37	2.03	0.92
plastic_limit	1.34	0.62	0.93
plasticity_index	9.38	2.22	0.89
clay_content	12.16	2.50	0.86
sand_content	7.52	1.81	0.96
silt_content	4.26	1.19	0.96
Water_content_0.75	0.82	0.46	0.97
Water_content_0.60	2.59	0.80	0.95
Water_content_0.85	0.66	0.42	0.97

8.2.2 Clay compaction data: auto encoded

Following the same pattern, the auto-encoder model well reconstructed the clay compaction variables. Figure 43 and Table 24 show the results. The MAE and the MSE results of the Dry density proctor show greater values than other variables. However, this is because the density numerical values are around 1500 [kg/m3].

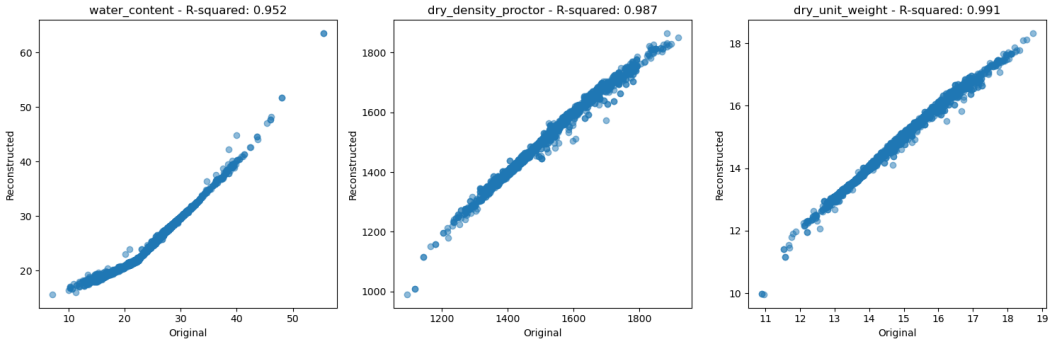


Figure 43: CC auto encoded variables: Water content, Dry density proctor, Dry unit weight

Table 24: Auto-encoders Error Metrics and R-squared Values of Clay compaction variables

Variable	MSE	MAE	R-squared
Water content	1.47	0.57	0.95
Dry density proctor	209.44	8.38	0.98
Dry unit weight	0.01	0.06	0.99

8.2.3 Reconstructed classifications:

The data has been reconstructed by auto-encoder. The position of the data points will differ as they are reliable on the clay and sand content for the triangular relationship and the liquid and plastic limits for the plasticity graph. Figures 44 & 45 show the positions of the data points before and after the autoencoders. These figures show that the reconstructed points are more focused in the Ks2, Ks3, and Ks4 classes, while the original data is slightly more spread towards the ks1 class.

On that point, the reconstructed plasticity graph in Figure 47 also shows more concentration towards the center of the points, making them less spread across the graph. It is worth mentioning that the reconstructed data points have some points scattered above the U line in the plasticity grap, which were removed to not affect the analysis.

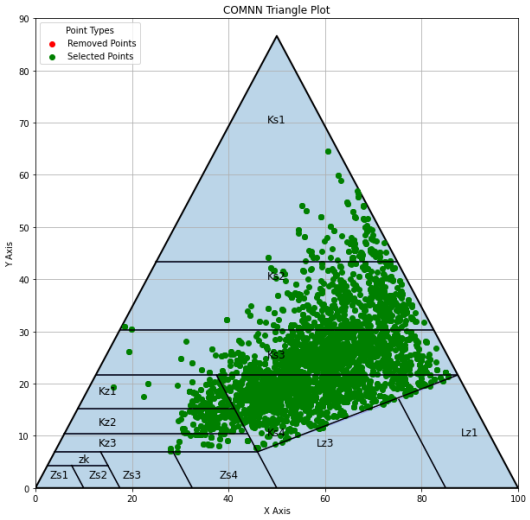


Figure 44: Triangular classification before auto encoder

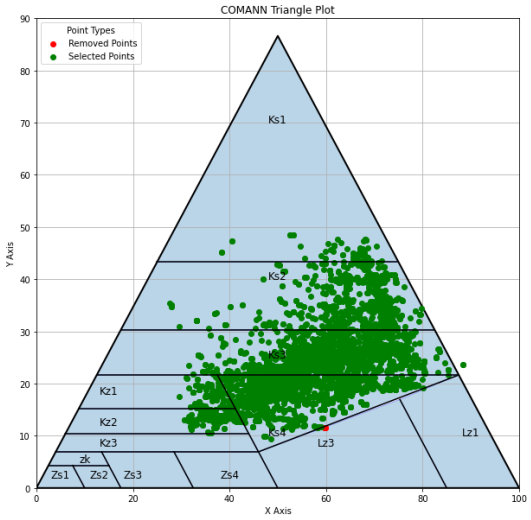


Figure 45: Reconstructed Triangular classification

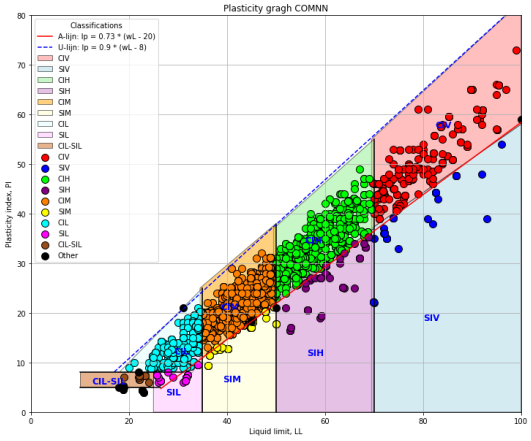


Figure 46: Plasticity graph before auto encoder

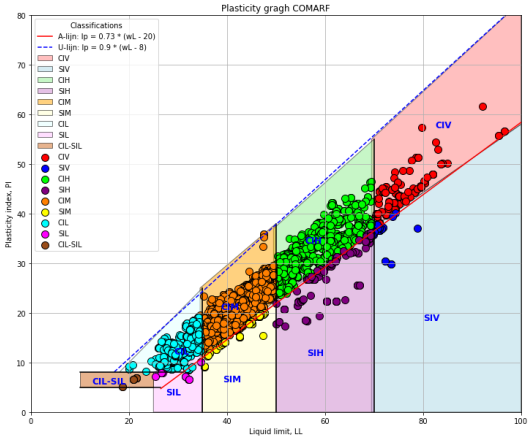


Figure 47: Reconstructed Plasticity graph

8.3 Machine learning performace:

The machine learning performance is tested on 4 main scenarios; those scenarios form the available data features that a designer might have in the design phase for the ML models to make predictions. Again, those scenarios are defined as knowing the following:

1. The Atterberg limits of the soil sample [two data features as input].
2. Clay contents, that is, clay, sand, and silt contents [three data features as input].
3. Triangular clay class [one data feature as input].
4. Plasticity diagram class [one data feature as input].

To emphasize the point, the performance of the ML models was tested on these scenarios for the clay inspection data set, and then the models were tested using the water content variable (as a known feature) for the clay compaction dataset, as it is the common variable between the clay inspection and clay compaction datasets. This is done on both the original and reconstructed data sets.

8.3.1 Clay Inspection data:

Using the optimized models, Table 25 shows the scored R-squared results for the clay inspection data. This table shows that using the Atterberg limits scores the best results, followed by the scenario using the clay contents and the one with the plasticity diagram class. Having the triangular class as a known variable scored weak results, making this scenario unreliable at this stage of the results.

Note: **Bold** variables names in the following four tables indicate the known feature; therefore, no r-squared values are mentioned. Furthermore, the water content measures refer to the average value for the three water content measures calculated using the 0.6, 0.75, and 0.85 consistency indices, Section 5.3.4

Comparing the original and reconstructed data sets, the predictions are weaker when using the original data. However, it is noticeable that the Random Forest had more reliable results using the different scenarios, except the one with the triangular class, as there was not much difference between the two models. This makes the reconstructed data set more predictable than the original one, especially in cases where the soil contents are known as the prediction was enhanced from weak predictions to moderate ones.

Table 25: Prediction of original clay inspection data using NN and RF

CI: Original data	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
	NN	RF	NN	RF	NN	RF	NN	RF
Clay content	0.56	0.82	**CC**	**CC**	0.86	0.59	0.46	0.50
Sand content	0.44	0.78	**SC**	**SC**	0.44	0.56	0.29	0.40
Silt content	0.18	0.65	**SiC**	**SiC**	0.07	0.22	0.11	0.11
Liquid limit	**LL**	**LL**	0.57	0.79	0.49	0.5	0.75	0.83
Plastic limit	**PL**	**PL**	0.29	0.63	0.18	0.19	0.41	0.61
Plasticity Index	0.94	0.99	0.58	0.81	0.52	0.53	0.74	0.77
Water content measures	0.95	0.99	0.44	0.71	0.35	0.35	0.62	0.78
Tri class	0.14	0.49	0.96	0.99	**Tri class**	**Tri class**	0.26	0.41
PI class	0.93	0.94	0.28	0.65	0.28	0.46	**PI class**	**PI class**

Table 26: Comparison of Auto Encoded Data Using NN and RF

CI: Reconstructed	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
	NN	RF	NN	RF	NN	RF	NN	RF
Clay content	0.87	0.97	**CC**	**CC**	0.68	0.69	0.60	0.70
Sand content	0.5	0.78	**SC**	**SC**	0.4	0.53	0.27	0.47
Silt content	0.21	0.64	**SiC**	**SiC**	0.06	0.17	0.03	0.12
Liquid limit	**LL**	**LL**	0.85	0.93	0.52	0.53	0.65	0.82
Plastic limit	**PL**	**PL**	0.65	0.79	0.16	0.17	0.32	0.61
Plasticity Index	0.95	0.99	0.88	0.96	0.62	0.63	0.59	0.71
Water content measures	0.9	0.99	0.79	0.87	0.30	0.35	0.55	0.72
Tri class	0.22	0.6	0.36	0.78	**Tri class**	**Tri class**	0.27	0.52
PI class	0.22	0.6	0.36	0.78	**PI class**	**PI class**	**PI class**	**PI class**

8.3.2 Clay compaction data:

Using the water content as the input of this data frame led to good prediction results, as Table 27 shows. Both the original and the reconstructed datasets scored good prediction values. On that note, the reconstructed scored slightly better results, again giving it the advantage over the original data set.

Note: Only the clay inspection data set has three measures of water content. The clay compaction data set has only one water content variable, an original variable (reported in the tests). It is not estimated like the case of the clay inspection data frame.

Table 27: Prediction of original clay compaction data using NN and RF

CC: Original Data	NN	RF
Water content	**Water content**	**Water content**
Dry unit weight	0.89	0.91
Dry density in situ	0.74	0.77
Dry density proctor	0.79	0.82

Table 28: Comparison of reconstructed data using NN and RF

CC: Reconstructed	NN	RF
Water content	**Water content**	**Water content**
Dry unit weight	0.88	0.95
Dry density in situ	0.77	0.89
Dry density proctor	0.8	0.91

8.3.3 Combining the datasets:

The water content is a common variable between the clay inspection and compaction data frames. The Clay inspection data frames have three assumed measures of the water content, each calculated using a separate compaction index, namely, 0.6, 0.75, and 0.80. Therefore, the dry unit weight was predicted using these measures, which resulted in Figures 48 & 50 & 52. These three figures show that both models predicted the dry unit weight well when compared with the compaction data's water content and dry unit weight (the original data).

On the other hand, it is noticeable that the assumed measures 0.60 and 0.75 did lead to low dry unit weight & proctor dry density predictions, which was not the case for the assumed measure 0.85. This is because the statistics (distribution and data intervals) of the water content (0.85) are similar to those of the original water content variable of the Clay compaction data. Check Tables 14 and 15 and Appendix B for a visual check.

While both models show good performance, it's noticeable that the Random Forest better predicted the dry unit weight value for new unseen data. This is shown on the green spike of Figures 48 & 50, which shows that the Random Forest model captures the limits of the original variable (of the compaction clay data frame, shown as black points in the figures) for new unseen data. These spikes are visible around a dry unit weight of 12 [kN/m3] or a dry density value of 1200 [kg/m3]. This also proves the R-squared values of RF are better than those of the Neural Network model. This is clear to see when compared with the blue lines (NN predictions) which extend for values lower than the original values (black points)

Following the same course, the prediction of the proctor dry density led to similar results, with also wrong predictions of high water content values, which is seen in Figures 54 & 56. Furthermore, the scenario of 0.85 did not experience the same inaccuracy of predictions.

On that note, using the reconstructed data frames also leads to similar results. However, the predictions of the reconstructed data seem inaccurate at extremely high and low water content values. Again, this is apparent for scenarios 0.60 and 0.75. Scenario 0.85 only experienced some inaccuracies in prediction at low water content values.

Summarily, Looking at Figures 48 till 59, Random Forest's performance on the combined reconstructed clay inspection data captures the spread of the original compaction data better than the Neural Network's predicted variables for the original clay inspection data set. This means the Random Forest model delivers more certainty in combining the data frames than the NN, making the random forest's combined data frames (original or reconstructed) more reliable for further analysis. More on this aspect is explained in Section 8.4.1 and Section 9.

Lastly, to make the predictions fall into the physical range of the original data sets. The projections of all scenarios are corrected by replacing the predictions' low and high values with the original data's limits. The low dry unit weight & proctor dry density values were replaced with a minimum value taken from Table 15, 10.90 [kN/m3] for dry unit weight and 1094.00 [kg/m3] for the proctor dry density. On the other hand, the high prediction was replaced with a max value of 1919 [kg/m3] for the proctor dry density and a value of 18.73 [kN/m3] for the dry unit weight.

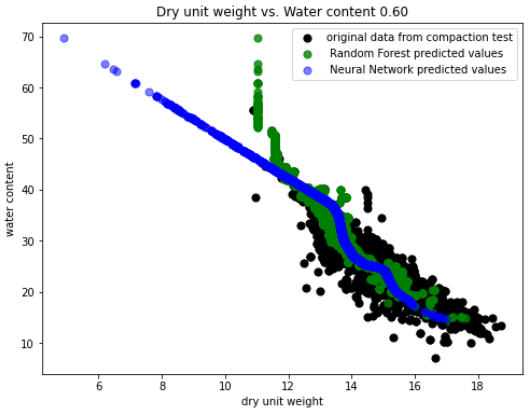


Figure 48: Predicted unit weight 0.60

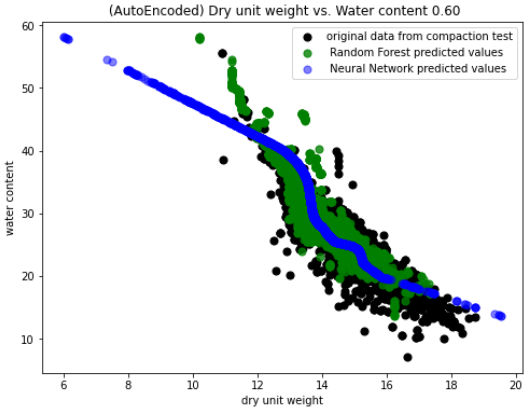


Figure 49: (Reconstructed) Predicted unit weight 0.60

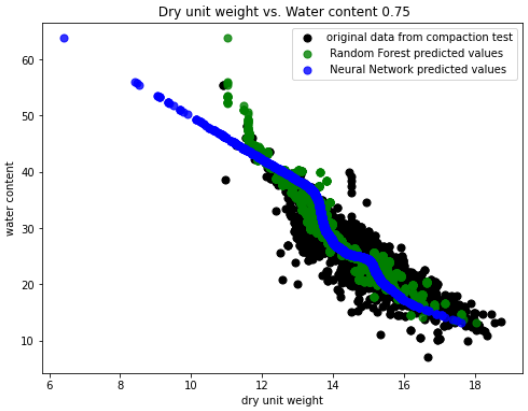


Figure 50: Predicted unit weight 0.75

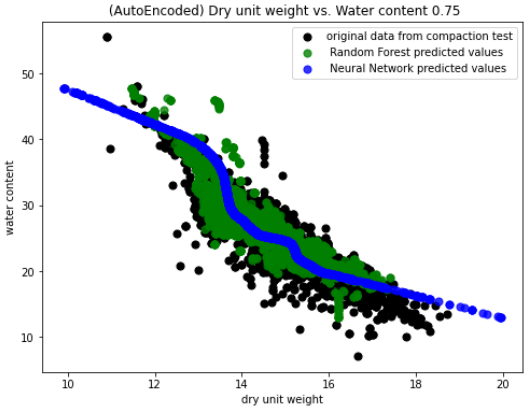


Figure 51: (Reconstructed) Predicted unit weight 0.75

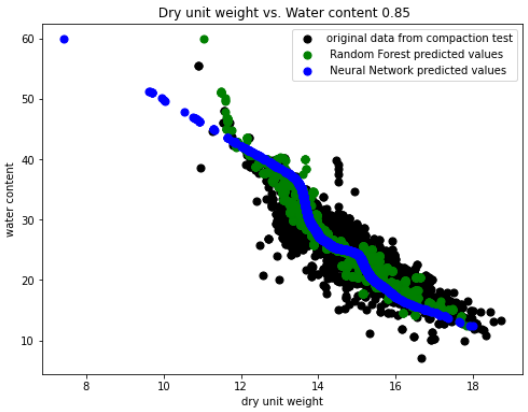


Figure 52: Predicted unit weight 0.85

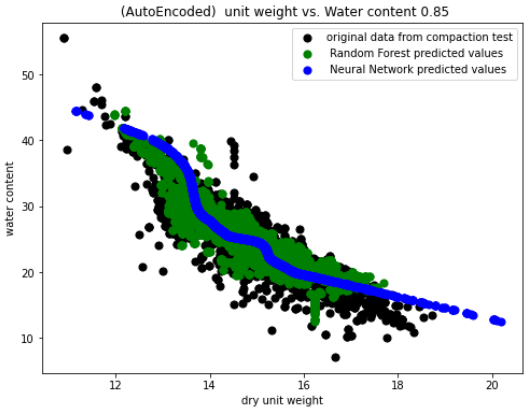


Figure 53: (Reconstructed) Predicted unit weight 0.85

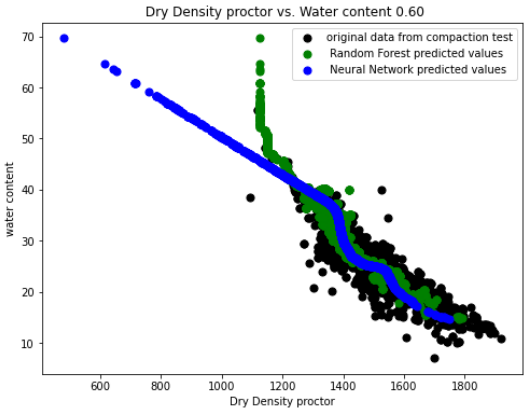


Figure 54: Predicted Dry density proctor 0.60

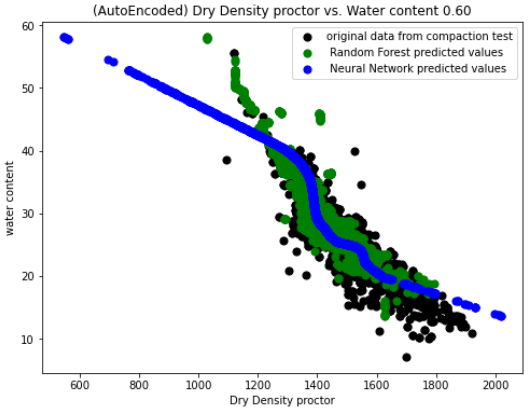


Figure 55: (Reconstructed) Predicted Dry density proctor 0.60

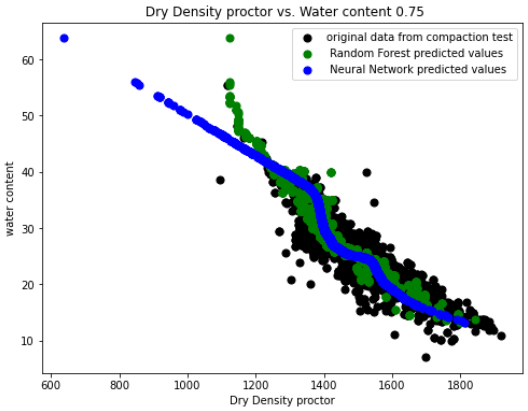


Figure 56: Predicted Dry density proctor 0.75

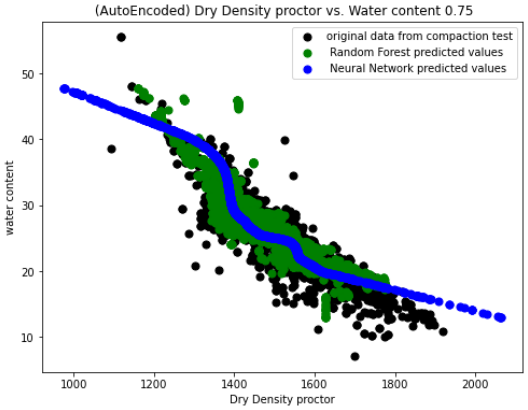


Figure 57: (Reconstructed) Predicted Dry density proctor 0.75

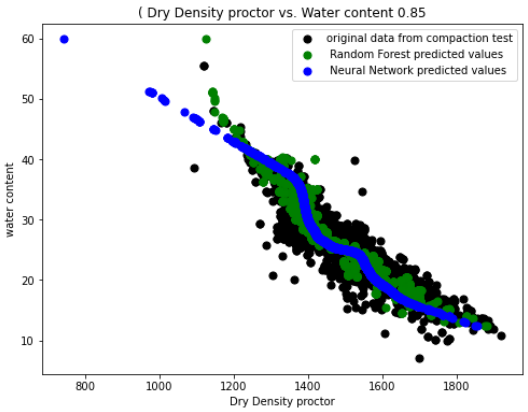


Figure 58: Predicted Dry density proctor 0.85

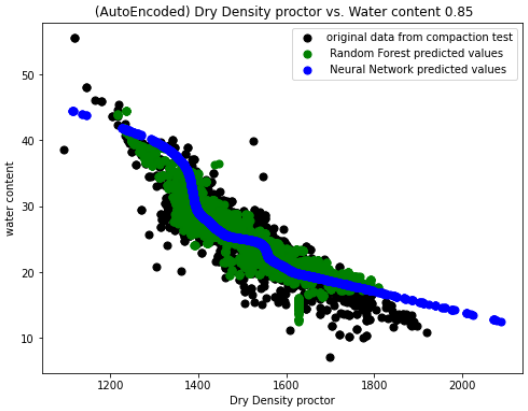


Figure 59: (Reconstructed) Predicted Dry density proctor 0.85

8.4 ML performance on the combined Data frames:

From this point forward, the machine learning models are tested on four different data frames: a Combined Neural Network (COMNN), a Combined Random Forest (COMRF), a combined reconstructed Neural Network (COMANN), and a combined reconstructed Random Forest (COMARF). These data frames were used to train the models upon them individually and then compare whether a dataset led to better predictions.

8.4.1 Performance of machine learning of the combined data frames

As previously mentioned in Section 7.7, both models are tested using four different scenarios. These scenarios define what features the designer might know about a soil sample and what is not. Again, these scenarios are defined as follows:

- Scenario 1: the Atterberg limits of the sample are known as the Liquid Limit and plastic limit [two input features].
- Scenario 2: the contents of the clay are known, which are the clay, sand, and silt content according to the Triangular classifications [three input features].
- Scenario 3: the clay class is known according to the Triangular classification (NEN 5014) [one input feature].
- Scenario 4: According to the Plasticity graph classification, the clay class is known [one input feature].

The detailed results of these scenarios contain a large amount of numbers. Therefore, it was decided to summarize the results in a couple of split findings per scenario, comparing the original and reconstructed data frames. This choice makes it easier for the reader to comprehend the vital idea of the findings.

The reconstructed data frames produce better results than the original ones, similar to the findings found when the data frames were combined. This is valid for both the Neural Network and the Random Forest model. However, it is worth mentioning that the scenarios that knew the Atterberg limits or plasticity graph class slightly worsened when using the reconstructed data on Neural Networks. This was relevant for both scenarios 1 and 4, Table 29.

On this point, the random forest model had better prediction results across all the scenarios and all the data frames. This is followed by the model's rapid training and validation time (computational time). Furthermore, the RF models showed more consistent prediction values than the NN models.

On top of that, the best predictions were those applied to the COMARF, which again stands for Combined reconstructed Random Forest data frame. Slightly similar results were obtained from the COMANN, the Neural Network counterpart, but the predictions made using the COMARF data frame were still better. On top of that, the difference between these two data frames was only apparent in the dry unit weight and proctor dry density predictions. This is logical since the model combined that part of the data frames while the other variables were the original data.

Since the Atterberg limits were used to calculate the water contents (using the consistency index), and water content was the connecting variable (when the data frames were combined). This gave the Atterberg limits scenario an advantage, allowing the ML models to capture the relationship between the Atterberg limits and the rest of the variables.

Lastly, to make the comparison more straightforward and not confuse the reader with many numbers. The results of the following subsections are of the COMARF data frame, as the ML models showed the best results using this data frame. Furthermore, compared to the original data frame, the results of the COMRF were used, as the models performed better than their counterparts (COMNN). Figure 60 visually represents a summary of the performance of the models. Appendix H shows a table with the results of RF on all the data frames.

Table 29 shows the R-squared results of RF and NN models on the COMARF and COMRF data frames. For **scenario 1 (Atterberg limits)** both models showed good results across the predicted variables. However, it is clear that the RF model scores better on both data frames. This is very

apparent in the variables with a weak R-squared value, like silt content, where the NN model scored around 0.2, while the RF model scored around 0.50. Furthermore, both models scored excellent results for the water content, dry unit weight, and dry density proctor measures.

Scenario 2 (Clay contents) shows weak results overall, with only the triangular class scoring good results when using the RF model. This also applies to the dry unit weight, where all models scored weak to average r-squared results.

Moving on to **scenario 3 (Triangular class)** & **scenario 4 (plasticity graph)** then, both models did not show promising results. in the case of the third scenario, then only the variable clay content showed promising results. At the same time, only the liquid limit showed acceptable R-squared results in the case of the 4th scenario. Hence making both scenarios unreliable for predictions.

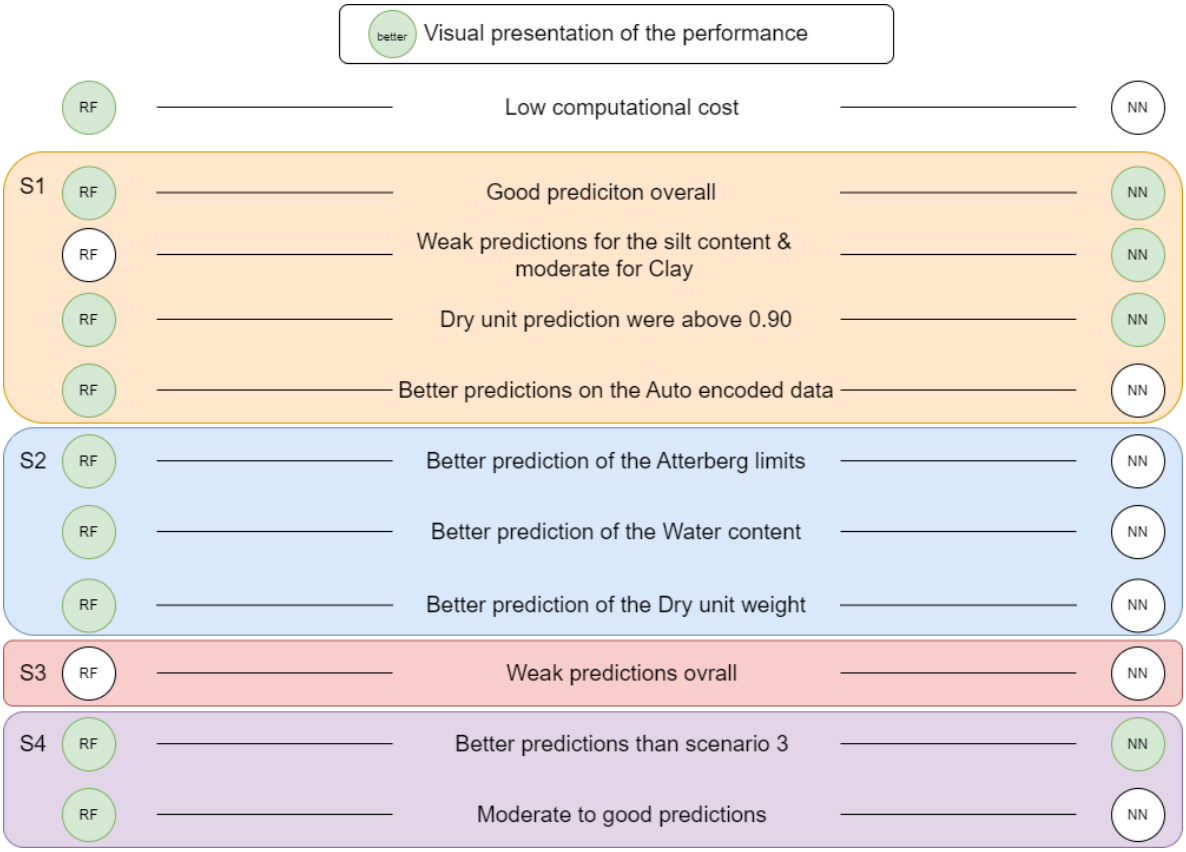


Figure 60: Summary of scenarios performance on NN and RF

Table 29: Results of RF and NN models on the COMARF data frame

Variable	RF: COMARF	NN: COMARF	RF: COMRF	NN: COMRF
Scenario 1: Atterberg limits [Two inputted features]				
Plasticity index	0.97	0.90	0.99	0.92
Clay content	0.97	0.91	0.76	0.59
Sand content	0.74	0.56	0.69	0.43
Silt content	0.50	0.23	0.52	0.19
Water content (*)	0.99	0.95	0.99	0.96
Dry unit weight (*)	0.90	0.84	0.97	0.90
Dry density proctor (*)	0.90	0.85	0.97	0.92
Triangular clay class	0.85	0.53	0.69	0.34
Plasticity graph class	0.94	0.88	0.94	0.94
Scenario 2: Clay contents [Three inputted features]				
Liquid limit	0.88	0.78	0.71	0.58
Plastic limit	0.61	0.38	0.50	0.31
Plasticity index	0.95	0.83	0.72	0.58
Water content (*)	0.80	0.63	0.60	0.48
Dry unit weight (*)	0.70	0.50	0.60	0.44
Dry density proctor (*)	0.65	0.52	0.60	0.43
Triangular clay class	0.94	0.75	0.92	0.46
Plasticity graph class	0.80	0.30	0.65	0.26
Scenario 3: Triangular class [One inputted feature]				
Liquid limit	0.69	0.66	0.50	0.51
Plastic limit	0.23	0.21	0.19	0.21
Plasticity index	0.79	0.74	0.53	0.54
Clay content	0.87	0.84	0.89	0.85
Sand content	0.54	0.41	0.56	0.45
Silt content	0.20	0.04	0.22	0.05
Water content (*)	0.43	0.40	0.35	0.39
Dry unit weight (*)	0.38	0.34	0.37	0.33
Dry density proctor (*)	0.35	0.38	0.35	0.33
Plasticity graph class	0.54	0.23	0.46	0.28
Scenario 4: Plasticity class [One inputted feature]				
Liquid limit	0.83	0.65	0.82	0.75
Plastic limit	0.62	0.32	0.61	0.41
Plasticity index	0.71	0.60	0.77	0.74
Clay content	0.70	0.60	0.50	0.46
Sand content	0.48	0.27	0.40	0.29
Silt content	0.13	0.03	0.11	0.04
Water content (*)	0.72	0.52	0.78	0.62
Dry unit weight (*)	0.62	0.36	0.75	0.51
Dry density proctor (*)	0.61	0.37	0.74	0.50
Triangular clay class	0.52	0.29	0.41	0.28

8.4.2 Performance of Linear regression on the combined data frames (COMARF)

Figure 61 shows the linear correlations between the variables of the combined data frame. The figure shows an R2 value of 0.88 between the liquid limit and Clay content and a value of 0.94 with the plasticity index. The Clay content shows no promising correlations with the dry unit weight or dry density. As stated before, the Atterberg limits were used to calculate the water contents, which were then used to combine the Clay Inspection and Clay Compaction data frames. The Atterberg limits show moderate to good correlations with the other variables across the data frame, with weaker correlation on the contents side, except the correlation between the clay content & liquid limit and the plasticity index & clay content. Appendix F shows the complete list of linear equations.

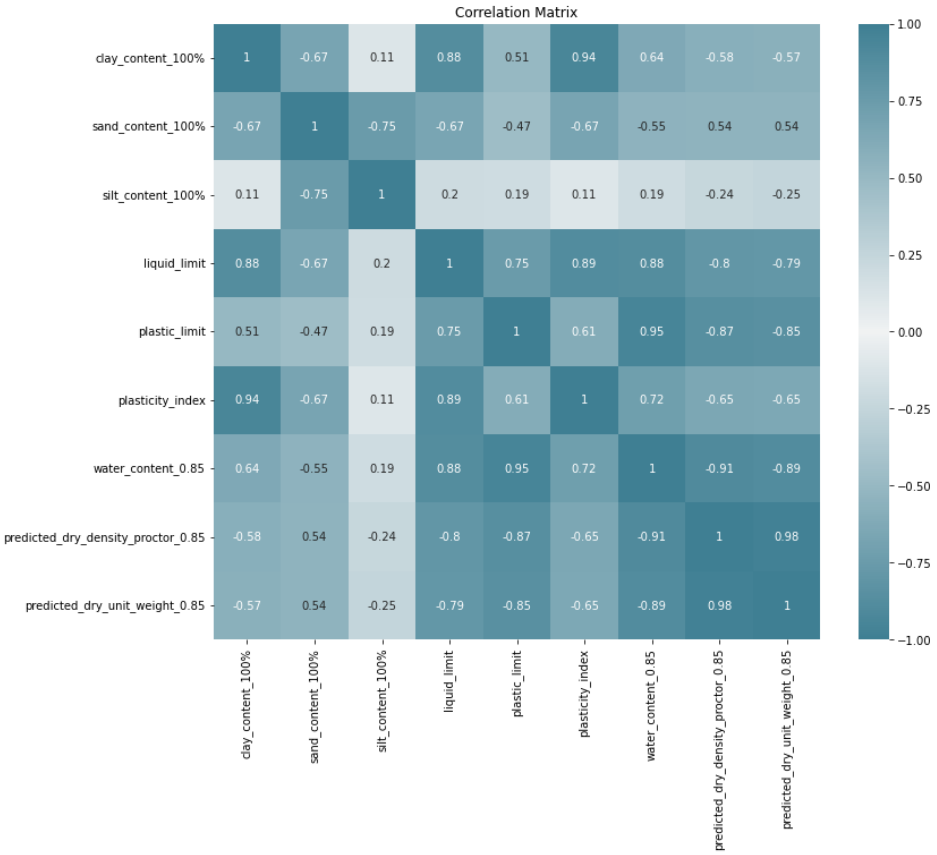


Figure 61: Linear correlation in the COMARF Data frame

8.5 Comparison with the literature-based correlations

The previous subsection shows that Clay content and liquid limit correlate well. Figure 62 shows the performance of RF against the original data and found correlations in the literature. The figure shows that RF predicted the values of clay content well using the second scenario (knowing Atterberg limits). Furthermore, the linear correlations show promising performance; however, the equation does not capture the spread of the original data, giving the advantage to the ML performance. On that note, it is not mentioned which clays Polidorli, 2007 studies, as the author mentioned many constants to capture a wide spread of clay types. However, the correlations found in the literature could not predict the values of the data, as they either over- or underestimated the clay contents' variables, making them unusable in this context of Dutch clays. More on this is mentioned in Section 4

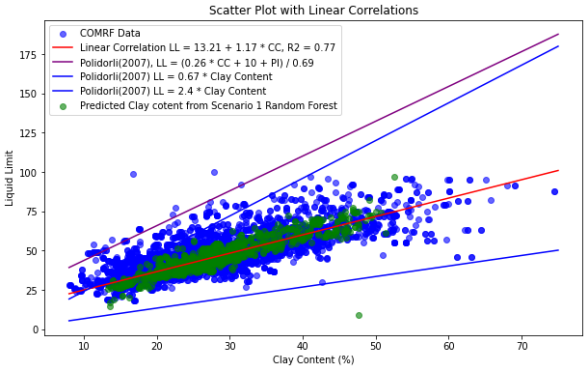
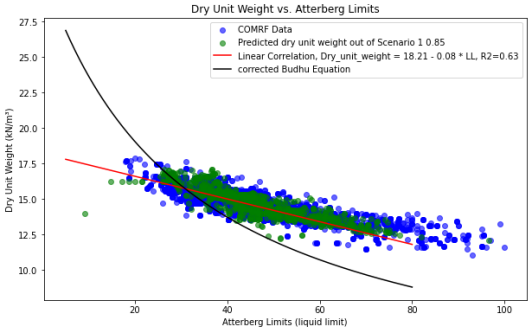


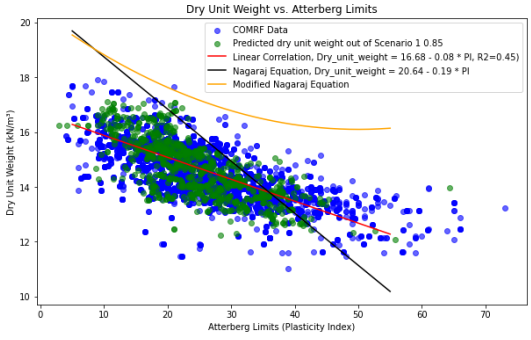
Figure 62: Scatter plot LL and CC

Furthermore, RF also better predicted the dry unit weight using the Atterberg limits, as shown in Figures 63a, 63b & 64b. The linear equation does predict fit the data, indicating reliability in their predictions. On the other hand, some literature equations do have some reliability in their predictions in some determined intervals. In the case of the Dry unit weight and liquid limit, Figure 63a. Then, it is possible to use the equations for liquid limits between 25 and 35, as the black lines show. However, it is necessary to mention that the equation slightly underestimates the variable.

Moving to Figure 63, then the equation of Nagaraj (developed for natural soils) could also be used to predict the dry unit weight for plasticity index values between 20 and 40. However, this equation overestimates the variables. On the other hand, the modified equation of Nagaraj did not fit the data, even though it was modified to account for more sand fraction in the data, which was the case in some soil samples in this study. This shows that Dutch soil behaves differently from natural soils collected from various locations. More on these soil types can be found in Nagaraj correlations



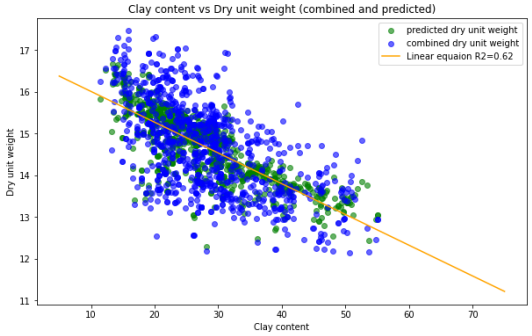
(a) Scatter plot dry unit weight and liquid limit



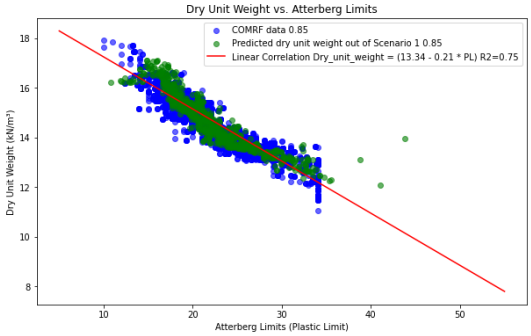
(b) Scatter plot dry unit weight and plasticity index

Figure 63: Comparison of dry unit weight and Atterberg limits (liquid limit and plasticity index)

Unfortunately, no correlations in the literature capture the correlations between the clay fraction and dry unit weight. However, a moderate/weak correlation was found in the data of this research with an R2 value of 0.62, shown in Figure 64a. This Figure shows that the reconstructed and the predicted data have a wide scatter, indicating a weak correlation between the two variables. However, as indicated previously, RF had an R2 value of 0.75, which is considered a moderate correlation, producing less noisy data than the reconstructed one, as the figure shows.



(a) Scatter plot dry unit weight and clay content



(b) Scatter plot dry unit weight and plastic limit

Figure 64: Scatter plots of dry units, plastic limit, and clay content.

Upon examining the distribution of the actual and predicted data, it is evident that the variability introduces uncertainty when making predictions based solely on linear correlations, representing average values. The original data show a wide range of potential values for each variable (such as Atterberg limits or clay contents). Therefore, based on the graphs, if more precise predictions with lower uncertainty are required, the Random Forest model effectively accounts for this variability. This uncertainty is further explored in Section 9, where numerical results are presented.

8.5.1 Summary: performance of ML

- Auto-encoder had an excellent performance on the data frame CI and CC individually and on the combined data frame. This made the reconstructed clay categorizations more centered and enhanced the overall performance of NN and RF.
- Scenario 1 [Atterberg limits] had the best performance followed by the plasticity graph classes scenario. This was apparent in the CI data frame individually and combined. Using the water content to predict the CC variables also showed excellent results.
- Random forest had better performance combing the data and predicting the variables using the mentioned scenario. Furthermore, RF was significantly less computationally expensive than NN, making RF superior to NN in this research.
- COMARF (combined reconstricted random forest) data frame had the best prediction, making this data frame more useful for prediction and the other tests data frames.
- Internal linear correlation only captured the average of the needed variables. However, they did not capture the uncertainty of the variables.
- Literature correlations did not capture the data used in this research, making them unusable for predictions.
- RF showed promising results capturing the uncertainty of the data.

9 Uncertainty of the results:

Research Question 4: What is the level of uncertainty of the found correlations?
Since the COMARF showed the best results when training the ML code on it. The following section shows this data frame's R-squared, MSE, MAE, and standard deviation results. This will be done for scenarios one (Atterberg limits), four (plasticity diagram classifications), and a new scenario, combining the previous two scenarios. This is done to mimic the order of the current practices. As explained in Section 2.3, the current practice determines the clay class, and then, based on a consistency index, average values are estimated. Therefore, this new scenario determines the clay class according to the plasticity diagram class & the Atterberg limits and shows results depending on these two known information of the clay sample.

9.1 Scenario 1 (Atterberg limits):

As previously mentioned, this scenario has the best results among all the others. Based on Table 30, the RF model shows strong performance across most variables. The R-squared is generally high, indicating that the model can explain a significant portion of the variance in the dependent variable. The MSE and MAE values are relatively low, suggesting accurate and minimal error. On that note, the Clay content variable shows strong R-squared and MAE values; however, the high standard deviation 8.26) indicates that the model's predictions might be less consistent for some clay values. Furthermore, the model does not perform well on the dry proctor density scenarios as the model exhibits high values, indicating a wide spread of the predicted values. On the other hand, the model shows excellent metric values for the dry unit weight, which is the most important one for initiating a stability analysis.

Table 30: Performance metrics for different variables using scenario 1

Variable	R2 Score	MAE	MSE	Std.
Clay content	0.97	0.97	2.30	8.26
Sand content	0.74	5.70	55.37	11.46
Silt content	0.50	5.67	51.89	6.11
Water content 0.75	0.99	0.23	0.31	5.87
Water content 0.60	0.99	0.28	0.47	6.69
Water content 0.85	0.99	0.21	0.18	5.18
Dry density proctor 0.75	0.93	23.31	933.54	106.09
Dry unit weight 0.75	0.91	0.25	0.11	1.00
Dry density proctor 0.60	0.92	21.15	940.21	102.99
Dry unit weight 0.60	0.90	0.24	0.11	0.96
Dry density proctor 0.85	0.91	25.24	1156.60	111.33
Dry unit weight 0.85	0.89	0.28	0.14	1.04
Tri class	0.85	0.26	0.20	0.98
Plasticity diagram class	0.94	0.18	0.15	1.42

9.2 Scenario 4 (plasticity diagram class):

Following a similar story of the previous scenario, this one indicates weaker model performance across the variables. This is apparent in the standard deviation and R-squared values. However, the critical thing to note here is that the standard deviation values of the dry unit weight are low, indicating a little spread of the predicted data regardless of the moderate R-squared values Table 31 indicates.

Table 31: Performance metrics for different variables using scenario 4

Variable	R2 Score	MAE	MSE	Std.
Liquid limit	0.83	3.81	23.17	10.54
Plastic limit	0.62	1.94	6.52	3.58
Plasticity index	0.71	3.27	20.50	7.26
Clay content	0.70	3.69	21.54	7.23
Sand content	0.48	8.28	111.71	9.73
Silt content	0.13	7.61	91.17	3.57
Water content 0.75	0.73	2.13	8.14	5.14
Water content 0.60	0.74	2.41	10.65	5.90
Water content 0.85	0.70	2.01	6.89	4.44
Dry density proctor 0.75	0.64	53.89	4479.46	88.75
Dry unit weight 0.75	0.62	0.53	0.44	0.84
Dry density proctor 0.60	0.62	48.47	3967.49	86.98
Dry unit weight 0.60	0.61	0.46	0.36	0.82
Dry density proctor 0.85	0.59	54.16	4986.18	91.63
Dry unit weight 0.85	0.57	0.53	0.48	0.85
Tri class	0.52	0.64	0.61	0.83

9.3 Scenario combined (1 & 4):

Combining both scenarios does not improve the values compared to the first scenario's results. This indicates a more significant influence of the Atterberg limits on the results than the plasticity diagram scenario. However, this scenario was added to compare it with the current variable estimation practice, which is explained in Section 9.4.

Table 32: Performance metrics for different variables using the combined scenario

Variable	R2 Score	MAE	MSE	Std.
Clay content	0.97	0.98	1.81	8.33
Sand content	0.74	5.88	61.14	11.51
Silt content	0.49	5.76	55.61	6.07
Water content 0.75	0.99	0.25	0.21	5.87
Water content 0.60	0.99	0.28	0.38	6.70
Water content 0.85	0.99	0.24	0.15	5.17
Dry density proctor 0.75	0.93	24.07	1046.76	106.05
Dry unit weight 0.75	0.90	0.27	0.13	1.00
Dry density proctor 0.60	0.92	19.69	745.17	102.95
Dry unit weight 0.60	0.90	0.23	0.09	0.96
Dry density proctor 0.85	0.91	24.05	1079.55	111.05
Dry unit weight 0.85	0.88	0.27	0.14	1.04
Tri class	0.84	0.28	0.25	0.99

9.4 Implications on the stability analysis:

The following section compares the current estimated unit weight values and those found in the research according to the forth scenario. This is important as this difference will affect the results of the stability analysis applications. Furthermore, the section will elaborate on the differences between the clay classes or erosion categories (vertically) and between the consistency index measures (horizontally). This showed to what extent these variations are essential for designers in the design phase of a dike project. On that note, it is necessary to mention that the unit weights were not predicted directly. The values were calculated from the predicted water contents and dry unit weights.

Reminder of the current practice:
As explained in Section 2.3, the current practices estimate the unit weight of clays by first defining the admixture of the clay. This could be clean, weak sandy, strong sandy, or organic. Then, the current practices categorize it into three consistency states: weak, moderate, and fixed/solid. Then, the unit weight is estimated, summarized in Table 33 (shown again for easier access). On that note, it is noticeable that the values of clean clay are between 14 and 20 [kN/m3], weak sand between 20 and 21 [kN/m3], and strong sandy unit weight values are between 18 and 20 [kN/m3].

Table 33: Current practice values(double for easier access to the reader)

Admixture	Consistency	unit weight
Clean	Weak	14
	Modeate	17
	Fixed/solid	19 or 20
weak sandy	Weak	15
	Moderate	18
	Fixed/solid	20 or 21
strong sandy	-	18 or 20

As explained in Section 9, since the predictions of the third scenario were weak, this study could not produce reliable results following the same order. Hence, the combined scenario defines the clay class depending on the plasticity diagram classifications along the Atterberg limits' respective values (liquid and plastic limits). This was also done using three consistency indices: 0.60, 0.75, and 0.85. Table 35 shows the average predicted values of these unit weights with the predicted respective max and min values in this research.

Predicted data count:
Before delving into the predicted values, it is essential to note that using K-fold affected the predicted data count. While the original size of the dataset is around 4406 rows of data, the predicted dataset is equal to 882 rows of data (features), Section 7.3.1. Furthermore, some clay classes did not have as many data points as those in the plasticity graph diagram, Figure 21. Therefore, some clay classes did not have enough predicted data points to make them reliable for analysis. This is seen in Table 34, where classes CIL-SIL, SIL, and SIV had only a few points predicted for them. Therefore, while their respective unit weight values are shown in the following tables, these classes are excluded from the comparison.

Table 34: Count of Predicted Unit Weights by Clay Class and Erosion Categories (EC)

Class	CIH	CIL	CIL-SIL	CIM	CIV	SIH	SIL	SIM	SIV	EC1, EC2, EC3
Count	231	117	1	438	34	22	1	17	5	34, 438, 161

Current vs. Predicted Unit Weights of plasticity diagram classes:
Comparing the predicted values with the current estimations shows differences in the range limits of both approaches. First, it is noticeable that the smallest predicted value for the different clay classes is 16.94 [kN/m3], which is 17% greater than the estimated value for a weak clean clay (equals 14

[kN/m3]). On the other hand, the max predicted values did not score more than 20.10 [kN/m3], which is 4.5% smaller than the maximum estimated unit weight of the used practices, which is 21 [kN/m3].

However, when comparing the averages with the current estimation, the lowest average value is 17.31 [kN/m3], which is 23% greater than the lowest estimated value (14 [kN/m3]) of the current practices. On the other hand, these average values scored similar to the moderate estimation values of the current practices, which is either 17 or 18 [kN/m3] (depending on the clay type). This is noticeable since the predicted values are between 17.31 and 19.31, which is between 1 - 11 % difference compared to 17 [kN/m3].

Differences between the plasticity diagram clay classes:
Looking at the average values of the clay class individually, class CIV has the lowest values, with averages of 17.31, 17.54, and 17.65 [kN/m^3] for consistency indices of 0.60, 0.75, and 0.85, respectively. Furthermore, class CIL exhibits the highest values: 18.72, 19.04, and 19.31 [kN/m^3] for the same indices.(both are *Italic* in Table 35. The difference between the lowest and highest average unit weight is approximately 1.5 [kN/m^3], or 8%. This gap narrows to 5% when comparing the maximum and minimum predicted values.

Table 35: Unit Weights of Different Clay Classes (Average, Maximum, and Minimum)

	Average Unit Weight			Maximum Unit Weight			Minimum Unit Weight		
Class	0.60	0.75	0.85	0.60	0.75	0.85	0.60	0.75	0.85
CIH	17.80	17.88	17.92	18.62	18.64	18.63	16.97	17.11	17.30
CIL	*18.72*	*19.04*	*19.31*	19.91	20.10	20.09	17.84	17.36	17.64
CIL-SIL	18.99	18.70	18.59	18.98	18.70	18.59	18.98	18.70	18.58
CIM	18.00	18.28	18.58	19.16	19.17	19.94	17.51	17.52	17.26
CIV	*17.31*	*17.54*	*17.65*	19.24	19.19	18.21	16.94	17.09	17.29
SIH	17.54	17.75	17.82	18.49	18.74	18.40	17.06	17.15	17.23
SIL	19.60	19.56	19.80	19.59	19.56	19.79	19.59	19.56	19.79
SIM	17.71	17.67	17.76	17.97	18.73	18.61	17.47	17.30	17.24
SIV	17.39	17.96	17.50	17.73	19.07	17.77	17.09	17.18	17.35

Note: 0.60 = min. consistency index (in dike core), 0.75 = min. in dike revetment, 0.85 = best guess to match the CC data

Differences between the Erosion category classes:
The clay classes were grouped into three erosion categories (EC) by converting plasticity diagram classes using an IF function in Excel, as shown in Table 36.

The average unit weight values vary across the erosion categories, with EC1 having the lowest values and EC3 the highest. For a consistency index (CI) of 0.85, EC1 has an average unit weight of 17.64 [kN/m^3], compared to 18.87 [kN/m^3] for EC3, reflecting a difference of 1.23 [kN/m^3], or approximately 6%. This trend is consistent across the other consistency indices (CI 0.60 and 0.75).

On that point, the ECs' results are similar to those of the plasticity classes when compared with the results of current practices. The lowest estimated unit weight (16.94 [kN/m3]) is 16% higher than the lowest current unit weight value (14 [kN/m3]). Similarly, the lowest average unit weight (17,30 [kN/m3]) is 19% higher. Furthermore, there is a 1.57 [kN/m3] difference between the lowest and the highest average, which is around 8% difference. Furthermore, these two values score around 1-9% higher difference than 17 [kN/m3](the moderate value of the current estimation).

However, these results are influenced by the predicted data count, which varies significantly across the categories. EC2 has the largest dataset with 438 points, while EC1 and EC3 have only 34 and 161 points, respectively, Table 34. This discrepancy affects the distribution of values, leading to cases where EC2 shows higher minimum unit weights than EC3, as the *Italic* numbers indicate in Table 36).

The maximum unit weight values also follow a similar trend. The difference between EC1 and EC3 ranges from 1% to 9%, depending on the consistency index. For the CI 0.60 measure, the difference is only 1%, while for the CI 0.75 measure, it increases to 4%.

Table 36: Unit Weights of Different Erosion Categories (Average, Maximum, and Minimum)

	Average Unit Weight			Maximum Unit Weight			Minimum Unit Weight		
EC	0.60	0.75	0.85	0.60	0.75	0.85	0.60	0.75	0.85
EC1	17.30	17.53	17.64	19.24	19.19	18.12	16.94	17.09	17.29
EC2	17.98	18.26	18.56	19.16	19.17	19.94	*17.51*	*17.52*	*17.26*
EC3	18.39	18.67	18.87	19.59	20.10	20.09	*17.06*	*17.15*	*17.24*

Note: 0.60 = min. consistency index (in dike core), 0.75 = min. in dike revetment, 0.85 = best guess to match the CC data

Difference between the consistency index measures:
The general trend is that the higher the consistency index, the greater the predicted value for the unit weight. This is clear from the average values for both the ECs and the clay classes. However, the predicted data court has also been affected by showing greater unit values for some 0.75 predicted unit weight values, similar to the situation explained with the erosion categories.

Regardless of this inconsistency in the results, the difference between the consistency index measures does not significantly affect the results, as does between the clay classes or the erosion categories, where the difference between the CI measures is between 1-3%.

Influence on stability analysis:
The previous results enlighten a designer on the variations between the clay classes when testing the safety of a dike design. While the new predicted variable shows more variety in the unit weight values, the old estimation practice seems rougher and less accurate than the newly found results. This means that a designer must be more precise when selecting a clay class and, therefore, a unit weight value for D-stability (the software being used for stability calculations), as this will reflect on the resulting safety factor. Lastly, the standard deviation of the dry unit weight of the clay classes and erosion categories are shown in Appendix I.

9.5 Summary: Uncertainty & implication on the stability analysis:

- A combined scenario was added to have a comparable estimation order as the current practices, which was done by combining the plasticity graph classes with the respective Atterberg limits (liquid and plastic limits).
- predicting unit weight and water content measures showed low uncertainty as indicated by the standard deviation, MSE, and MAE values. However, proctor dry density showed more uncertainty, making the prediction slightly less reliable.
- Current practice estimates unit weights based on clay type (clean, weak sandy, strong sandy) and consistency (weak, moderate, fixed/solid) with clean clay. In contrast, the new ones depend on the plasticity clay class or the erosion category.
- Average predicted values (17.31–19.56 [kN/m^3]) align with moderate current practice values. However, they are 19% higher than the lowest current estimated value. Clay class CIV had the lowest average unit weight (17.31–17.65 [kN/m^3]), while CIL had the highest (18.72–19.31 [kN/m^3]).
- Similar differences are found between the erosion categories (EC1, EC2, EC3) with Ec2 scoring the highest unit weights. These Ecs score similar differences as the clay classes compared to the current results.
- Differences between estimated and research-derived unit weights impact the stability analysis for dike design.

10 Discussion:

The following section reflects on the research process and the steps taken to achieve the results. It begins by evaluating data quality, addressing the first research sub-question: **What is the state of the data?** Next, it examines the internal correlations within the data, answering the second sub-question: **How are these data internally correlated?** The third sub-question, **To what extent can ML models predict specific variables?**, is then discussed, focusing on the effectiveness of the machine learning models and the challenges encountered. Lastly, the reflection covers the performance of these models, particularly in relation to uncertainty, addressing the final sub-question: **What is the level of uncertainty in the identified correlations?**

10.1 Data quality:

The original plan was to find documented data containing the clay inspection results with compaction and triaxial tests. However, because the data reported were not labeled, it was challenging to find clay data with the corresponding tests mentioned. For this reason, it was decided to collect the tests individually and use ML to find correlations between the data sets. This resulted in three data frames: clay inspection (physical variables), compaction, and triaxial (mechanical variables).

Both clay inspection and compaction data are of good quality. The clay inspection test data set contained 4406 rows of data, while the clay compaction and triaxial test data sets contained 2463 and 130 data rows, respectively. This made the clay inspection and clay compaction datasets suitable for ML applications. On top of that, the distribution of the variables of these data sets were either Gamme or Lognormal distributions, which fitted the normal physical ranges of the variables. This is apparent in Tables 14 & 15.

On the other hand, the triaxial test data had scattered data, making its histograms inconsistent; figures are shown in Appendix B. Therefore, because this dataset had a low data count and lousy quality, the data set was considered unsuitable for ML applications. This was also the case for the internal linear correlation in the dataset, which scored weak linear correlations. Furthermore, this data frame was also tested on the machine learning models. However, the models could not predict triaxial variables because of their quality, making the dataset unreliable for both applications.

The triangular classifications of clay inspection data consisted mainly of different silty clays and a few sandy ones, which gave more advantages to silty clays when processing them in machine learning. Furthermore, plotting the clay inspection data into the plasticity diagram classification resulted in most clays being in erosion category two. This is to be seen when comparing Figure 21 with Figure 12a, which shows that most of the data spread over the A line (red line). Furthermore, the reported classification of the triaxial dataset did not cover all the classes in the clay inspection dataset, weakening the quality of this dataset even more.

Furthermore, comparing the triangular classification with the plasticity diagram classification shows that old classifications spread over many new classes, which indicates the physical difference of both approaches; this might be the reason why the old classification system (triangular) is not used anymore, as such a comparison shows the old system's inaccuracy. This was also demonstrated when the new system's class (plasticity diagram) was used for prediction, as it outperformed the old one.

10.2 Linear correlations of the original data:

Most original data scored weak to moderate linear correlations, as shown in Section 6.4. Typically, the best R-squared values of the CI data frame scored between 0.6 and 0.76, the latter being the correlation between plasticity index and clay content. In clay compaction, good correlations were only found between the data's water content and unit weight, which scored between -0.86 and -0.93 for unit and dry unit weights, respectively.

Lastly, the Triaxial data only scored weak correlations between the variables, except for water content and dry unit weight, similar to the clay compaction data. Therefore, the triaxial test data added no new findings to the previous data frames.

10.3 Training & performance of the Machine learning models:

The Neural Networks and Random Forests models performed well, combining the clay inspection and compaction data. However, Random Forests captured the spread of the data and predicted new unseen data better than Neural Networks. This was shown in Figure 48 until Figure 59. This produced two different data frames, COMNN and COMRF, which stand for combined Neural Network data frame and its counterpart. These 2 data frames were then reconstructed, resulting in two new data frames, COMANN and COMARF, which stand for Combined reconstructed Neural Network data frame and its counterpart.

On that note, the ML models were trained using a set of four scenarios. Scenario one indicates knowing the Atterbeg limits; scenario two indicates knowing the clay content of clay samples, while scenarios three and four indicate knowing the triangular classifications and the plasticity diagram classifications, respectively.

On that note, Scenario 3 (knowing the clay contents) scored weakly for the following reasons: first, the clay samples were documented using clay, silt, sand, salt, organic, and mass loss contents, while the triangular classifications only used the clay, silt, and sand content, which were then considered as 100% of the total content. This differs from the documented data as these three variables scored around 85-90% (organic, salt, and mass loss scored around 10-15%). This broke the original ratios of the data, which might have caused the predictions for this scenario weak.

This led to Scenario One (Atterberg limits) having the best predictions across all the variables. It is essential to say that it is logical that this scenario had the best results because of the way the variables were produced. As mentioned in Section 5.3.4, the water contents 0.60, 0.75, and 0.85 were produced using Equation 2, which used the Atterberg limits & consistency index. These Atterberg limits became the connecting variables for most others. This strengthened the relationships between the variables and these limits, allowing the ML to efficiently capture these correlations and improve the accuracy of the first scenario's predictions.

Two scenarios were combined, namely, scenarios 1 and 4, which combined the Atterberg limits and plasticity diagram classifications. This led to a similar estimation of the unit weight depending on the clay class and Atterberg limits (similar current estimation process). However, it differs, as the clay class belongs to the PI diagram classes, which use the clay and silt contents, while the current practice uses the clay and sand content.

Furthermore, all scenarios showed the best results using the COMARF data frame and the Random Forest model. Random forests outperformed Neural Networks, combining the data frames and predicting the values. However, since auto-encoders are a form of neural networks, the architecture of neural networks (in this case, Auto-encoders) enabled the research to capture the relations between the data better and produce better predictions by removing noise from the data, giving them a significant role in this study.

10.4 Comparison with the literature-based correlations and linear correlations:

The linear correlations applied to the combined data frame showed moderately good correlations between the [liquid limit & clay content], [liquid limit & dry unit weight], and [plasticity index & dry unit weight]. However, Section 8.5 shows clearly that the ML models outperform the linear correlation models. This was also applicable in the case of the [clay content & dry unit weight] despite not finding any correlations in the literature, where this research found moderate correlations between the two. On that note, machine learning models offer a better way of predicting the variables with less uncertainty. The linear correlation predictions are suitable only for an average value but unreliable enough for stability analysis estimations as they are rough and do not capture the spread of the data (more uncertainty).

10.5 Uncertainty of the Results & implications of stability analysis:

Section 8.4.1 shows the performance of the scenarios used in this research, where scenario 1 was the best-performing scenario. Furthermore, Section 9 shows the standard deviation values of scenarios 1

(Atterberg limits) and 4 (plasticity graph class), along with the uncertainty of the results when both these scenarios are combined.

Most of the predictions had acceptable uncertainties. Some variables scored higher uncertainty, like the variations of the dry density proctors, which show a standard deviation of around 100. This was the case for the first and the combined (1&4) scenarios. Furthermore, the difference in the reported MSE and MAE values shows that there are outliers in the prediction, as the MSE values (which puts more weight on the outliers) are more significant than the MAE ones.

Furthermore, the water content and dry unit weight show excellent standard deviation, MSE and MAE scores for these three scenarios. However, scenario 4 (plasticity graph class) shows weak to moderate results for the water content and dry density weight, where they scored R-squared values around 0.70 for the water content predictions and around 0.65 for the dry unit weight, making this scenario not as reliable for predicting as the first (Atterberg limits) or the combined one (1&4).

For this reason, using the combined scenario, it was possible to reliably predict unit weights for the different clay classes. This was shown in two distinct categories. The first category shows the unit weight results according to the different plasticity diagram classes. The second shows the unit weight results according to the erosion categories. Both scored different unit weight values for each consistency index measure.

This prediction reflects the variance between the reported/used values for the unit weight and the results found in this research. While Table 3 reported values of 14-15 for a weak consistency index, 17-18 for **a moderate consistency index (CI)**, and 19-21 for a fixed consistency index, the new prediction is more accurate and cover a broader range of clay classifications that are also a part of the new classification system (plasticity graph classes).

Table 35 presents new predicted unit weight values for these classes, showing significant differences from current estimates. The minimum predicted unit weight is 17% higher than the lowest current estimate. Additionally, the lowest average predicted unit weight is 23% greater than the current average. However, the maximum predicted unit weight difference is less pronounced, as these values are closer to the existing estimates. Furthermore, the average predicted values align around the **moderate (CI) current estimation unit weights** with a 1-11% greater predicted unit weights, Section 9.4.

Furthermore, Table 36 shows the difference in results between the erosion category classes, which shows a similar difference to the minimum current unit weights as the clay classes. Additionally, when comparing the average predicted values with **moderate (CI) current unit weights**, the difference is between 1-9% more significant. This gap narrows down to 5% when looking at the maximum and minimum predicted values.

The previous findings were also affected by using the K-fold method (which was necessary for training the ML methods); this method lowered the number of predicted variables to 1/5 of the original data count, resulting in 882 predicted values out of 4406 data points. This made the results of some clay classes unreliable for the estimation. However, the original count of these clay classes was also low, making the issue more of a data quality than the used method. These clay classes were not considered for the comparison as discussed in Section 9.4.

Furthermore, using different consistency index measures did not result in a difference as significant as using different clay classes or erosion categories, as this difference was between 1 - 3% between the predicted unit weight values. This could result from the implemented correction after combining the clay inspection and clay compaction data frame, where the lower the upper limits of the dry unit weight and proctor dry density were corrected by the limits of the original clay compaction data frame, check Section 8.3.3.

On that note, combining the found R-squared results with the uncertainty measures (standard deviation, MSE, and MAE) makes it reliable to say that this research could produce more accurate unit weight values than those reported in the current estimations.

Therefore, combining the standard deviation results with the predicted maximum & minimum values along with the found distributions in Tables 14 & 15 (mostly fit the Log normal distribution) will

result in having more reliable stability analysis in the design phase than using the current estimation for the unit weight. In contrast, as the available triaxial data limited the research, it was impossible to predict the friction angle, making it necessary to rely on the current estimation methods for this variable.

11 Conclusion:

The Netherlands is a hub for different dike projects; these are the first defense against water nationwide. Therefore, designers are always searching for ways to make their designs more reliable for future risks. At that point, these designers do not have the clay available for testing during the project's design phase. Therefore, they rely on regional knowledge and experience to estimate the needed variables for making a dike design, mostly rough estimations. This research studied implementing Machine learning models to estimate better clay variables used for stability calculations, which answers the main research question: **How can the dike embankment soil variables for dike stability calculations be better predicted based on borrowed source information available in the design phase (before construction)?**

The research used Dutch data from three soil tests: clay inspection, compaction, and triaxial tests. It intended to collect data from projects with all three soil tests available and then correlate them with each other, which was not the case as most of the collected data was not labeled. Furthermore, the triaxial data collected suffered from low data count and low quality, making it unusable for ML and regression applications.

Machine Learning (ML) models, including Neural Networks and Random Forests, were also trained on individual datasets. These datasets were later combined into a single data frame, using water content as a common variable. While water content was reported in the clay compaction tests, it had to be estimated using three consistency index values (0.60, 0.75, 0.85) for the clay inspection dataset. Therefore, the combined dataset had predicted dry unit weight and Proctor dry density for each consistency index. Furthermore, Auto-encoders were applied to the dataset to enhance the ML models' ability to capture relationships between the variables.

Training the ML models was done using a set of four scenarios, each representing what a designer might know about the clay sample in the design phase. Scenario 1 uses the Atterberg limits as known variables, while scenario 2 uses the clay's contents. Scenarios 3 and 4 use the triangular and plasticity diagram classifications as known variables, respectively. The research found that using Scenario 1 yielded excellent results, while Scenario 4 yielded moderate prediction metrics.

Because the current estimation uses a clay class, scenarios 1 and 4 were combined into a new scenario, which resulted in excellent prediction results for dry unit weight and water content measures. This made it possible to predict the unit weight according to the plasticity graph classes, which were also turned into erosion category classes.

The study's results for the unit weight predictions proved more accurate than current regional knowledge or practice values. The minimum predicted unit weight was 17% greater than the lowest current estimation, and the average prediction was 23% higher. Furthermore, the average predictions (for both the clay classes and the erosion categories) were between 1-11% more significant than the average used value for a moderate consistency index. These findings suggest that current practices rely on rough estimates that can impact the accuracy of stability calculations.

Regardless of the limitations faced in this research, this study shed light on promising results. Firstly, it showed that using Machine learning models to predict clay variables for stability analysis (unit weight) is valid. However, this did not follow the order of the current practices, where the type of clay is mentioned (according to its clay and sand admixtures). The new approach shows that it is possible to predict according to the plasticity diagram classes, which depend on the clay and silt content.

In conclusion, this research provides greater certainty (than the current practice) in the unit weight design value when compaction and triaxial tests are unavailable during the design phase. It also demonstrated the effectiveness of predicting Proctor dry density, although with slightly lower accuracy. However, predictions of the friction angle were less successful due to limitations in the triaxial data.

Ultimately, designers can leverage these predictions, incorporating unit weight distributions (average, maximum, minimum, and standard deviation) to achieve more accurate stability analyses than current estimation methods allow.

11.1 Recommendations:

This research showed good performance of machine learning models, with unexpected results, such as Random Forest outperforming Neural Networks. It also showed that variables can be predicted using a lower level of knowledge of the soil sample. On that note, this research also revealed a space for improvements that could take place in the future, which should enhance the performance of the ML models to give better insights into the data correlations. On that note, the following section suggests action that could be taken on both the academic and the practical levels.

Acadamic/Research level:

Firstly, data collection could be enhanced by adding the locations of the data. This should allow for new categorizations in the data frames, like directional categorizations, such as East Dutch and Central Dutch clays, which might show different correlations. Furthermore, the research lacked Triaxial data, which was not used because of its low quality. Therefore, searching for triaxial data from other resources is recommended so that clay inspection, compaction, and triaxial data frames can be used to have one combined data frame. Then, this global data frame would be used to predict models linearly or using machine learning models.

The triangular classifications of the clay could be defined using the soil sample's different contents, including salt, organic, and mass loss contents. This might preserve this class connection to the other variables.

Furthermore, testing other machine learning models, like Bayesian neural networks, is viable. On that note, start with a simple model and increase the complexity if needed. Support vector Machines could be skipped as they did not perform well in this research.

On that note, it would be perfect to have enough data reports that contain the clay inspection, clay compaction, and triaxial test data, which will give more insights into the actual relationships of the variables, other than depending on machine learning to combine the data frames. Furthermore, it might be a good idea to have a separate study that focuses only on triaxial data as it is clear that there is a massive shortage of these data types owing to the complexity of the test and the fact that it is resource- and time-consuming.

If the previous approach is not possible, then generating triaxial data with statistics of actual data might also be an option. This was not done in this research as the quality of the collected triaxial data did not allow for that and had a broad variation, making such an approach undesirable/unreliable.

Practical level:

The research found variations in the unit weight depending on clay classes and erosion categories across different water content measures. Furthermore, this variation in the unit weight was accompanied by max, min, and standard deviation values. The original data could also be fit in either a Gamma or log-normal distribution. Therefore, by combining the previous results, a designer can better estimate the sensitivity of macro stability analysis to the found variations in the unit weight by applying both a deterministic and a probabilistic approach. This will show the impact of the unit weight variable on the safety factor of the dike design and/or the fragility curve (depending on whether the test is done deterministically or probabilistically).

References

Highway Research Board. (1962). *Compaction and correlation between compaction and classification data*. National Academy of Sciences-National Research Council.

Delft. (1996). Clay for dikes. www.minvenw.nl/dww/home/

O'Sullivan. (1997). *Nutrient disorders of sweet potato. aciar monograph no. 48, australian centre for international agricultural research, canberra, 136 p.*

Lubking. (2000). Grondgedrag feiten, norman en waarden met betrekking tot grond in de praktijk van de geotechniek.

Polidorli. (2007). *Relationship between the atterberg limits and clay content.*

Budhu, M. (2011). Soil mechanics and foundations.

Knappett, J., & Craig, R. (2012, February). *Craig's Soil Mechanics, Eighth Edition*. CRC Press.

Akayuli, C. F. A., Ofosu, B., Nyako, S. O., & Opuni, K. O. (2013). The influence of observed clay content on shear strength and compressibility of residual sandy soils. *International Journal of Engineering Research and Applications (IJERA)*. http://csirspace.csirgh.com:80/handle/123456789/810

Gurtug. (2015). *Prediction of compaction behaviour of soils at different energy levels.*

Nagaraj, H., Reesha, B., Sravan, M., &, S. (2015). Correlation of compaction characteristics of natural soils with modified plastic limit. *Transportation Geotechnics*, *2*, 65–77. https://doi.org/10.1016/j.trgeo.2014.09.002

Ameratunga, J., Sivakugan, N., & Das, B. M. (2016, January 1). *Correlations of soil and rock properties in geotechnical engineering*. https://doi.org/10.1007/978-81-322-2629-1

Ahmed, S. M. (2018). Assessment of clay stiffness and strength parameters using index properties. *Journal of Rock Mechanics and Geotechnical Engineering/Journal of rock mechanics and geotechnical engineering*, *10*(3), 579–593. https://doi.org/10.1016/j.jrmge.2017.10.006

Becker, D. (2018). *Rectified linear units (relu) in deep learning*. Retrieved May 7, 2018, from https://www.kaggle.com/code/dansbecker/rectified-linear-units-relu-in-deep-learning

Halter, W., et al. (2018). *Handboek dijkenbouw* [This handbook was made possible and published by the Hoogwaterbeschermingsprogramma (HWBP), Postbus 2232]. Hoogwaterbeschermingsprogramma (HWBP).

Ali et al. (2019). *A correlation between compaction characteristics and soil index properties for fine-grained soils.*

Jasim, M. M., Al-Rumaithi, A., & Al-Khaddar, R. M. (2019). Prediction of bearing capacity, angle of internal friction, cohesion, and plasticity index using ann(case study of baghdad, iraq). *International Journal of Civil Engineering and Technology (IJCIET)*, *10*, 2670–2679. http://www.iaeme.com/IJCIET/index.asp2670http://www.iaeme.com/ijciet/issues.asp?JType=IJCIET&VType=10&IType=1http://www.iaeme.com/IJCIET/issues.asp?JType=IJCIET&VType=10&IType=1

Katuwal, S., Knadel, M., Norgaard, T., Moldrup, P., Greve, M. H., & de Jonge, L. W. (2020). Predicting the dry bulk density of soils across denmark: Comparison of single-parameter, multi-parameter, and vis–nir based models. *Geoderma*, *361*, 114080. https://doi.org/10.1016/J.GEODERMA.2019.114080

Van der Meij. (2020, March 3). *D-stability, installation manual, tutorial and scientific manual*. https://open.rijkswaterstaat.nl/open-overheid/@50330/functioneel-ontwerp-stability-versie/

Arama, Z. A., Yücel, M., Akın, M. S., Nuray, S. E., & Alten, O. (2021). Prediction of soil plasticity index with the use of regression analysis and artificial neural networks: A specific case for bakırköy district. *6th International Conference on Harmony Search, Soft Computing and Applications*, 281–893.

Kumar, A. (2021). *Neural network*. Retrieved December 2021, from https://towardsdatascience.com/neural-network-74f53424ba82

Raad. (2021). *Factors influencing the shear strength of clays: A review.*

Shimobe, S., & Spagnoli, G. (2021). Relationships between strength properties and atterberg limits of fine-grained soils. *Geomechanics and geoengineering*, *17*(5), 1443–1457. https://doi.org/10.1080/17486025.2021.1940317

WBI. (2021, May 28). *Schematiseringshandleiding macrostabiliteit*. https://www.helpdeskwater.nl/onderwerpen/waterveiligheid/primaire/beoordelen/@205756/schematiseringshandleiding-

macrostabiliteit/#:~:text=De%20schematiseringshandleiding%20Macrostabiliteit%20beschrijft%20de,waarmee%20de%20toets%20wordt%20uitgevoerd.
Yıldırım, S. (2021). Hyperparameter Tuning for Support Vector Machines — C and Gamma Parameters. https://towardsdatascience.com/hyperparameter-tuning-for-support-vector-machines-c-and-gamma-parameters-6a5097416167
Pentoś, K., Mbah, J. T., Pieczarka, K., Niedbała, G., & Wojciechowski, T. (2022). Evaluation of multiple linear regression and machine learning approaches to predict soil compaction and shear stress based on electrical parameters. *Applied Sciences (Switzerland)*, *12*. https://doi.org/10.3390/app12178791
PovDGG. (2022). Technisch kader toepassing gebiedseigen grond.
Team, O. (2022, April 15). *Optuna: An automatic hyperparameter optimization framework*. https://odsc.com/blog/optuna-an-automatic-hyperparameter-optimization-framework/#:~:text=Optuna%20is%20a%20software%20framework,can%20be%20used%20in%20Python.
Alexander, G. (2023). *What is validation set?* Retrieved August 6, 2024, from https://www.techtarget.com/whatis/definition/validation-set
Chala, A. T., & Ray, R. (2023). Assessing the performance of machine learning algorithms for soil classification using cone penetration test data. *Applied Sciences (Switzerland)*, *13*. https://doi.org/10.3390/app13095758
Geeksforgeeks. (2023, May). Introduction to dimensionality reduction. https://www.geeksforgeeks.org/dimensionality-reduction/
Gillis, A. S. (2023, September). validation set. https://www.techtarget.com/whatis/definition/validation-set#:~:text=Validation%20data%20sets%20use%20a,the%20behavior%20of%20the%20model.
IBM. (2023). *Neural networks*. https://www.ibm.com/topics/neural-networks
Kanade, V. (2023, April 3). *What is linear regression?- spiceworks*. https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-linear-regression/
Saxena, S. (2023, August). A Beginner8217;s Guide to Random Forest Hyperparameter Tuning. https://www.analyticsvidhya.com/blog/2020/03/beginners-guide-random-forest-hyperparameter-tuning/
Aimil Ltd. (2024, February 20). *What is soil testing and how is it done for bridges, buildings, and dams? - aimil corporate blog*. https://www.aimil.com/blog/what-is-soil-testing-and-how-is-it-done-for-bridges-buildings-and-dams/
DataRobot. (2024, February 15). *Data preparation for machine learning — datarobot artificial intelligence wiki*. https://www.datarobot.com/wiki/data-preparation/
DeepChecks. (2024, May 27). *What is mean absolute error? formula significance*. https://deepchecks.com/glossary/mean-absolute-error/
Dhadse, A. (2024). Removing outliers. understanding how and what behind the magic. https://medium.com/analytics-vidhya/removing-outliers-understanding-how-and-what-behind-the-magic-18a78ab480ff
, G. (2024, August 7). *R-squared in regression analysis in machine learning*. https://www.geeksforgeeks.org/ml-r-squared-in-regression-analysis/
Gupta, A. (2024, June 26). *Mean squared error: Overview, examples, concepts and more*. https://www.simplilearn.com/tutorials/statistics-tutorial/mean-squared-error
IBM. (2024). What is Random Forest? — IBM. https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems.
Jain, S. (2024, June 18). *Regularization in deep learning with python code*. https://www.analyticsvidhya.com/blog/2018/04/fundamentals-deep-learning-regularization-techniques/
Pandian, S. (2024, September). K-Fold Cross Validation Technique and its Essentials. https://www.analyticsvidhya.com/blog/2022/02/k-fold-cross-validation-technique-and-its-essentials/
Rendyk. (2024, January). Tuning the hyperparameters and layers of neural network deep learning. https://www.analyticsvidhya.com/blog/2021/05/tuning-the-hyperparameters-and-layers-of-neural-network-deep-learning/

A Appendix A: Keyword & collection of data

Table 37: Overview of Keywords

Keyword	Data	file type
FysischKlei	Clay inspection data	PDF, XLS, XLSM
FLCS	Compaction clay data	PDF
B_BIT, B_BUT, B_BUK	Triaxial test Data	PDF

Table 38: Removed Keywords

Keyword	Data	Keyword location
Clay inspection data	Vak, Gloei, ATT, MM, TAUW, HB, Depot, M1, B01+B04, Bouwstoffen, M1, concept	filename
Compaction clay data	Nuclaire, Nuclear, zand, sand	Inside the pdf

A.0.1 Cleaning data:

Cleaning data was applied to both clay inspection and clay compaction data, while only duplicate removal was applied to the triaxial test because of the quantity of the found samples. For both cases, missing values were filled in using the mean value of the column. This is because most of the data missed around 10 values, and only two columns missed around 100 values, with a small interval; therefore, using the average values would not affect the correlations significantly. Table 39 shows the quantity of the disappeared value per variable.

Table 39: missing values

Variable	quantity of missing values	filling method
Dry density	9	average value
Organic content	64	average value
Mass loss	100	average value

Furthermore, outliers were detected using the interquartile method, namely, the 25th and 75th, as most data distributions were skewed. On that note, the following number of outliers was found, and all replacements were applied to the upper limit of the column.

Table 40: missing values

Variable	quantity of missing values	upper limit
salt content	33	0.72
organic content	130	6.27
plastic limit	64	34

B Appendix B: Statistics and data quality

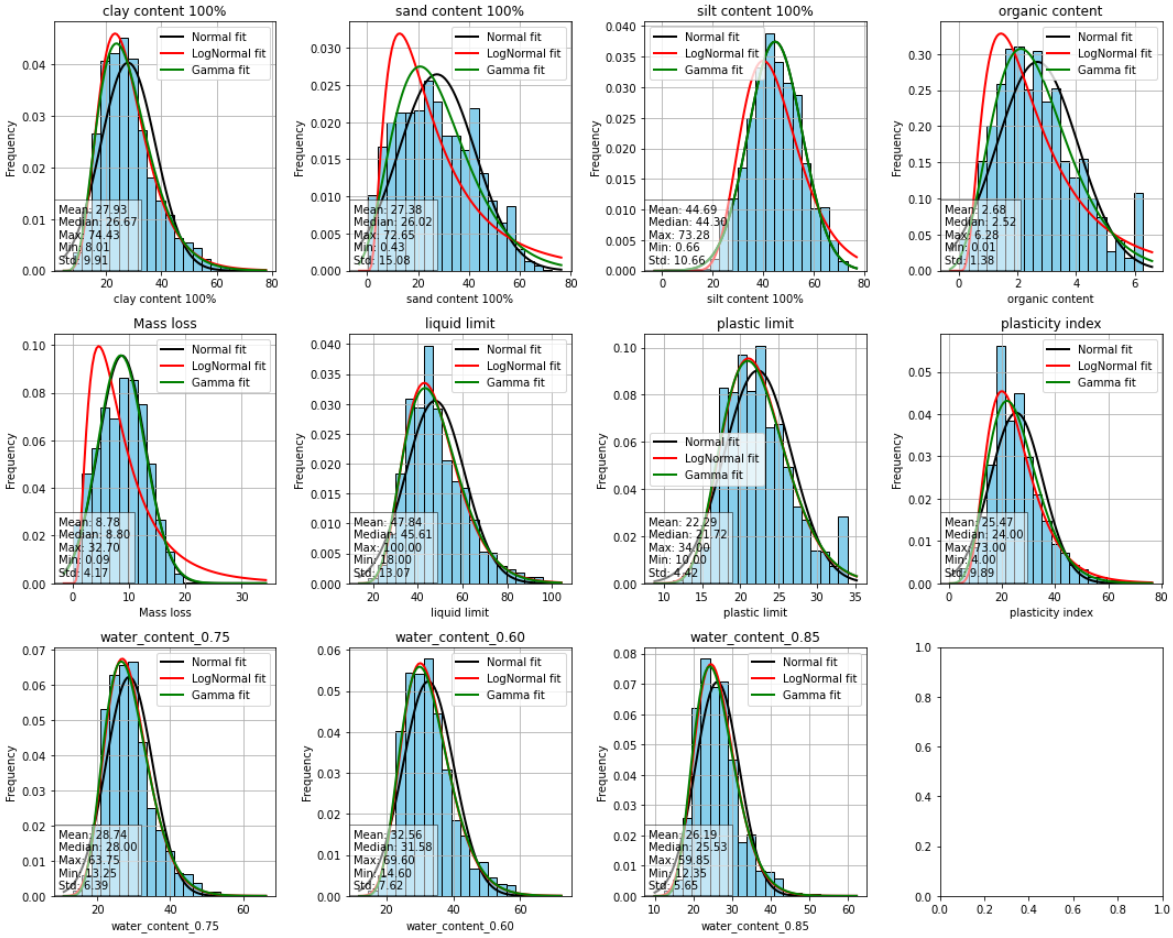


Figure 65: Distributions of clay inspection variables

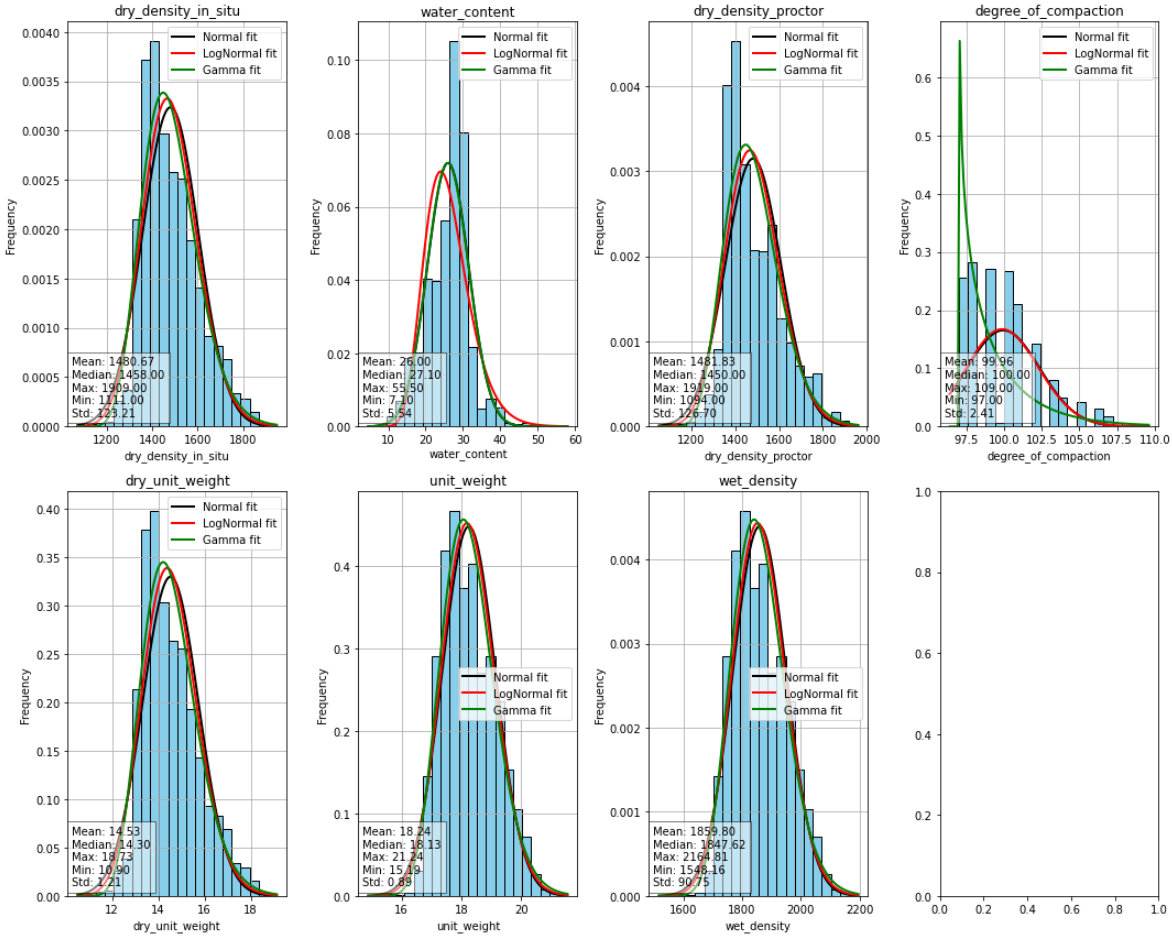


Figure 66: Histograms clay compaction

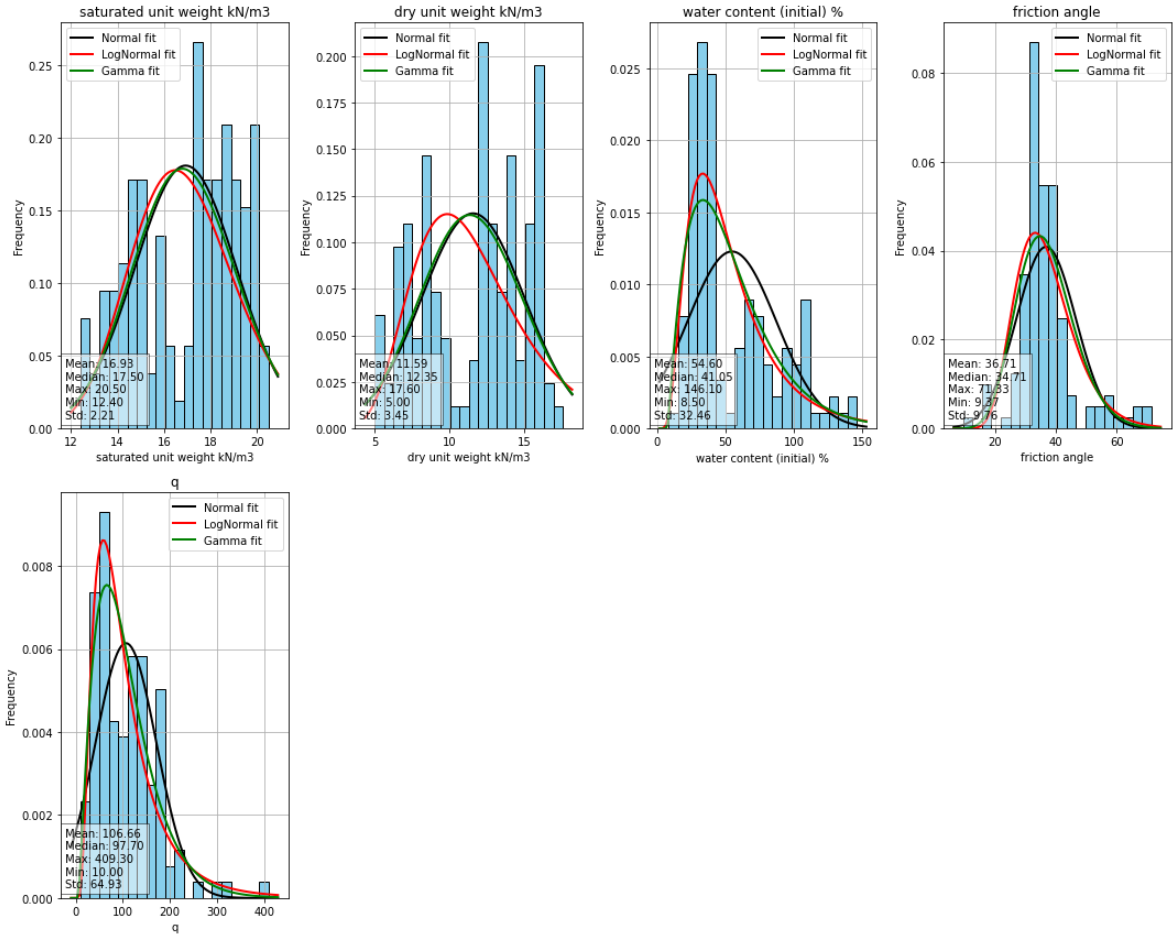


Figure 67: Histograms Triaxial tests

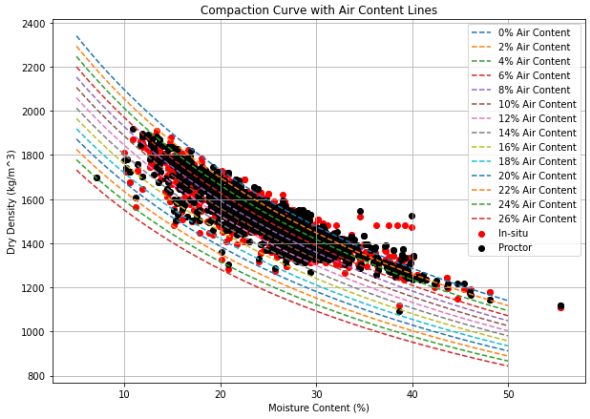
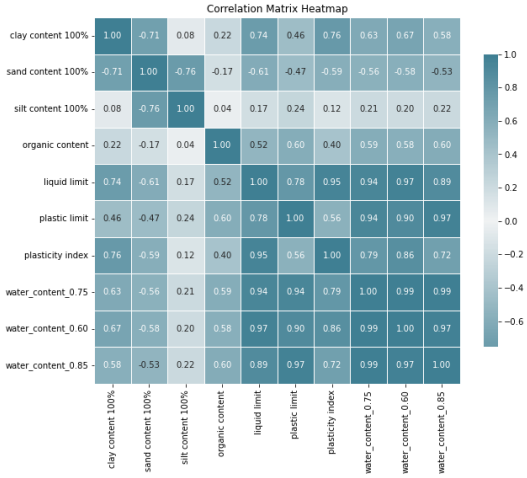
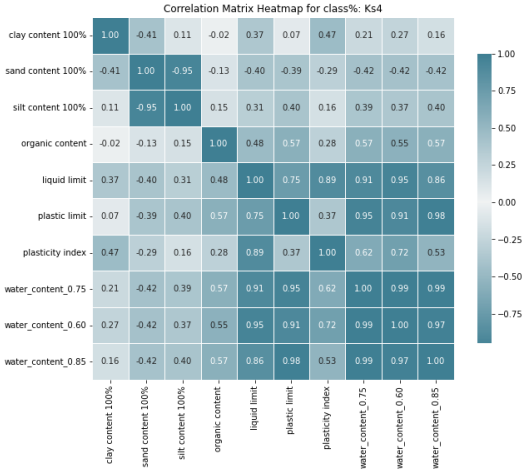


Figure 68: Compaction curve with air content lines.png

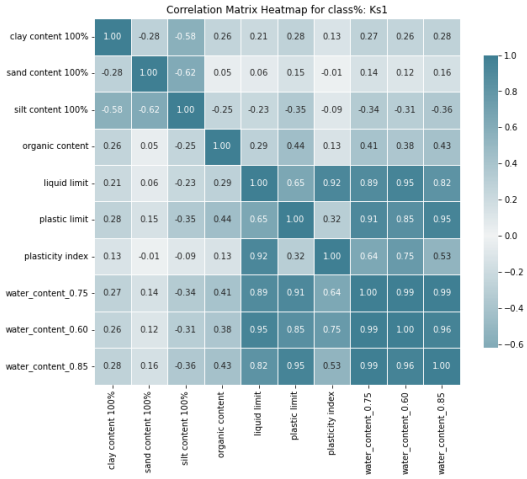
C Appendix C: Heatmaps



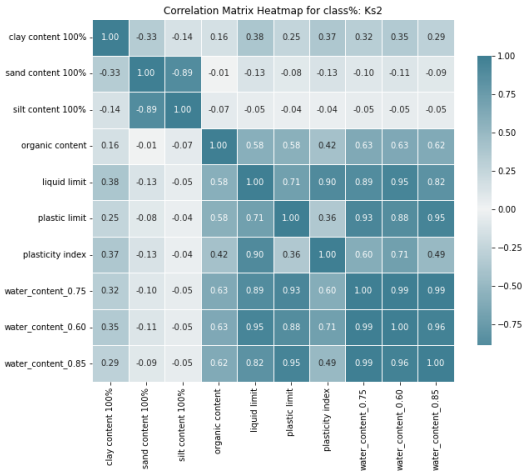
(a) Heatmap clay inspection data



(b) Ks4 Heatmap

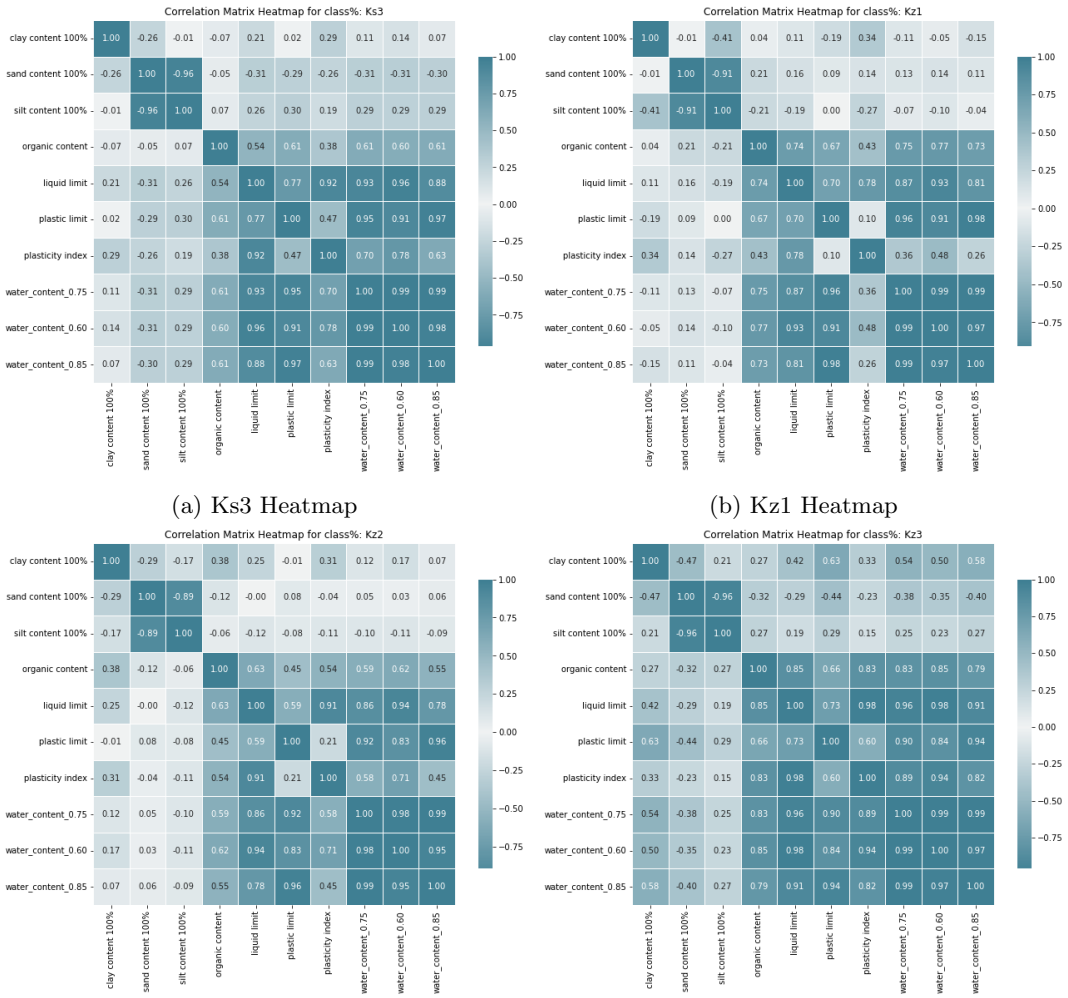


(c) Ks1 heatmap



(d) Ks2 Heatmap

Figure 69: Correlation Matrix Heatmaps for Different Clay Types: part 1



(a) Ks3 Heatmap

(b) Kz1 Heatmap

(c) Kz2 Heatmap

(d) Kz3 Heatmap

Figure 70: Correlation Matrix Heatmaps for Different Clay Types: Part 2

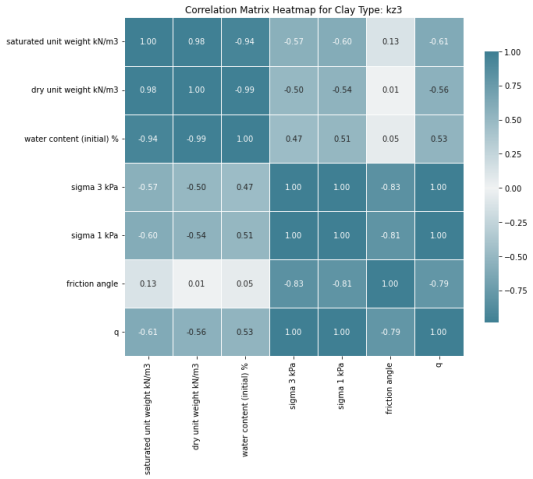
(a) Triaxial ks1 heatmap

(b) Triaxial ks2 heatmap

(c) Triaxial ks3 heatmap

(d) Triaxial kz1 heatmap

Figure 71: Correlation Matrix Heatmaps for Different Clay Types: Triaxial: part 1



(a) Triaxial kz3 heatmap

Figure 72: Correlation Matrix Heatmaps for Different Clay Types: Triaxial: Part 2

D Appendix D: Examples of the different reports

FUGRO INGENIEURSBUREAU BV
Materiaalkundig Laboratorium

ONDERZOEKSRAPPORT			
Project	Lateraalkanaal te Roermond		
Opdrachtgever	Van den Biggelaar B.V.	Opdrachtnummer	7209-0319-001
Contact persoon	de heer W.J. de Vos	Datum rapport	15-07-2010
Monstername	Uitgevoerd door Fugro Ingenieursbureau bv	Datum ontvangst	08-07-2010

ONDERZOEK MONSTERS		
Monster	Omschrijving	Diepte in meters t.o.v. maaiveld
1	HB 3.215	
2	HB 3.500	
3	HB 3.800	
4	HB 4.100	
5	HB 4.400	
EISEN	bestekseis	

RESULTATEN									
Parameter/Verrichting		Monster					Eisen	Eenheid	Methode van onderzoek
		1	2	3	4	5			
Watergehalte (A)		19	15	20	15	20	--	%(m/m)	proef 161.1 Std RAW/NEN5112
Gehalte > 63µm	Q	11.5	26.5	11.6	44.9	18.2	≤50	%(m/m)	proef 2 Std RAW
Gehalte < 2µm	Q	-	-	-	-	-	--	%(m/m)	proef 125 Std RAW
Gehalte organische stof	Q	3.7	3.7	3.9	3.4	3.4	≤5	%(m/m)	proef 158 Std RAW
Massa verlies bij HCl-beh.	Q	4.8	4.3	5.1	3.7	5.4	≤25	%(m/m)	proef 159 Std RAW
Geleidingsvermogen	Q						--	µS/cm	proef 122 Std RAW
Vloeigrens (W_l)	Q	38	31	43	24	31	--	%(m/m)	proef 15 Std RAW
Uitrolgrens (W_p)	Q	21	16	19	13	21	--	%(m/m)	proef 15 Std RAW
Plasticiteits-index (I_p)	Q	17	15	24	11	10	—	--	proef 15 Std RAW
A-lijn		13	8	17	3	8	--	--	berekend als 0,73*(W_l-20)
Zoutgehalte bodemvocht		0.55	0.58	<0.05	<0.05	<0.05	≤4	NaCl g/l	[1])
Wmax		25	20	25	16	24	--	%(m/m)	berekend als W_p + 0,25 I_p
Consistentie-index (I_c)		1.12	1.08	0.97	0.88	1.06	$I_c \geq 0,75$	--	berekend als (W_l-A)/(W_l-W_p)
Vloeibaarheidsindex (I_l)		-	-	0.03	0.12	-	--	--	berekend als 1-I_c
voldoet aan bestekseis		ja	ja	ja	ja	ja	--	--	berekend als 1-I_c

OPMERKINGEN

De met "Q" gemerkte verrichtingen zijn geaccrediteerd door RvA.
[1]) Uitgevoerd door Alcontrol Laboratories B.V. te Hoogvliet
bestekseis opgegeven door opdrachtgever

Figure 73: Clay inspection report

RAPPORTAGE LABORATORIUMONDERZOEK

Project	Wolfreren Sprok in situ 2021	Projectnummer	1719-0340-010
Opdrachtgever	Combinatie De Betuwse Waard V.O.F.	Datum rapport	2023-02-20
Contactpersoon	Hr Bouwens	Datum monstername	2022-12-14
Monstername	Uitgevoerd door Fugro NL Land B.V.	Datum herkeuring	

VERDICHTINGSONDERZOEK IN-SITU KLEI STEEKRINGMETHODE

#	Monster ID	Coördinaten		Diepte-MV	Locatie	Materiaal	Opmerkingen
		X	Y				
K1	F-FN19ZB-W1HG	183727.1	432189.5	1.60	DS12b-BI-TA_laag1	klei	Cat3
K2	F-JN19ZB-QUSX	183725.8	432189.8	1.20	DS12b-BI-TA_laag2	klei	Cat3
K3	F-ON19ZB-O6WT	183726.5	432188.5	0.80	DS12b-BI-TA_laag3	klei	Cat3
K4	F-BSOF2C-SP9E	183721.7	432187.9	0.40	DS12b-BI-TA_laag4	klei	Cat2

Standaard RAW 2020 Artikel 22.02.16 Klei verwerking en verdichting & Artikel 22.02.17 Klei watergehalte

RESULTATEN

Parameter	Proefnummer										Eenheid	Methode van onderzoek	Q
	K1	K2	K3	K4									
Droge dichtheid in-situ	1830	1737	1819	1543							kg/m³	Proef 6	Q
Watergehalte in-situ	15.0	16.8	13.7	24.6							% (m/m)	Proef 6	Q
Droge dichtheid bij aanw. watergeh.	1838	1792	1866	1578							kg/m³	Proef 9	Q
W-Max*	24	24	24	28							%(m/m)		
Verdichtingsgraad	100	97	97	98							% (m/m)	Proef 3	
Voldoet aan eis min ≥ 97%	ja	ja	ja	ja									
Eis Wopt ≤ Wn ≤ Wmax	ja	ja	ja	ja									

GEBRUIKTE APPARATUUR

Proctor machine	OCH-011

Figure 74: Compaction test report

Opdracht 02P008157
Project Grondonderzoek Waalbandijk Neder-Betuwe

Triaxiaalproef conform NEN-5117

Boring	DT118.+054_B_BUK	Soort proefstuk	ongeroerd uit steekbus
Monster	mo-08	Testmethode	CAU
Diepte	3,55 [m-mv]		
	9,10 [m tov NAP]		

Klassificatie: Klei, zwak zandig, zwak humeus [conform NEN-5104]

Initiële eigenschappen:	symbool		eenheid
Hoogte	h_i	134	mm
Diameter	D_i	67	mm
Nat volumegewicht	γ_n	19,5	kN/m^3
Droog volumegewicht	γ_{dr}	15,3	kN/m^3
Watergehalte	W_i	27,9	%

Verzadigingsfase:

Verzadigingsspanning	U_b	300	kPa
Verzadigingsfactor	B_i	1,00	-

Consolidatiefase:		Isotroop	Anisotroop	
Effectieve celdruk	σ'_c	35,7	36,1	kPa
Effectieve axiale druk	σ'_c	35,7	65,4	kPa
	K0	1,00	0,55	[-]
Monsterhoogte	h_c	133,5	132,4	mm
Monsteroppervlak	A_c	35,3	35,4	cm^2
Monstervolume	V_c	471,2	468,6	cm^3
Nat volumegewicht	γ_n	19,5	19,6	kN/m^3
Droog volumegewicht	γ_{dr}	15,3	15,4	kN/m^3
Watergehalte	W_c	27,8	27,4	%
t100	t100	24		min

Belastingsfase:			
Axiale reksnelheid	v	0,90	%/uur
Deviatorspanning	$q_{u\ (max)}$	154,0	kPa
Axiale rek bij maximale deviatorspanning	$\varepsilon_{qu(max)}$	26,1	%
Ongedraineerde schuifsterkte	f_{undr}	77,0	kPa
	s'	142,1	kPa
Effectieve spanning Axiaal	σ'1	219,1	kPa
Effectieve spanning Radiaal	σ'3	65,1	kPa
Bij 50% max. bezwijkdeviatorspanning:			
Ongedraineerde elasticiteitsmodulus	$E_{undr;50}$	1,5	MPa
Axiale rek	$\varepsilon_{b;50}$	4,16	%

Eindresultaat beproeving:			
Monstervolume	V_b	468,6	cm^3
Watergehalte	W_e	25,2	%

Percentage rek	*0*	*2*	*5*	*15*	*25*	*Rek bij bezwijken 26,1*	***Maximale rek 26,1***	
s'	50,6	63,5	83,4	127,9	140,6	142,1	***142,1***	kPa
t	14,6	35,3	49,3	71,8	76,5	77,0	***77,0***	kPa
σ'3	36,0	28,2	34,1	56,1	64,1	65,1	***65,1***	kPa
σ'1	65,3	98,7	132,7	199,7	217,2	219,1	***219,1***	kPa
Eundr	[-]	2,06	1,39	0,76	0,50	0,48	***0,48***	MPa

INPIJN-BLOKPOEL ingenieursbureau

Figure 75: Triaxial report

E Appendix E: difference between class systems

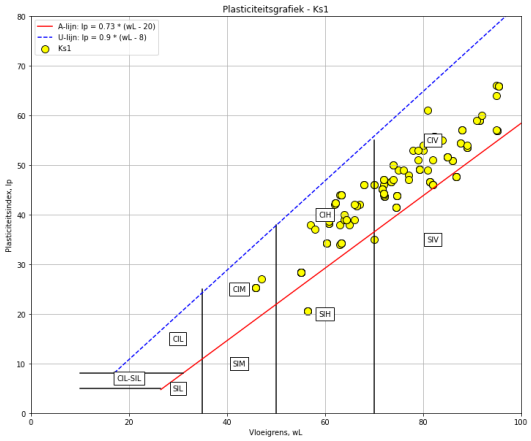


Figure 76: plasticity diagram classifications Ks1

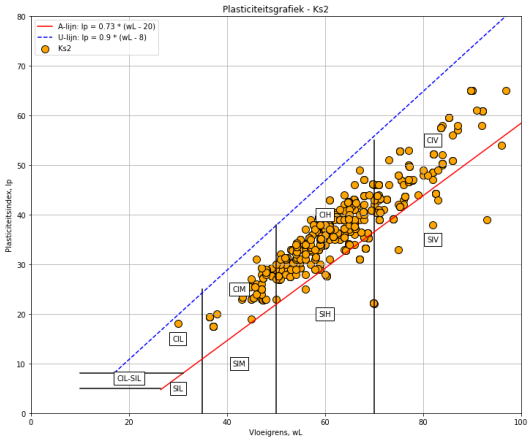


Figure 77: (plasticity diagram classifications Ks2

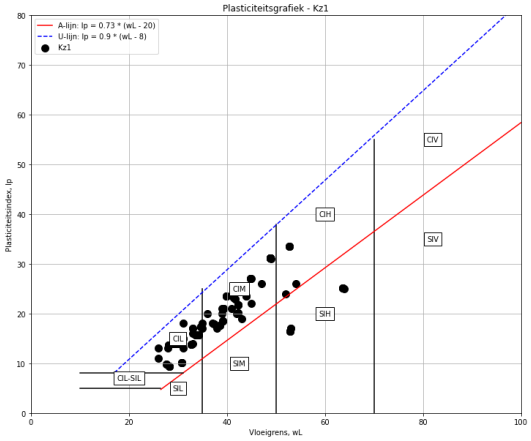


Figure 78: plasticity diagram classifications Kz1

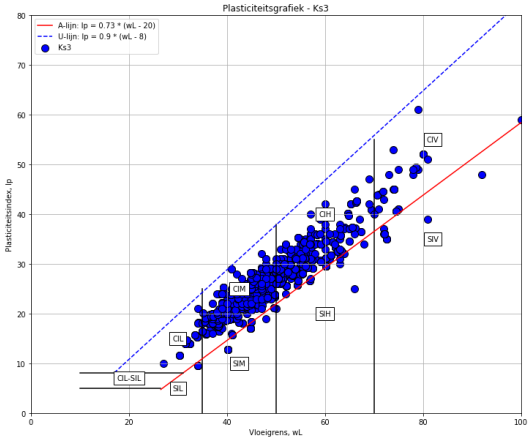


Figure 79: plasticity diagram classifications Ks3

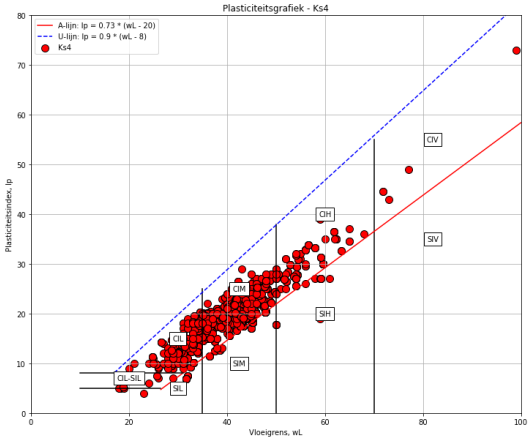


Figure 80: plasticity diagram classifications Ks4

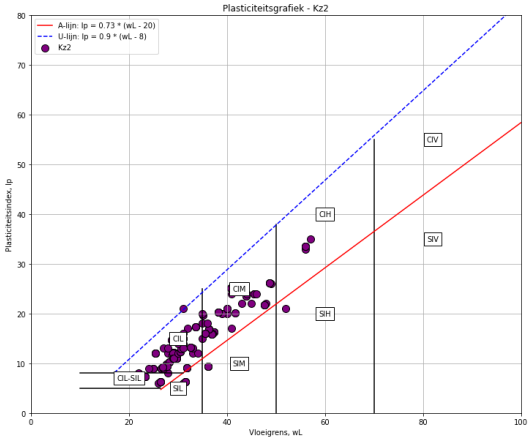


Figure 81: plasticity diagram classifications Ks4

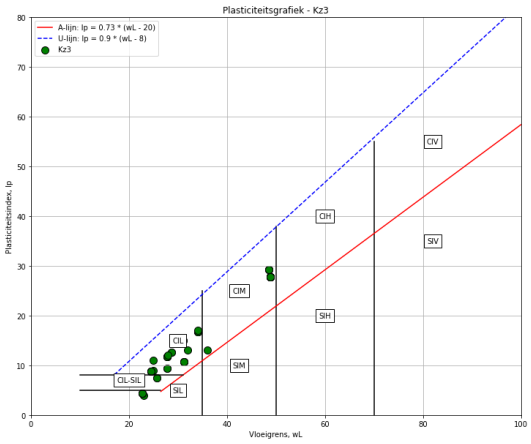


Figure 82: plasticity diagram classifications Kz3

F Appendix F: Linear correlations of the combined data sets

$$\text{sand\_content} = 58.97 - 1.12 \cdot \text{clay\_content} \quad (R^2 = 0.45)$$
$$\text{silt\_content} = 40.65 + 0.13 \cdot \text{clay\_content} \quad (R^2 = 0.01)$$
$$\text{liquid\_limit} = 13.21 + 1.17 \cdot \text{clay\_content} \quad \mathbf{(R^2 = 0.77)}$$
$$\text{plastic\_limit} = 14.80 + 0.26 \cdot \text{clay\_content} \quad (R^2 = 0.26)$$
$$\text{plasticity\_index} = -2.17 + 0.93 \cdot \text{clay\_content} \quad (R^2 = 0.88)$$
$$\text{water\_content\_0.85} = 14.90 + 0.39 \cdot \text{clay\_content} \quad (R^2 = 0.41)$$
$$\text{predicted\_dry\_density\_proctor\_0.85} = 1714.91 - 7.83 \cdot \text{clay\_content} \quad (R^2 = 0.34)$$
$$\text{predicted\_dry\_unit\_weight\_0.85} = 16.75 - 0.07 \cdot \text{clay\_content} \quad (R^2 = 0.33)$$
$$\text{silt\_content} = 58.88 - 0.53 \cdot \text{sand\_content} \quad (R^2 = 0.56)$$
$$\text{liquid\_limit} = 60.91 - 0.53 \cdot \text{sand\_content} \quad (R^2 = 0.44)$$
$$\text{plastic\_limit} = 26.26 - 0.15 \cdot \text{sand\_content} \quad (R^2 = 0.22)$$
$$\text{plasticity\_index} = 35.09 - 0.40 \cdot \text{sand\_content} \quad (R^2 = 0.45)$$
$$\text{water\_content\_0.85} = 31.33 - 0.20 \cdot \text{sand\_content} \quad (R^2 = 0.30)$$
$$\text{predicted\_dry\_density\_proctor\_0.85} = 1371.84 + 4.42 \cdot \text{sand\_content} \quad (R^2 = 0.29)$$
$$\text{predicted\_dry\_unit\_weight\_0.85} = 13.51 + 0.04 \cdot \text{sand\_content} \quad (R^2 = 0.30)$$
$$\text{liquid\_limit} = 36.57 + 0.22 \cdot \text{silt\_content} \quad (R^2 = 0.04)$$
$$\text{plastic\_limit} = 18.55 + 0.08 \cdot \text{silt\_content} \quad (R^2 = 0.04)$$
$$\text{plasticity\_index} = 20.31 + 0.09 \cdot \text{silt\_content} \quad (R^2 = 0.01)$$
$$\text{water\_content\_0.85} = 21.55 + 0.10 \cdot \text{silt\_content} \quad (R^2 = 0.04)$$
$$\text{predicted\_dry\_density\_proctor\_0.85} = 1614.56 - 2.77 \cdot \text{silt\_content} \quad (R^2 = 0.06)$$
$$\text{predicted\_dry\_unit\_weight\_0.85} = 15.86 - 0.03 \cdot \text{silt\_content} \quad (R^2 = 0.06)$$
$$\text{plastic\_limit} = 8.69 + 0.29 \cdot \text{liquid\_limit} \quad (R^2 = 0.56)$$
$$\text{plasticity\_index} = -6.29 + 0.66 \cdot \text{liquid\_limit} \quad (R^2 = 0.78)$$
$$\text{water\_content\_0.85} = 7.27 + 0.40 \cdot \text{liquid\_limit} \quad (R^2 = 0.78)$$
$$\text{predicted\_dry\_density\_proctor\_0.85} = 1870.21 - 8.15 \cdot \text{liquid\_limit} \quad (R^2 = 0.64)$$
$$\text{predicted\_dry\_unit\_weight\_0.85} = 18.21 - 0.08 \cdot \text{liquid\_limit} \quad (R^2 = 0.63)$$
$$\text{plasticity\_index} = -1.62 + 1.16 \cdot \text{plastic\_limit} \quad (R^2 = 0.37)$$
$$\text{water\_content\_0.85} = 1.30 + 1.10 \cdot \text{plastic\_limit} \quad (R^2 = 0.89)$$
$$\text{predicted\_dry\_density\_proctor\_0.85} = 1996.12 - 22.62 \cdot \text{plastic\_limit} \quad (R^2 = 0.75)$$
$$\text{predicted\_dry\_unit\_weight\_0.85} = 19.34 - 0.21 \cdot \text{plastic\_limit} \quad (R^2 = 0.72)$$
$$\text{water\_content\_0.85} = 15.21 + 0.44 \cdot \text{plasticity\_index} \quad (R^2 = 0.52)$$
$$\text{predicted\_dry\_density\_proctor\_0.85} = 1708.66 - 8.95 \cdot \text{plasticity\_index} \quad (R^2 = 0.43)$$
$$\text{predicted\_dry\_unit\_weight\_0.85} = 16.69 - 0.08 \cdot \text{plasticity\_index} \quad (R^2 = 0.42)$$
$$\text{predicted\_dry\_density\_proctor\_0.85} = 2015.98 - 20.23 \cdot \text{water\_content\_0.85} \quad (R^2 = 0.82)$$
$$\text{predicted\_dry\_unit\_weight\_0.85} = 19.54 - 0.19 \cdot \text{water\_content\_0.85} \quad (R^2 = 0.79)$$
$$\text{predicted\_dry\_unit\_weight\_0.85} = 0.75 + 0.01 \cdot \text{predicted\_dry\_density\_proctor\_0.85} \quad (R^2 = 0.96)$$

G Appendix G

G.0.1 Soil contents & class fo clay: clay inspection tests

After collecting the clay inspection data, it was noticed that the data has two different PDF files; the first one mentioned the $content < 63\mu m$ and $content < 2\mu m$, while the other mentions $content > 63\mu m$ and $content < 2\mu m$. In other words, these two different contents had the exact location in the table, collected under the same column. Therefore, the contents needed to be calculated differently between the two categories. In other words, these two categories of data were mixed as $content63$ and had to be separated. Therefore, a new variable called content-check was introduced (only for the separation process). This variable was calculated as:

$$content - check = content63 + content < 2 + OC + M_{loss} \tag{38}$$

This content-check variable varied between 20 and 160 [%]. On that note, the two groups were separated around the percentage of 80% as indicated by the files, which was checked by doing a visual manual check of the files. Therefore, this assumption was made to calculate the contents further.

For this purpose, the contents were calculated as follows:

Table 41: Equations for calculating clay content

Group with content < 63
$CC = \text{content}_{<2}$
$SiC = \text{content}_{<63} - \text{content}_{<2} - SaC - OC - M_{\text{Loss}}$
$SC = 100 - SiC - CC - OC - M_{\text{Loss}} - SaC$
Group with content > 63
$CC = \text{content}_{<2}$
$SiC = 100 - SC - CC - SaC - OC - M_{\text{Loss}}$
$SC = \text{content}_{>63}$

Table 42: Clay contents symbol descriptions

Symbol	**Description**
CC	Clay content [%]
SiC	Silt content [%]
SC	Sand content [%]
OC	Organic content [%]
SaC	Salt content [%]
M_{loss}	Mass loss [%]
$\text{content}_{<63}$	Mix between data groups [%]

On that note, it was possible to determine the class of the clay, which was done by first plotting the triangle from 5a in Python and then plotting the x and y points according to the following equations; these two equation in Table 10were taken from a Fugro excel sheet that was used to determine the classes of clay samples.

Table 43: Equations for the clay classes coordinates in the classification triangle from Figure 20

	point coordinates
x	$(100 - SC) - CC * \cos(\frac{1}{3} * \pi)$
y	$\sin(\frac{1}{3} * \pi)$

After that, a function was made that gives the point the polygon's name.

G.1 Data collection:

Data collection has been through 3 different phases:

1. Collection of files based on keywords
2. filtering files with relevant data
3. Extraction of data and data reviewing

The first phase involved identifying the keywords of the files provided by Fugro. These keywords varied between the different data types. For the clay inspection, files were found in PDF, XLS, and XLSM formats. On the other hand, only PDF files were found for the clay compaction data and triaxial data. A table of the keywords is added in Appendix A along with the Python used to perform this operation.

The second phase was more about filtering the data relevant to the research. In the case of the clay inspection test, the search was mainly done on the standard form of the clay inspection test. However, there are files removed that had specific keywords in the name of the files (table 38, appendix A). In the case of the clay compaction test, the files have been removed based on the existence of multiple keywords inside the PDF (provided in the previous table). Lastly, for the triaxial test, the found PDF files were merged, and irrelevant pages were removed using a PDF reader called PDFgear, done manually.

The third phase took the most time in this research. Extracting data out of PDFs was time-consuming, and the extracted data needed cleaning and reviewing to ensure that everything was correctly extracted. This also included fixing encountered problems, like randomly merged data columns, which were fixed manually using the data-to-text Excel function, as Python was unable/inconsistent at solving this issue. On the other hand, some data were not extracted; therefore, some were manually added from the files.

H Appendix H: Results of RF on the different Dataframes

Table 44: Summary of Random Forest Performance, (*) = values are equal to the average value of the three measures (0.60, 0.75, 0.80)

Scenario	Variable	COMRF	COMARF	COMNN	COMANN
Scenario 1: Atterberg limits [Two inputted features]	Plasticity index	0.99	0.97	0.99	0.96
	Clay Content	0.76	0.97	0.77	0.97
	Sand Content	0.69	0.74	0.70	0.74
	Silt Content	0.52	0.50	0.56	0.52
	Water Content (*)	0.99	0.99	0.99	0.99
	Dry Unit Weight (*)	0.97	0.90	0.99	0.99
	Dry Density Proctor (*)	0.97	0.90	0.99	0.99
	Triangular Clay Class	0.69	0.85	0.70	0.85
	Plasticity Graph Class	0.94	0.94	0.94	0.91
Scenario 2: Clay contents [Three inputted features]	Liquid Limit	0.71	0.88	0.72	0.88
	Plastic Limit	0.50	0.61	0.50	0.61
	Plasticity index	0.72	0.95	0.72	0.96
	Water Content (*)	0.60	0.80	0.62	0.77
	Dry Unit Weight (*)	0.60	0.70	0.60	0.72
	Dry Density Proctor (*)	0.60	0.65	0.60	0.75
	Triangular Clay Class	0.92	0.94	0.92	0.93
	Plasticity Graph Class	0.65	0.80	0.65	0.79
Scenario 3: Triangular class [One inputted feature]	Liquid Limit	0.50	0.69	0.50	0.68
	Plastic Limit	0.19	0.23	0.19	0.23
	Plasticity Index	0.53	0.79	0.53	0.78
	Clay Content	0.89	0.87	0.89	0.86
	Sand Content	0.56	0.54	0.56	0.54
	Silt Content	0.22	0.20	0.22	0.19
	Water Content (*)	0.35	0.43	0.35	0.45
	Dry Unit Weight (*)	0.37	0.38	0.30	0.35
	Dry Density Proctor (*)	0.35	0.35	0.32	0.36
	Plasticity Graph Class	0.46	0.54	0.46	0.53
Scenario 4: Plasticity class [One inputted feature]	Liquid Limit	0.82	0.83	0.80	0.82
	Plastic Limit	0.61	0.62	0.61	0.61
	Plasticity Index	0.77	0.71	0.77	0.70
	Clay Content	0.50	0.70	0.50	0.70
	Sand Content	0.40	0.48	0.40	0.47
	Silt Content	0.11	0.13	0.11	0.12
	Water Content (*)	0.78	0.72	0.78	0.72
	Dry Unit Weight (*)	0.75	0.62	0.73	0.63
	Dry Density Proctor (*)	0.74	0.61	0.72	0.62
	Triangular Clay Class	0.41	0.52	0.41	0.52

I Appendix I

Table 45: Average and Standard Deviation of Dry Unit Weights

Clay Class	Avg_0.60	Avg_0.75	Avg_0.85	StdDev_0.60	StdDev_0.75	StdDev_0.85
CIH	12.96	13.48	13.85	0.39	0.26	0.41
CIL	15.23	15.75	16.15	0.55	0.62	0.57
CIL-SIL	16.37	16.27	16.23	#DIV/0!	#DIV/0!	#DIV/0!
CIM	13.94	14.51	14.99	0.45	0.58	0.63
CIV	11.73	12.33	12.81	0.52	0.44	0.36
SIH	12.34	12.84	13.22	0.59	0.50	0.24
SIL	16.28	16.31	16.60	#DIV/0!	#DIV/0!	#DIV/0!
SIM	13.27	13.51	13.76	0.17	0.45	0.33
SIV	11.78	12.52	12.55	0.22	0.55	0.17

Table 46: Standard Deviation of Unit Weights by Erosion Categories

Erosion Category	StdDev 0.60	StdDev 0.75	StdDev 0.85
EC1	0.51	0.51	0.21
EC2	0.25	0.39	0.45
EC3	0.64	0.76	0.81