Proceedings of the 7th International Conference on Geotechnical and Geophysical Site Characterization (ISC)
June 18-21, 2024, Barcelona, Spain

Machine learning tools for the treatment of offshore site investigations

Bruno Stuyts[1,2*]

[1] *Vrije Universiteit Brussel, OWI-Lab, Pleinlaan 2 1050 Brussels, Belgium*
[2] *UGent, Civil Engineering Departement, Technologiepark 68 Zwijnaarde, Belgium*
[*] *Corresponding author: bruno.stuyts@vub.be*

ABSTRACT. During geotechnical and geophysical site characterisation for large infrastructure projects, significant data volumes are being collected which need to be processed and interpreted. Due to the limited budgets available for site characterisation and the various sources of uncertainty, the interpretation relies on a combination of data from various sources (e.g. in-situ and laboratory tests), the use of parameter correlations from the literature and expert judgement. In recent years, modern data science techniques have become increasingly accessible to practicing engineers and researchers and they offer the possibility to improve several aspects of the site characterisation and parameter selection process. Machine learning models can be trained on high-quality datasets and expert judgement can also be internalised in the model formulations. In this contribution, the role of data science and machine learning for geotechnical site characterisation is discussed based on several example applications using datasets from offshore wind farm projects. The role of data coverage and data quality is discussed as well as the role of geophysical data for interpolating geotechnical point measurements in a quantitative way. Supervised and unsupervised machine learning techniques are explained and illustrated on the provided datasets. Finally, a perspective is given on the role of the emerging Large Language Models (LLM) for geotechnical site characterisation applications.

Keywords: Data science; Machine learning; Geotechnical site characterisation

1 Introduction

In recent years, digitisation has permeated nearly every aspect of modern society and it is continuing to transform the way in which people interact with their environment. Although the digital transformation is most noticeable in computer science and related fields, the way in which engineers work is also being impacted. Building predictive models for physical processes has always been one of the core tasks of the engineering practice. Machine learning allows engineers to build more sophisticated predictive models and to leverage ever larger datasets. While datasets used for the development of foundation design methods and geotechnical parameter correlations are generally relatively small, the increasing availability of digital data allows engineers to increase the knowledge base on which such models are built. For example, the test program which formed the basis for the development of the widely used lateral pile design methods by the American Petroleum Institute (API) consisted of just four model piles (Reese et al., 1974) (Reese et al., 1975). Digital data from such tests is much easier to obtain these days due to the availability of digital instrumentation. In the determination of geotechnical parameter correlations from CPT testing, the amount of available data plays an important role. Jamiolkowski et al. (2003) developed a correlation for the relative density of sand from CPT testing. The researchers recognised the value of carefully collecting data and storing it in a database for further analysis. A database of 484 CPT tests in a calibration chamber was thus obtained which allowed meaningful data analysis. Database technology has evolved since those days and the increasing capabilities of personal computers supplemented by the available of versatile cloud computing platforms create the necessary prerequisites for widespread digitisation. Even though the required building blocks for a thorough digitisation are in place, the adoption of data science and machine learning (ML) in geotechnical engineering is not ubiquitous. Several researchers and practicioners are actively exploring these techniques (Phoon and Zhang, 2023) and ISSMGE TC309 provides a platform for sharing knowledge, but there is no concensus on best practices for using ML. Moreover, ML can also produce unreliable models when used without proper un-

derstanding of the underlying principles, leading to a reluctance for introducing these models in daily practice. Overall, the workflow for geotechnical parameter selection and foundation design shown in Figure 1 has not fundamentaly changed and as such, many engineers do not see the necessity to deviate from existing design guidelines, calculation tools and algorithms.

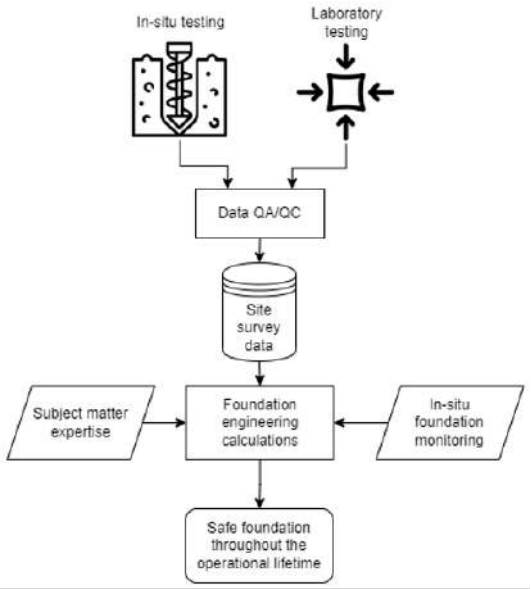

Figure 1. Flowchart for selection of geotechnical parameters and foundation design (Stuyts and Suryasentana, 2023)

In this contribution, data science and ML techniques are presented with a view to using them for geotechnical site characterisation. The paper aims to show that through a proper understanding of these novel techniques combined with subject matter expertise, ML models can be a very useful supplement to the toolkit of the geotechnical engineer. In Section 2, the relevant aspects of site investigation data for data science and ML are discussed. Machine learning models are highly dependent on the data which are fed to them and having a carefully curated high-quality dataset forms an essential starting point for any modelling effort. In Section 3, the datasets used in the examples are presented. Sections 4 and 5 provide the background for supervised and unsupervised learning models, two types of ML algorithms which encompass the majority of models in use today. Large Language models (LLMs) are described in Section 6. LLMs have received a lot of attention since the arrival of ChatGPT in 2022 (Deng and Lin, 2022) and a perspective is given on how they might in the geotechnical engineering profession. Finally, applications of supervised and unsupervised learning are provided in Sections 7 and 8.

2 Site characterisation data

Having appropriate site investigation data is crucial for any geotechnical assessment. For the use of these data in ML algorithms, there are a number of considerations which will determine whether ML modelling efforts can be attempted.

2.1 Data management

The data need to be available to the engineer and retrievable in a uniform digital format. Although several efforts have been undertaken to propose uniform data sharing standards, there is currently no uniformity in the way in which geotechnical data is transferred between different stakeholders. The practices depend very much on the national context.

2.1.1 File-based formats

The Association of Geotechnical and Geoenvironmental Specialists (2017) (AGS) has proposed a file-based standard for the transfer of geotechnical data. This standard is widely used in the offshore sector and also onshore in the United Kingdom. It facilitates collaboration on offshore wind farm projects and large onshore infrastructure projects. The standard models the relations between geotechnical tests and the location where they were performed. The standard is also extensible in case additional data fields or test types need to be captured. For geotechnical laboratory tests, summary outputs can be saved but detailed time series for e.g. oedometer tests or triaxial tests are currently not supported.

In the United States, the DIGGS standard was developed by the GeoInstitute of ASCE (Cadden and Keelor, 2017). This open-source data transfer standard uses the Geography Markup Language, a geospatially enabled extension of the eXtensible Markup Languag (XML), to create file-based representations of geotechnical data. The XML schema has a high level of detail and supports many different in-situ and laboratory tests. The schema is also being extended to include foundation tests such as pile load tests. The creators provide a data conversion tool to allow conversion of AGS files and other file formats to DIGGS. Although file-based transfer is suitable for smaller-scale projects, files are typically stored in project folders which are archived after a project's completion. This

may lead to data loss and reduces the potential to learn from historical data.

2.1.2 Geotechnical databases

To improve upon file-based data transfer, cloud-based geotechnical databases can be developed for the storage of factual and interpreted geotechnical data. Both commercial and open-source initiatives exist which all implement the basic relations between the data types involved in a geotechnical project. Stuyts et al. (2023) outline the development of a semi-structured database for geotechnical data and present the relations between the different entities (Figure 2). Construction projects encompass one or more geotechnical surveys and each surveys has one or more testing locations where in-situ tests are performed or samples are taken. On those samples, further laboratory tests can be performed. These entity relations were implemented by the authors in a PostgreSQL database with PostGIS extension for geospatial functionality. The schema of cloud-based databases cannot be altered by the database user (database fields cannot be added) so differences in the structure of the data have to be accomodated in an alternative manner. This can be achieved by allowing so-called *unstructured* database fields in which the user has freedom to store data with varying structure. JSON fields are used in the database proposed by Stuyts et al. (2023) to accomodate differences in CPT data formats or triaxial test outputs. To allow further use of the JSON fields, standardisation of the JSON format is required but this standardisation is not enforced by the database itself. The user is then responsible to adhere to agreed data formatting practices (e.g. always using the column name *qc* for storing cone tip resistance listings).

Using a geospatially enabled database allows data to be stored and retrieved in a structured manner. Data loss is prevented since data from historical projects remains available in the database for further use. When the database is stored in the cloud, it is accessible 24/7 and world-wide allowing all stakeholders to effectively collaborate on the data. To ensure that the database contains high-quality data, a data QA/QC process is required. This can either happen before uploading the data to the database or by creating versions of the data with different levels of approvals. Only the fully approved data should then be used in a design process.

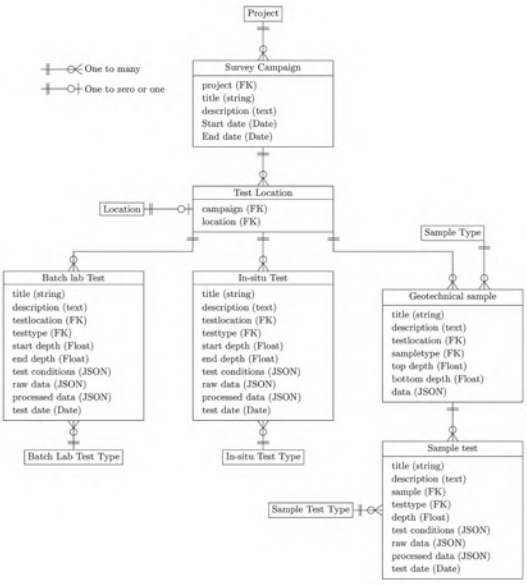

Figure 2. Entity relations between geotechnical data types (Stuyts et al., 2023)

2.1.3 Open data

Several initiatives exist for making geotechnical data available in the public domain. These initiatives are often led by governments who disclose the data obtained with tax-payer funds. For students, researchers and practicing engineers, this data can be a very useful supplement to commercially obtained site investigation data. Various schemes for sharing data in the public domain exist, ranging from file-based data sharing on web pages to providing access to cloud-based geotechnical database through XML or GIS interfaces.

The tendering process for offshore wind farm projects in Europe has played a significant role in making more offshore site investigation data available to the public. Several governments provide geotechnical survey data as part of the public tender for offshore wind farm sites. This data is available for the following countries:

- The Netherlands (`https://offshorewind.rvo.nl/`): Reports and digital file-based data for offshore wind farm projects in the Dutch Exclusive Economic Zone (EEZ).
- Germany (`https://pinta.bsh.de/?lang=en`): Reports and digital file-based data for BSH tenders for offshore wind farm projects in the German North Sea and Baltic Sea.
- Belgium (`https://offshore.digital-database.`

economie.fgov.be): File-based geotechnical and geophysical data and GIS application for data viewing for the public tender for the Belgian Princess Elizabeth Zone.

- United Kingdom (https://www.marinedataexchange.co.uk/): Geotechnical and geophysical data for offshore wind farm projects which are fully commissioned. The offshore wind farm developers are obliged to submit this data after wind farm construction is complete.

The Flemish goverment (https://dov.vlaanderen/be) makes CPT and borehole data for the entire Flemish region available through a web-based viewer or a XML Application Programming Interface (API). Such APIs allow geotechnical data to be retrieved based on complex geospatial data queries and allow automation of the geotechnical parameter selection and design workflows. A similar platform is available for the Netherlands (https://www.dinoloket.nl/ondergrondgegevens). These data platforms provide the large datasets which form a pre-requisite for data science and machine learning tasks.

2.2 Data quality assessments

Before starting a machine learning workflow, the quality of the data always needs to be checked. Several aspects of the data acquisition process can affect the data quality. Geotechnical testing creates disturbances in the soil which will have an impact on the test results. Sample disturbance during field sampling can be assessed by measuring the change in void ratio during the consolidation phase of triaxial testing (Lunne et al., 1998). Where possible, such checks should always be performed and the the triaxial testing results should be supplemented with meta-data describing any possible effects of the sampling technique. For reconstituted samples, the reconstitution technique can lead to differences in the behaviour during testing (Fearon and Coop, 2000). Although the majority of geotechnical datasets will only contain the geotechnical parameter which is the result of the test (e.g. peak drained friction angle), being aware of the meta-data will allow data to be differentiated based on the basis of sample quality. Sample quality can be encoded as a categorical feature. For example, the sample quality for cohesive samples is subdivided into four categories by Lunne et al. (1998). However, no concensus currently exists on a uniform scoring system which would apply to all soil mechanical tests.

Even when a dataset is gathered with high-quality sampling and testing techniques, statistical variations will still exist within the geological formation. Describing the statistical properties of the dataset with both numerical metrics (e.g. mean, median, standard deviation) and graphical representations (boxplots, histograms) is recommended to allow the engineer to gain in-depth knowledge of the dataset they are working with. Geological formations with large statistical variations can be discerned from those with more uniform properties. A good understanding of the site geology is also relevant in this respect. Low energy depositional environments with more fine grained material will typically show a greater uniformity of test results than high energy environments where more coarse grained material is present. This is illustrated with a CPT trace from the Princess Elizabeth Zone (PEZ) offshore Belgium. This dataset is described further in Section 3. Figure 3 shows that from surface to 11m depth, a layer of dense to very dense sand with heterogenous cone resistance is observed. Below 11m, stiff Tertiary clay of the Formation of Kortrijk with more uniform cone resistance is identified.

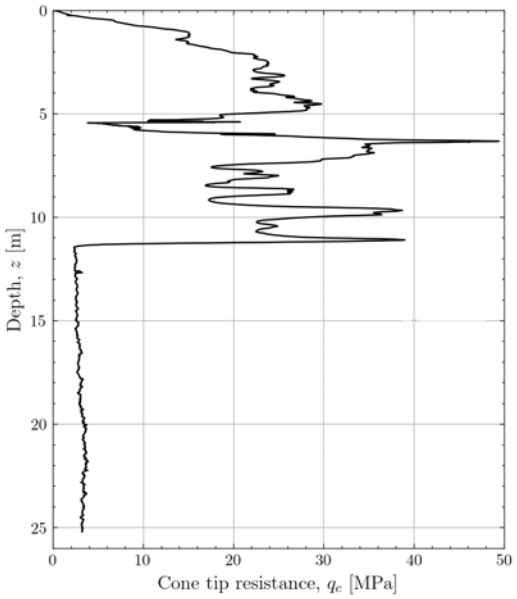

Figure 3. Cone tip resistance trace for a location from the Belgian Princess Elizabeth Zone.

During the exploration of the dataset, it is also important to identify correlations between geotechnical parameters. Pearson's correlation coefficient is often used to calculate a metric for the correlation between features of a dataset. It should however be noted that Pearson's coeffcient determines the amount of linear

correlation between features. If the data shows non-linearity, the correlation coefficients are misleading. Rather than calculating correlation coefficients numerically, it is recommended to create a scattermatrix plotting the feature values against each other. This reveals non-linear correlations in the dataset. An example of such a scattermatrix for the ISFOG pile driving dataset (Stuyts, 2020) is shown in Figure 4. The diagonal subplots show a histogram of the individual features and the off-diagonal subplots reveal correlations between the features.

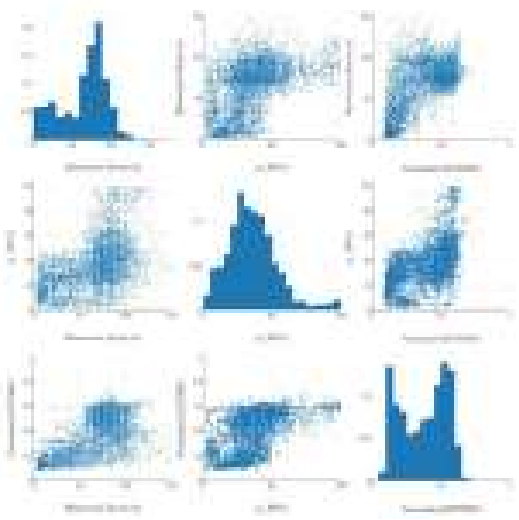

Figure 4. Scattermatrix for the ISFOG pile driving dataset (Stuyts, 2020).

The geospatial data coverage also impacts the quality of a dataset. While sparse datasets may capture large-scale variations, the statistical properties of such data may be less relevant when only considering the soil volume around a small subset of the boreholes. In addition to conventional statistical analysis, variograms capture the geospatial uncertainty on the data (Chiles and Delfiner, 2012). Equation 1 shows the mathematical formula for an experimental variogram. For a number of point pairs with separation distance $\overrightarrow{h}$, the sum of the differences between the values of the function F at the first points and second points are taken and divided by the number of pairs. Figure 5 shows the variogram for axial pile resistance at 30m depth for 2.5m diameter tubular piles at a North Sea wind farm. The figure shows that at separation distances less than 1000m, reduced variability is observed. For larger separation distances, the variation is equal to the variance of the entire site. The determination of a high-quality variogram is only possible when a dense site investigation coverage is available.

$$\gamma^*\left(\overrightarrow{h}\right) = \frac{1}{N\left(\overrightarrow{h}\right)} \sum_{i=1}^{N\left(\overrightarrow{h}\right)} \left[F\left(\overrightarrow{x_i} + \overrightarrow{h}\right) - F\left(\overrightarrow{x_i}\right)\right] \tag{1}$$

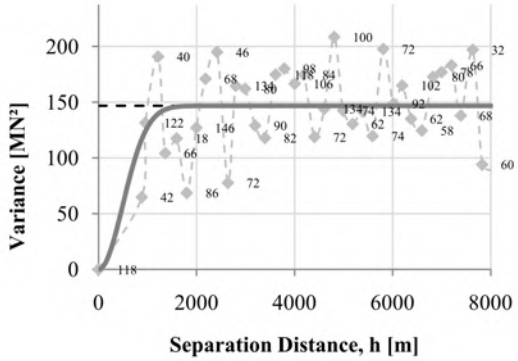

Figure 5. Variogram of axial pile resistance at 30m depth for 2.5m diameter tubular piles at a North Sea wind farm (Stuyts et al., 2010). The number of pairs are shown for each spacing.

Machine learning model building should only start after the engineer has gained a throrough understanding of the dataset in terms of how it was acquired, which geospatial trends and correlations may exist and what the inherent statistical variations are.

2.3 Combining data from various sources

When developing the datasets for the determination of geotechnical parameter correlation, data from various sources need to be combined. For example, when developing the database for determining a correlation for effective friction angle data in silts and clays from CPT data, Ouyang and Mayne (2018) had to determine the CPT measurements for depths corresponding to the depths of the triaxial testing samples. Another example is the determination of shear wave velocity from CPT data using the seismic CPT. In the offshore environment, the majority of tests is carried out using a dual-geophone setup (Figure 6). The value of V_s derived from the signals at the two geophones is typically assigned to a depth coordinate half-way between the two geophones. When comparing this shear wave velocity with CPT data, it should be taken into account that V_s represents the average wave propagation velocity in the region between the two geophones. It is therefore recommended to average the CPT measurements in this depth range. Such scale effects need

to be considered when combining data from different sources.

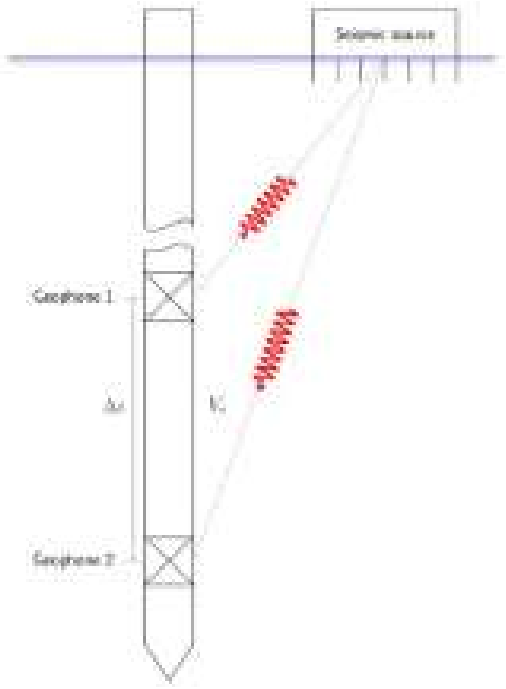

Figure 6. Schematic representation of a seismic cone with a dual-geophone setup.

A good characterisation of the elastic properties of the seabed through V_s measurements and CPT-based correlations for V_s allows the use of seismic inversion techniques. The geophysical data is then used in a quantitative manner to interpolate the available V_s measurements. Karkov et al. (2022) describe the use of amplitude vs offset seismic inversion for obtaining syntetic CPTs at locations where no geotechnical tests have been performed.

Another example where data from various sources has to be combined is when the same geotechnical parameter is measured using several methods with varying measurement uncertainty. For example, in offshore site investigations, undrained shear strength S_u may be derived from offshore laboratory tests with hand-held tools (torvane, pocket penetrometer and miniature vane tests) or from more accurate onshore laboratory tests such as direct simple shear (DSS) or consolidated undrained triaxial (CU) tests. CPT tests results can also be used to determine S_u using the proportionality factor N_{kt} between S_u and the net cone resistance q_{net} (Lunne et al., 2002). This is illustrated in Figure 7. The dataset on undrained shear strength which is shown in the figure contains data with varying levels of confidence. This needs to be taken into account during machine learning model building. Techniques such as Multi-fidelity data fusion (MFDF) (Stuyts and Suryasentana, 2023) can be used to work with such heterogeneous datasets and still make meaningful predictions.

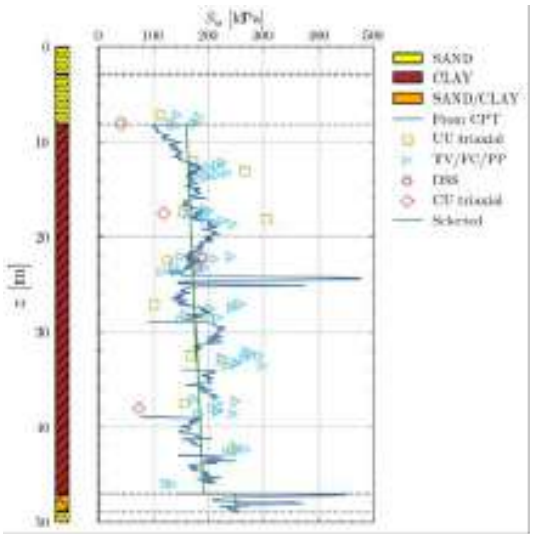

Figure 7. Undrained shear strength measurements from various data sources.

2.4 Data dimensionality

Geotechnical datasets typically contain a large amount of features to describe soil type, depth range, raw and normalised CPT parameters, ... When a dataset is partioned to look at selected combinations of soil type, depth range, cone resistance range, ... the amount of data in the dataset quickly shrinks and training a machine learning model becomes difficult.

Moreover, when a machine learning model is trained on a dataset with a large number of features, model coefficients have to be determined to capture the influence of each of those features. This is called the *curse of dimensionality* and may lead to ML models that perform poorly. Geotechnical dataset are often limited in size with hundreds or thousands of samples. Datasets used for training advanced machine learning models have millions or even billions of samples. To make machine learning models perform well, the number of samples needs to be much larger than the number of features. When evaluating a dataset, the features which have the most meaningful impact on the model outcome should be isolated and less meaningful features may be discarded during the model training process.

3 Datasets

Several machine learning techniques are illustrated in this paper based on a three example dataset from offshore wind farm projects. The datasets are summarised in the following sections and are provided on GitHub: `https://github.com/snakesonabrain/isc7_datasets`.

3.1 Downhole PCPT data on a sandy site

CPT testing can be performed in a continuous manner from the seabed, but if the cone tip encounters a hard stratum or the friction on the cone rods becomes too large (typically at deeper depths) refusal may occur and the test needs to be terminated. To mitigate this issue, offshore CPTs are often performed from the bottom of the drillstring. In this *downhole* mode, the CPT trace is composed of a number of consecutive *strokes*. Figure 8 shows an example *downhole* CPT with strokes of 3m length. When the CPT stroke is started from the bottom of the borehole, the cone resistance will have to build up until the soil fails plastically and the cone can advance. This initial phase of the penetration is characterised by a steep increase of cone resistance. The initial parts of the stroke are marked in red for the example CPT in Figure 8. The initial part of the stroke is not representative for the actual penetration resistance and needs to be removed from the CPT trace when using the trace in e.g. CPT-based pile design methods. In this paper, this dataset is used to illustrated classification models. The dataset consists of 80 *downhole* CPTs in dense to very dense sand from the German sector of the North sea. The data is processed to include both the raw CPT data and the normalised parameters Q_t, F_r and B_q. The soil behaviour type index I_c according to Robertson and Cabal (2015) is also available. Out of the 80 locations, 15 are manually labeled to differentiate the initial part of the stroke from the remainder of the data in the stroke.

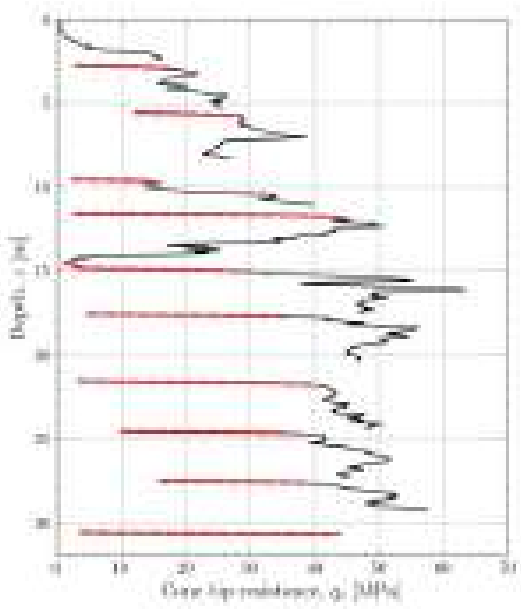

Figure 8. Example downhole CPT from the downhole PCPT dataset.

3.2 PCPT data from a site with stiff Tertiary clay

The second offshore wind farm zone in Belgium is called the Princess Elizabeth Zone (PEZ) and will be developed between 2024 and 2030. The Belgian federal government procured the geotechnical and geophysical surveys at the site and made the data available in the public domain.

An extensive CPT campaign was conducted at PEZ. Continuous CPT testing from the seabed (so-called *seafloor CPTs*) was performed until refusal. The deeper portions of the site were characterised by discontinuous *downhole* CPT testing. An example of a seafloor CPT and a downhole CPT at the same location is shown in Figure 9. The downhole CPT starts from the depth where the seafloor CPT terminates.

The PEZ site is characterised by a massive layer of stiff Tertiary clay from the Formation of Kortrijk which occurs across the entire site. Overlying the clay layer, there are varying amounts of sand cover. Within the Kortrijk Formation, certain marker horizons are identifiable from the geophysical data. The PEZ dataset is used to illustrate the detection of these marker horizons with anomaly detection algorithms.

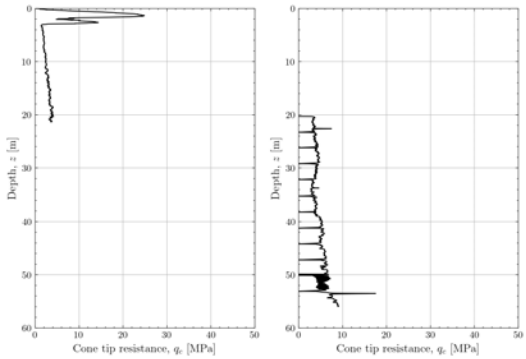

Figure 9. Example seafloor and downhole CPT for the PEZ.

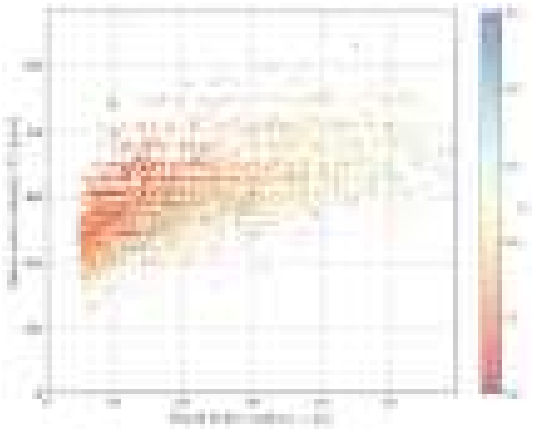

Figure 10. Overview of the North Sea V_s dataset.

3.3 S-PCPT shear wave velocity database

Determining shear wave velocity profiles from CPT data has been the subject of several studies (Rix and Stokoe, 1991), (Mayne and Rix, 1993), (Hegazy and Mayne, 2006), (Andrus et al., 2007). At the offshore wind farm sites in the Netherlands, extensive S-PCPT testing was performed which allowed the compilation of a dataset of 2905 S-PCPT measurements with corresponding CPT parameters. All S-PCPT tests in the dataset were executed with a seafloor CPT rig. The S-PCPT cone was equipped with a dual geophone setup. In accordance with the recommendations of the ISO 19901-8:2014 standard, all points shallower than 5m below seafloor were removed as these points may lead to inaccurate V_s estimates. Raw and normalised CPT parameters were compiled and the soil behaviour type index I_c was also calculated. All CPT parameters were averaged in the depth range between the two geophones.

Figure 10 shows an overview of the V_s dataset in terms of the depth, soil behaviour type index and shear wave velocity of the data. The data shows an increase of V_s with depth and smaller I_c-values (coarser grained soils) appear to be associated with higher V_s. The data shows significant scatter and a number of outliers. This dataset is used to illustrate regression models and dimensionality reduction techniques.

4 Supervised machine learning techniques

Machine learning algorithms which learn from known observations are known as supervised learning algorithms. These have several applications in geotechnical engineering such as the determination of soil parameter correlations, data-driven prediction of foundation resistance or stiffness and learning of foundation installation behaviour from recorded installation data. In this section, supervised learning techniques are described in terms of the model formulation, the model quality metrics and the application to the prediction of continuous and categorical target variables. Deep learning is also presented here in terms of supervised learning, although it should be noted that neural networks can also be used in unsupervised learning applications.

4.1 Formulation of the supervised learning problem

In a supervised learning algorithm, the machine learning algorithm takes a labelled dataset and learns patterns from this data during the training phase. The label is a value, also known as the *target*, which the user seeks to predict from the input. This *target* can either be discrete or continuous. The input variables on which the predictions are based are also known as *features*.

Figure 11 depicts the process of training a machine learning model using a dataset containing m labelled samples. This dataset comprises n features and a target variable y, which is known for the labelled samples. The primary goal of machine learning is to discover the

relationship between these features and the target, ultimately enabling accurate predictions for unseen features.

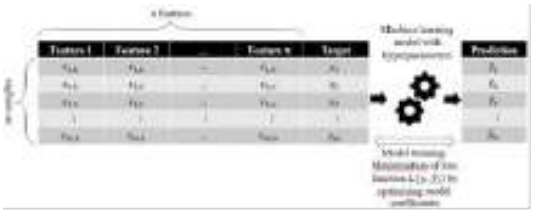

Figure 11. Schematic representation of a supervised learning problem.

The machine learning model is a mathematical construct that seeks to approximate the true relationship f between the target variable and the input features x_1 to x_n with an approximative relation $\hat{f}$. The prediction $\hat{y}$ generated by the machine learning model might deviate from the true value y of the target. During the training process, the model coefficients are adjusted to minimize the error ϵ, ensuring better alignment between predictions and observed outcomes. This is represented in Equation 2.

$$\hat{y} = \hat{f}\left(x_1, x_2, ..., x_n\right) = y + \epsilon \tag{2}$$

In machine learning, a *loss function* is minimized to determine the optimal set of model coefficients. This optimal fit relies on both the variability (scatter) in the training data and the model's capacity to capture the underlying patterns within the data.

Additionally, the model incorporates a set of *hyperparameters*; parameters that govern the model's behavior. These hyperparameters are fixed before the training process begins. However, they can be tuned by the user to enhance the overall performance of the machine learning model.

The training phase consists of minimising the loss function by optimising the model coefficients. This results in a minimal difference between the model predictions and the known target values of the labeled training dataset. Once trained, the model can be used to make predictions on unseen data.

4.2 Continuous target: Regression

When the target is a continuous variable, supervised learning is called *regression*. Basic regression algorithms such as linear regression are used in conventional geotechnical engineering. Modern data science libraries such as *scikit-learn* (Pedregosa et al., 2011) implement a wide variety of regression algorithms ranging from the linear regression model to more complex tree-based models such as random forests. Although the internal workings and numerical implementation of the models may be complex, the documentation of *scikit-learn* explains the underlying principles. Any user of such models should be aware of these principles as they provide an understanding of the strengths and weaknesses of the models. Discussing the wide variety of machine learning models types is beyond the scope of this paper. In the examples, two regression algorithms are applied to the V_s dataset; a basic linear regression model and the XGBoost algorithm (Chen and Guestrin, 2016).

The majority of machine learning models predict a scalar value for the target. The uncertainty on the estimate can then be derived by comparing the predicted and observed values of the target on the training dataset. Certain methods are however capable of calculating a confidence interval on the estimate. Gaussian Process Regression is a machine learning algorithm which is very similar to *kriging* techniques from geostatistics. Given samples with a certain multivariate distribution, the algorithm will learn patterns in the data but will also take into account the generalised distance between samples. The result is a prediction technique which does not only provide an estimate of the target but also the uncertainty on this estimate. These techniques can be very powerful on noisy datasets which contain non-linearities. Figure 12 shows a Gaussian process prediction of a 95% confidence interval on the shear wave velocity profile derived from CPT data. The algorithm has been trained on the V_s dataset discussed in Section 3. The estimate of the expected value of V_s from the CPT is shown as the green line and captures the typical trends of V_s reported in the literature (Cha et al., 2014). The predicted confidence interval shows a large uncertainty which encompasses the input data. The uncerainty increases at shallow depth, where no data is available. Gaussian Process Regression is very useful in highlighting the areas in which the estimate from a machine learning model is unreliable.

Figure 12. Prediction of the expected value of V_s and the associated confidence interval using Gaussian Process Regression.

4.2.1 Model quality metrics

When the model has been trained, the model accuracy can be calculated by evaluating the differences between the predictions $\hat{y}$ and the known target values$y$. Several accuracy metrics exist:

- Mean Absolute Error (MAE): $\text{MAE} = \frac{1}{n}\sum_1^n \|y_i - \hat{y}_i\|$
- Mean Square Error (MSE): $\text{MSE} = \frac{1}{n}\sum_1^n (y_i - \hat{y}_i)^2$
- Coefficient of determination (R^2): $R^2 = 1 - \frac{\sum_1^n (y_i - \hat{y}_i)^2}{\sum_1^n (y_i - \bar{y})^2}$

The coefficient of determination R^2 is an interesting metric for machine learning models as it is a measure for the proportion of the variance in the training data which is captured by the model. The score should be as close to 1 as possible, in which case the variance of the model predictions is exactly equal to the variance of the training data. R^2 can become negative if a model does not capture the underlying trends in the data.

By using appropriate loss functions, the accuracy metrics are optimised on the training dataset. However, this is no guarantee that the trained model will perform well on unseen data. To perform well on unseen data, the model must capture the underlying trends in the dataset and should not *overfit* the data. It is said that the model should *generalise* well. To evaluate this, a portion of the labelled dataset is withheld during the training phase. Typically 20% of the data is retailed as the *test* data. This *test* data is then used to evaluate the performance of the model on unseen data. This is called *train-test splitting*. If the model generalises well, it should show similar performance on the data used for training and on the test data which was not part of the training phase. The train-test split can be performed k times in which case it is called k-fold cross-validation. The accuracy scores of train and test set for each of the k folds are then analysed (Figure 13).

Figure 13. Illustration of k-fold cross-validation (Raschka, 2015).

4.3 Categorical target: Classification

When a supervised learning algorithm is tasked with predicting a discrete value, the model is called a *classification* model. The target is a class label which needs to be predicted from the feature values. Several geotechnical problems are categorical in nature, such as the prediction of a soil type from CPT data (Stuyts, 2020). If the number of classes is equal to two, the classification is called *binary classification*. Problems like the prediction of pile refusal during driving or the identification of the initial part of the stroke of a downhole CPT are binary classification problems.

Several machine learning algorithms exist such as, linear classifiers (logistic regression) and decision trees. These basic classification algorithms are illustrated on the downhole CPT dataset.

4.3.1 Prediction of the class label

All classification models will predict the class of the target variable but certain algorithms such as logistic regression and decision trees will also provide a class probability for each prediction. For example, in binary classification, the model will not predict whether the expected class label is 0 or 1, but it will instead predict the probability for each of these classes (e.g. the model will say that a sample has 80% chance of belonging to class 0 and 20% chance of belonging to

class 1). This allows the user to determine for which of range the feature values, the model is more reliable.

4.3.2 Model quality metrics

The model quality metrics are different from those used in regression models. The accuracy score is defined in Equation 3.

$$\text{Accuracy score} = \frac{\text{number of correct predictions}}{\text{number of samples}} \quad (3)$$

The accuracy of the model can also be displayed graphically in a confusion matrix. The matrix shows the true labels as matrix rows and the predicted labels as columns. In case of perfect predictions, all off-diagonal terms should be zero. This is illustrated in Figure 14 for a linear classifier prediction the soil type from CPT data (Stuyts, 2020). In this example, the model is accurate for soil type 3 (clays) but less accurate in differentiating clean sand (soil type 6) from silty sand (soil type 5) and silt (soil type 4).

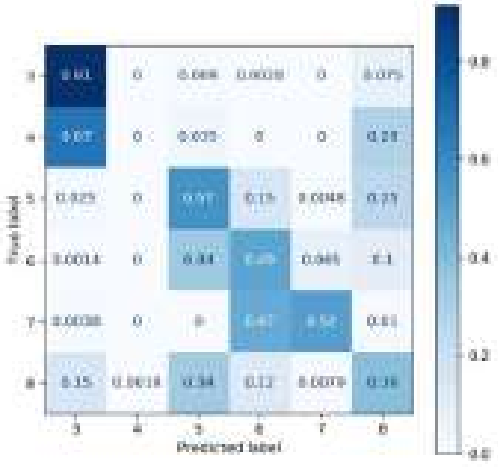

Figure 14. Confusion matrix for a linear classifier of soil type from CPT data (Stuyts, 2020).

4.4 Deep learning

Deep learning is a subtype of machine learning in which artificial neural networks are trained to predict a discrete or continuous outcome. Figure 15 shows the basic building block of a neural network. The inputs (feature values) are multiplied by weights and summed. This net input is then passed to an activation function φ to calculate the output. The activation function is usually non-linear and can be chosen to coerce the output into a specific format (e.g. the sigmoid function to ensure that outputs are between 0 and 1 or the RELU function to ensure that outputs are always positive). This basic buildling block can be combined multiple times to lead to complex networks with multiple layers. The training process for such a model consists of optimising the weights to minimise a loss function. This is done through back-propagation, in which the gradient of the loss function with respect to each weight is calculated. A gradient descent algorithm is then used to iteratively calculate the optimised weights. As neural networks become larger, this optimisation process can get very computationally expensive.

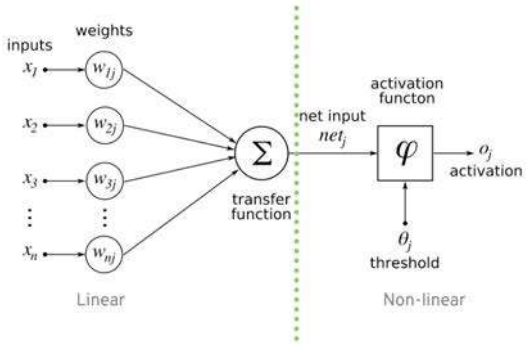

Figure 15. Basic building block for a neural network.

Neural networks can be applied to the regression and classificiation problems described above. They have been used to learn relations between geophysical data from seismic reflection surveys and CPT data (Sauvin et al., 2019). They also have applications in image or video analysis. In such cases, they use *convolutional* layers to extract features from the image. The image is encoded as a grid of pixels with an RGB-value (encoding the amount of Red Green Blue in the color) assigned to each pixel. Training a network on each individual pixel would be very time consuming. The convolutional layers apply filters which slide over regions of the image to extract features. These filters lead to a reduced set of features which make the training process more efficient. The convolutional neural networks (CNNs) have already been applied to microscope images of soil grains to identify the different minerals in a sample without requiring expensive mineralogical analysis. King et al. (2023) use a pre-trained CNN to extract the glauconite content of a sample without having to use the magnetic separator technique.

Neural networks can also be used to learn patterns from timeseries. The model then learns to predict future behaviour from past observations. LSTM

(Long short-term memory) models are popular for this type of applications. They are applied to forward predictions of pile driving by Stuyts and Suryasentana (2023).

Although deep learning is popular for building very advanced models from large datasets, geotechnical problems are often less suitable for their application. Because the datasets are small (thousands instead of millions of samples), determining a suitable network architecture and training the model can be challenging. The models can easily overfit small datasets and have more difficulty to learn general patterns from the data. This could result in model behaviour that is not physically meaningful. To enforce physically meaningful model behaviour, the governing differential equations of the physical phenomenon can be taken into account for the loss function. Model outputs which deviate substantially from what is physically meaningful are penalised in such Physics-Informed Neural Networks (PINNs). In any case, the model's accuracy and generalisation should be rigorously verified.

5 Unsupervised machine learning techniques

As data labelling may be a labour intensive process and may not be feasible within a reasonable timeframe for datasets with millions of samples, extraction of patterns from unlabelled data is often the target of a machine learning exercise. Two types of algorithms are discerned, clustering and principal component analysis which are discussed hereunder.

5.1 Clustering algorithms

For clustering algorithms, the primary goal is to distinguish clusters exhibiting significantly different behavior. For geotechnical engineers, this concept is most tangible when considering the clustering of foundation locations across a project site. These locations are grouped based on shared geotechnical conditions, such as depth to a load-bearing stratum or the presence of soft soil. The number of clusters needed to capture variations between individual location groups depends on the geological characteristics of the site. While geologically homogeneous sites may require only a few clusters, sites with strong heterogeneity may necessitate more.

In unsupervised clustering analysis, the feature space is partitioned into clusters based on similarities among individual data points. Among various clustering algorithms, the K-means clustering algorithm stands out as an intuitive choice. It calculates a generalized distance between cluster centers and each data point (as expressed in Equation 4). Optimal cluster centers are determined by minimizing the distance between points within a cluster while maximizing the distance between cluster centers. However, identifying meaningful clusters can be challenging for datasets with significant scatter.

$$\sum_{i=0}^{n} \min_{\mu_j \in C} \left(\|x_i - \mu_j\|^2 \right) \tag{4}$$

An example of such a clustering is shown in Figure 16 for the seafloor CPTs at the PEZ offshore wind farm site. The data which is associated with the surface sand layer is shown as red diamonds and the data associated with the stiff clay of the Kortrijk Formation is shown as blue circles. It is clear that the two soil type clusters can be discerned by looking at their cone tip resistance and sleeve friction. Indeed, the clays of the Kortrijk formation are expected to have a higher friction ratio.

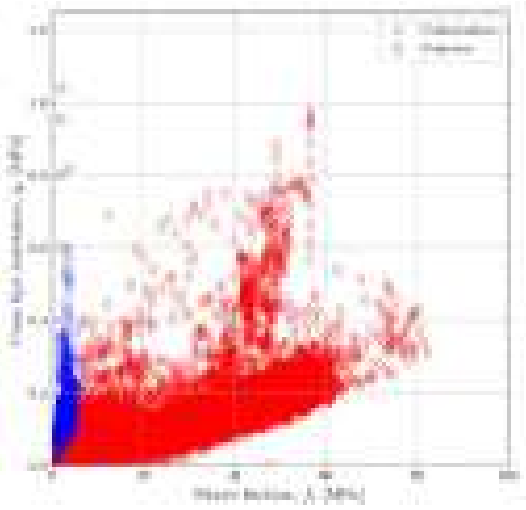

Figure 16. Clustering of CPT data from seafloor CPTs at PEZ.

When clusters have been identified, data which lies significantly outside of the clusters can be identified as *outliers*. Outliers detection algorithms are also considered as a part of unsupervised learning and are closely related to clustering algorithms. In Section 8, an example on the detection of cone resistance spikes will be illustrated on the PEZ downhole CPT dataset.

5.2 Principal Component Analysis

Geotechnical dataset often contain features which are correlated, so the dimensionality of the dataset can be reduced without losing substantial information. In Principal Component Analysis (PCA), the data from the original feature space is transformed into a new feature space with reduced dimension. The new feature space is identified by calculating a series of orthogonal unit vectors which represent the best-fitting lines through the data points. The unit vectors which captures the most significant variations are then retained while unit vectors with limited variance can be discarded.

PCA is illustrated in Section 8 on the shear wave velocity dataset. The use of Non-Negative Matrix Factorization (NMF), a technique for extracting physically meaningful feature combinations from data is also illustrated there.

6 Large Language Models

6.1 Model internal workings

Large Language Models have led to an acceleration of the use of AI across a range of applications. While AI was already widely in use at the time of publishing *chatGPT*, the user adoption of this model was unprecedented. GPT stands for *Generative Pre-trained Transformer*. These algorithms are able to generate sequences of meaningful text from a prompt. They use Transformer network architectures (Keita, 2022), a complex type of network in which word sequences are encoded by *tokenizing* words (assigning a unique number to a word or word part). The position of the word in a sentence is also encoded using a *context vector*. The Transformer architecture contains attention mechanisms to capture the contextual relations that exist between words in a given sequence.

In the pre-training phase, the transformers are trained using extensive volumes of text sourced from the internet. For instance, GPT-3 (Brown et al., 2020) was trained on 45 terabytes of text data and boasts an impressive 175 billion parameters. The expenses incurred during the model's training process are estimated to be approximately 4 million USD. During this pre-training phase, the algorithm discovers the statistical patterns of a language without ever being exposed to grammar rules.

Although chatGPT has received a lot of attention for its ability to create meaninful text output from user defined queries or *prompts*, the Transformer architecture is also able to create images or even video from text-based prompts.

6.2 LLM model enhancements

After the model is pre-trained, it can understand the basics of a language but it is not yet equipped to perform detailed tasks. During the model fine-tuning phase, the pre-trained model is used as a starting point and is then fine-tuned on a domain specific narrower dataset.

This fine-tuning often consists of a *reinforcement learning from human feedback* phase (Yang, 2023) as illustrated in Figure 17. In this phase, the pre-trained model is used to generate several outputs which are then ranked by a human agent in terms of appropriateness. A separate *reward* model is trained on this basis and this reward model is then used to further optimise the coefficients of the pre-trained LLM.

Figure 17. Explanation of Reinforcement Learning from Human Feedback (RLHF) (Yang, 2023).

Creating a LLM and fine-tuning it is beyond the capabilities of geotechnical engineers, but non-specialists can still increase their effectiveness in using LLMs by using *prompt engineering*. For specific example applications, software platforms exist which allow users to work with pre-trained LLMs and fine-tune them.

6.3 Prompt engineering

In Prompt Engineering, the user of a LLM algorithm crafts the text of the query which is fed to the LLM to ensure an optimal response. Although LLMs have a remarkable capacity for returning appropriate answer for direct questions or so-called *zero-shot prompts*, the output of an LLM can be enhanced by providing specific examples.

When asking Microsoft's CoPilot the question " **What is the CPT response for silt** ", the response in Figure 18 is returned. The response is not very targetted and includes elements of liquefaction resistance assessment which may not be relevant for all users.

Figure 18. Microsoft CoPilot response for zero-shot prompting.

By providing a couple of example answers to the LLM, the LLM can be guided to provide a more precise answer. The question from the zero-shot prompt was modified as follows: " **The CPT response for sand is a high cone resistance and hydrostatic pore pressure. The CPT response for clay is a low cone resistance and excess pore pressure. What is the CPT response for silt** ". Based on this so-called *few-shot prompt*, the response in Figure 19 is returned. Although the response may still not be appropriate for all silts, it is much more precise and returns the information which is desired by the user.

Figure 19. Microsoft CoPilot response for few-shot prompting.

A full discussion on prompt engineering is beyond the scope of this paper. The reader is referred to DIAR.AI (2024) for more details on prompt engineering techniques.

6.4 LLM fine-tuning for specific tasks

For very specific problems where domain knowledge is important, the LLMs which are available through APIs may not be sufficient. In such cases, the model needs to be fine-tuned by providing domain-specific knowledge to a pre-trained model and updating its weights. Domain-specific knowledge can be provided to an LLM by creating pairs of questions and answers which capture the knowledge that needs to be learned by the model. The pairs of questions and answers can either be specified as a prompt and an output, or additional input which contains elements of the answer can be specified with the prompt. The example below first shows a prompt and output without additional input. The second part does contain additional input.

The task of preparing prompt-output pairs can be quite labour-intensive but if accurate answers are sought, this may be indispensable.

```
{
"instruction": "Write a description of the CPT
dissipation test.",
"input": "",
"output": "A dissipation test during CPT testing
examines the decay of excess pore pressure
over time. The cone is kept at a constant depth
and the pore pressure is continuously measured.
The horizontal coefficient of consolidation can
be derived from this test."
},
{
"instruction": "Which of the following tests is
not a laboratory test?",
"input": "CU triaxial, Direct Simple Shear,
    Dilatometer, Oedometer",
"output": "Dilatometer"
}
```

When the additional prompt-output pairs have been defined, the model can be fine-tuned. The LLM transforms a text input which can be encoded as a vector $\vec{x}$ into an embedding $\vec{h}$. $\vec{h}$ is a vector which encodes the meaning and context of words. This transformation of inputs to outputs can then be expressed as a matrix multiplication (Equation 5).

$$\vec{h} = W \cdot \vec{x} \tag{5}$$

When fine-tuning the LLM, the weight matrix can be updated by summing the weight matrix W of the pre-trained model with a weight update matrix ΔW. LLMs can contain billions of weights and therefore the weight matrix can have a very large rank. Figure 20 shows two strategies for model fine-tuning. In regular fine-tuning, the weight update matrix ΔW has the same rank as the weight matrix W of the pre-trained model. Updating the model weights with this stragegy is very expensive in terms of computer time and resources. The updated embeddings can be expressed as shown in Equation 6.

To reduce the computational requirements and make LLM fine-tuning possible on a single Graphical Processing Unit (GPU), a Low-Rank (LoRa) approximation can be taken for the weight update matrix. In

this strategy, the weight update matrix ΔW is written as the product of two low-rank matrices W_A and W_B. If the weight matrix has dimensions $n \times m$, W_A is taken as a $n \times k$ matrix and W_B as a $k \times m$ matrix. The multiplication of these two matrices yields a $n \times m$ matrix (Equation 7). By choosing a low number for k, the computational requirements for the fine-tuning task are significantly reduced.

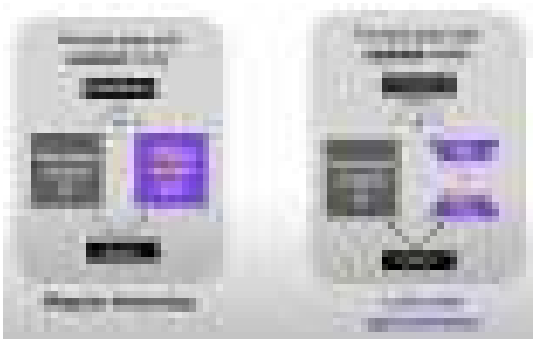

Figure 20. Explanation of Low Rank adaptation for LLM re-training (AI, 2023).

$$\vec{h}_{\text{updated}} = (W + \Delta W) \cdot \vec{x} \tag{6}$$

$$\vec{h}_{\text{updated,LoRa}} = \left(W_{n\times m} + W_{A_{n\times k}} \cdot W_{B_{k\times m}}\right) \cdot \vec{x} \tag{7}$$

Implementing LLMs and the code for fine-tuning them is a very complex task which cannot be undertaken by geotechnical engineers. However, there are software platforms such as LitGPT (AI, 2023) and Hugging Face (Hugging Face, 2023) which allow users to leverage the capabilities of pre-trained LLMs and perform model fine-tuning without having to master the underlying implementations. These platforms allow geotechnical engineers to focus on their subject matter expertise. Engineers can then concentrate their efforts on preparing high quality prompt-output pairs for fine-tuning and evaluating whether the fine-tuned model performs well.

A fine-tuned LLM is currently being developed by researchers at the Flemish Research Institute Vito. They are confronted with soil descriptions (in Dutch) which are highly dependent on the person logging the core. These descriptions need to be rationalised into a major and minor soil type in an automated manner. Because existing algorithms did not provide good results, the Dutch adaptation BERTje (De Vries et al., 2019) of the LLM BERT (Bidirectional Encoder Representations from Transformers) was fine-tuned on a number of human-processed soil descriptions using the Hugging Face platform. The initial results look promising, with superior performance to previously developed rule-based algorithms.

7 Example applications of supervised learning

In this section, the machine learning techniques of regression and classification are illustrated on basic example problems. The examples are kept relatively simple to allow engineers reading this paper to repeat the analyses themselves. As discussed in Section 3, all data is provided on GitHub.

7.1 Regression model for V_s

Deriving shear wave velocity profiles from CPT data is a common task in geotechnical parameter selection. Several correlations are proposed in the literature which all formulate a closed-form regression model to derive V_s (or G_{max}) from CPT data. Machine learning models can be trained on V_s data from S-PCPT with corresponding CPT measurements.

In this paper, two types of regression model are evaluated. First, a simple linear regression model is applied to the data. Because the relation between V_s and other features such as vertical effective stress is non-linear, feature transformation is used to linearise the relation between features and the target. Next, a more sophisticated XGBoost model is applied. The V_s dataset described in Section 3 is used for the regression.

7.1.1 Linearised model

Cha et al. (2014) identify a power-law relation between V_s and vertical effective stress. Although the authors make a distinction between the effective stress in the direction of wave propagation and the direction perpendicular to it, the relation can also be simplified in terms of the vertical effective stress as shown in Equation 8.

$$V_s = \alpha \cdot \left(\frac{\sigma'_{v0}}{1\text{kPa}}\right)^{\beta} \tag{8}$$

The coefficients α and β then depend on the packing density and mineralogy of the soil. Robertson and Cabal (2015) propose a correlation in which α has a linear relation with the soil behaviour type index I_c. The relation between V_s, σ'_{v0} and I_c can be expressed as shown in Equation 9.

$$V_s = 10^{a_0 + a_1 \cdot I_c} \left(\frac{\sigma'_{vo}}{1\text{kPa}} \right)^{\beta} \tag{9}$$

By taking the logarithm of each term in the equation, a linear relation between $\log_{10}(V_s)$ and I_c and $\log_{10}(\sigma'_{vo}/1\text{kPa})$ is obtained (Equation 10). A linear regression machine learning algorithm can be applied to optimise the coefficients a_0, a_1 and β.

$$\log_{10}(V_s) = (a_0 + a_1 \cdot I_c) + \beta \cdot \log_{10} \left(\frac{\sigma'_{vo}}{1\text{kPa}} \right) \tag{10}$$

The dataset was partitioned to use 75% of the data as training data and the remainder as a test set for checking the generalisation of the model. It should be noted that the location of the data was not considered in the partitioning. In many cases, it can be meaningful to make partitions which are based on the geospatial location of the data (Stuyts, 2020).

After optimisation, the model coefficients shown in Equation 11 are obtained as a result of the linear regression. With these coefficients, the R^2 score on the training set is 0.46 and on the test set, $R^2 = 0.48$ is obtained. Although this value is relatively low for R^2, the comparable scores on the training and test set show that the model generalises well.

$$V_s = 10^{2.0385 - 0.0694 \cdot I_c} \cdot \left(\frac{\sigma'_{vo}}{1\text{kPa}} \right)^{0.24669} \tag{11}$$

The accuracy of the model can also be represented graphically. Figure 21 shows a scatterplot of the measured V_s from the training set and the corresponding predictions. All points are colour-coded in term of the soil behaviour type index I_c to check for any trends with soil type. In case of a perfect prediction, the point lies on the grey dashed line. Inevitably, some scatter around this line is expected. An unbiased model should have a scatter which is equally distributed along the parity line. Figure 21 shows that the linear regression model shows a tendency for underprediction of higher V_s values. The model does not show any specific trends with soil types. Points of all colours are positioned equally around the parity line.

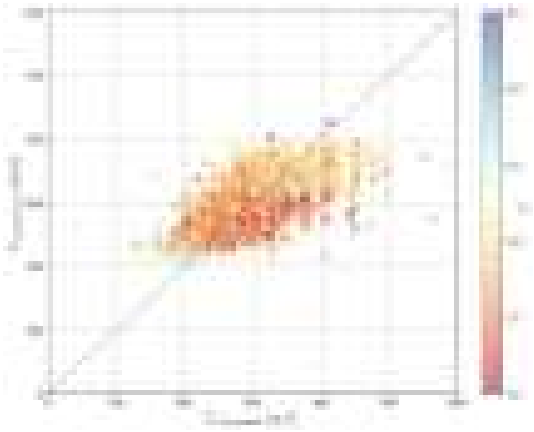

Figure 21. Graphical representation of the accuracy of the linear regression model for V_s.

The model performance can also be demonstrated on a single location. Here the CPT data from the location IJV171-SCPT from the IJmuiden Ver offshore wind farm zone is used. This location was excluded from the training data. The formula from Equation 11 is applied to the CPT data which was first averaged using a 0.5m moving average window. This averaging was necessary to take into account the 0.5m spacing of the geophones. The resulting prediction is shown in Figure 22 and can be compared with the V_s measurements from the seismic CPT. The predicted trend (in green) captures the general trend but some deviation can be observed for individual datapoints. The underprediction of V_s for depths greater than 24m is most noticeable.

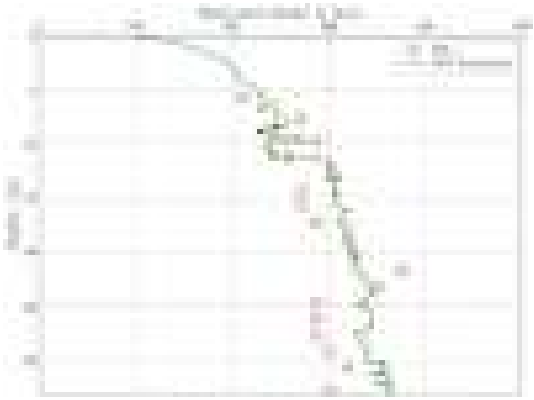

Figure 22. Prediction of V_s at location IJV171-SCPT with the linear regression model from Equation 11.

Even though the model is not too accurate, it still captures general trends. This was enforced by transforming the features according to the power-law relation which is known to be physically meaningful.

7.1.2 XGBoost model

The XGBoost algorithm (Chen and Guestrin, 2016) is a powerful algorithm based on building consecutive tree-based learners. The principle is illustrated in Figure 23. Each decision tree is trained to predict the residuals (difference between measured and predicted values) of the previous step. To avoid overfitting, *regularisation* is applied during model training which prevents the decision trees from making too many partitions. The user can also specify the maximum tree depth explicitly as a hyperparameter. The prediction error terms are not fully applied to the prediction. Rather, a *learning rate* hyperparameter is specified by the user which determines how much of the predicted residual is taken into account. This prevents overly agressive corrections to the predictions. The user sets how many iterations are performed by specifying the number of trees as a hyperparameter.

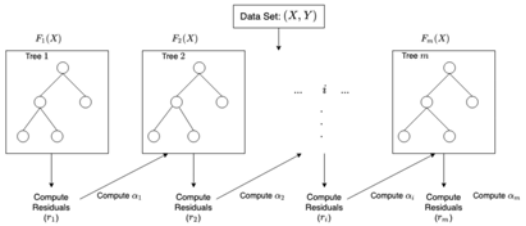

Figure 23. Underlying principle of the XGBoost algorithm (Chen and Guestrin, 2016).

For the V_s dataset, an XGBoost model with 50 trees with a maximum depth of 5 was trained. The learning rate was set to 0.1. The dataset was again split into a training set containing 75% of the data and a test set containing the remaining 25%. The model was trained on the features q_c, vertical effective stress, I_c, depth, f_s, u_2, Q_t, F_r and B_q. Because the XGBoost model is composed of consecutive decision trees, it does not have a closed-form mathematical formula.

The accuracy of the XGBoost model is shown in Figure 24. This shows a narrower spread along the parity line compared to the linear regression model. The R^2 score on the training is 0.69 which is higher than the linear regression model. The R^2 score on the test dataset is 0.49, which suggests that the model overfits the training data. This mismatch between the R^2 score of the training and test dataset can be reduced by e.g. reducing the learning rate or the maximum tree depth. This will come at the expense of a reduced R^2 score on the training dataset.

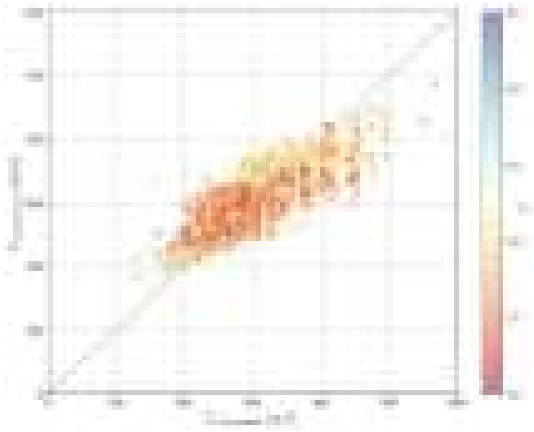

Figure 24. Graphical representation of the accuracy of the XGBoost model for V_s.

The model can again be applied to predict the V_s profile at the unseen location IJV171-SCPT. Figure 25 shows that although the model is far more complex, it provides predictions with similar deviations to the measurements compared to the linear regression model. Moreover, the model fails to capture the expected power law trend. This is especially clear for the depth range from 0 to 5m where no training data was available. The XGBoost model was not able to learn the underlying physical behaviour and should therefore be considered as an advanced interpolator inside the feature space. When the feature values deviate from those contained in the training dataset, the model will be unreliable. When extrapolation is expected, building a model which captures known physical behaviour appears to be the better approach.

Figure 25. Prediction of V_s at location IJV171-SCPT with the XGBoost model.

7.2 Classification model for the initial part of the CPT stroke

Identifying which part of the CPT trace in a downhole CPT stroke belongs to the initial build-up of resistance is a binary classification problem which can be used to illustrate machine learning classification models. The dataset of labelled downhole CPTs is used for training a linear classifier (logistic regression) and a tree-based classifier. Each of the techniques is discussed in the paragraphs below.

Before starting the machine learning modelling, meaningful features are extracted from the data. The initial build-up of resistance is limited to the first centimeters of the stroke, so the distance from the start of the stroke (Δz_{stroke}) is calculated as an additional feature. Cohesionless soils typically have higher stiffness than cohesive soils, so the soil behaviour type index I_c is also adopted as a feature. The absolute value of cone tip resistance also plays a role. At deeper depths, the build-up of resistance will go up to higher q_c values, so q_c is also retained as a feature. Finally, the initial part of the CPT stroke shows a steep increase of q_c with depth. Therefore, the cone resistance gradient with depth $\Delta q_c/\Delta z$ is calculated and retained as the final feature. The target of the classification model is a boolean which determines whether a point has to be removed from the stroke or not. A value of 1 is assigned if the point has to be removed, a value of 0 if the point is a meaningful part of the cone resistance trace which should be retained.

Figure 26 shows how the labeled data is separated in the feature space. The top panel shows that points with steeper cone resistance gradients are more likely to belong to the initial part of the stroke. In the center panel, a reduced distance for cone resistance build-up is noticed for more cohesive soils (higher I_c). The bottom panel shows that higher q_c values require larger mobilisation distances. A linear boundary is a reasonable approximation for separating the points in $\Delta q_c/\Delta z$-Δz_{stroke} space and in q_c-Δz_{stroke} space. In the I_c-Δz_{stroke} space, the boundary appears to be curved and a non-linear classifier will be required.

For the training of all classification models, the labelled data is split into a training set with 80% of the data and a test set with the remaining 20%. When performing the partition, one should ensure that there are sufficient samples for each class in the training and the test set.

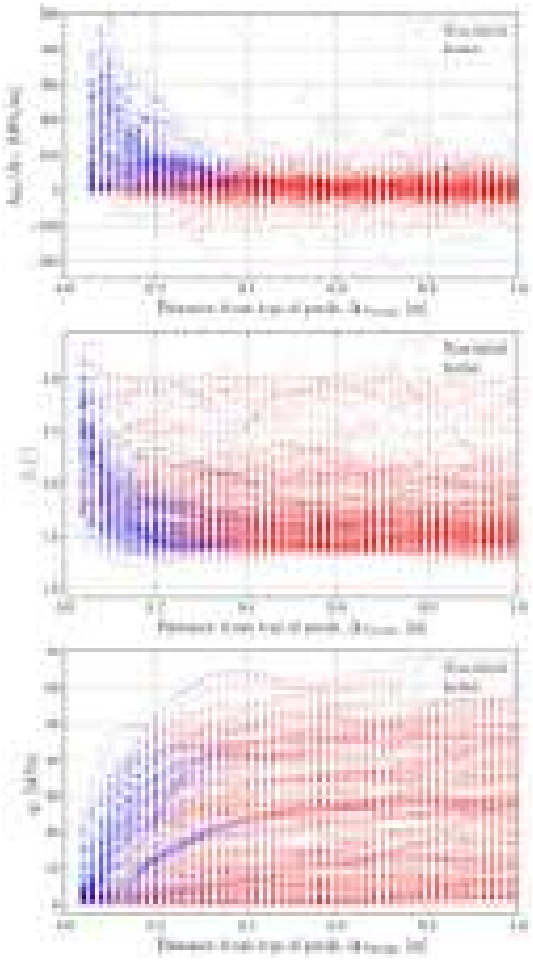

Figure 26. Overview of the labeled data for the classification problem.

7.2.1 Logistic regression

Although the name suggests otherwise, the logistic regression algorithm is actually a linear classification algorithm. For a binary classification problem, the probability of the class 1 $P(y_i = 1|X_i)$ is given in Equation 12. During the training phase, the weights w and intercept w_0 are optimised to maximise the accuracy of the model. Equation 12 immediately shows how the logistic regression algorithm provides probabilistic estimates of each class. The actual prediction of the algorithm is taken as the class for which a probability of more than 0.5 is obtained.

$$\hat{p}(X_i) = \frac{1}{1 + \exp\left(-X_i w - w_0\right)} \tag{12}$$

The predicted class is shown in Figure 27 for the training set. The predictions are shown in $\Delta q_c/\Delta z$-

Δz_{stroke} space. Because the true class is known for this data, separate points can be plotted for the following cases:

- True positives: Samples correctly classified as belonging to the initial part of the stroke;
- True negatives: Sample correctly classified as not belonging to the initial part of the stroke;
- False positives: Samples incorrectly classified as belonging to the initial part of the stroke;
- False negatives: Samples incorrectly classified as not belonging to the initial part of the stroke.

The figure shows clear clusters of points belonging to the initial part of the stroke and others not belonging to it. A region can also be identified on the boundary of the regions with true positives and true negatives where the algorithm is less certain of the predictions. Nevertheless, the model reveals a high accuracy on both the training and test set of 0.975 and 0.970 respectively. It should be noted that this number is slightly misleading because the majority of the data belongs to the non-initial part of the stroke. Hence, there are a lot of true negatives which increase the accuracy score. Figure 26 shows that the initial part of the stroke always happens within 0.6m from the start of the stroke. If the accuracy score is only calculated for this data, accuracies of 0.860 and 0.847 are obtained for the training and test set. As the scores are close together for the training and the test set, it can be said that the model generalises well.

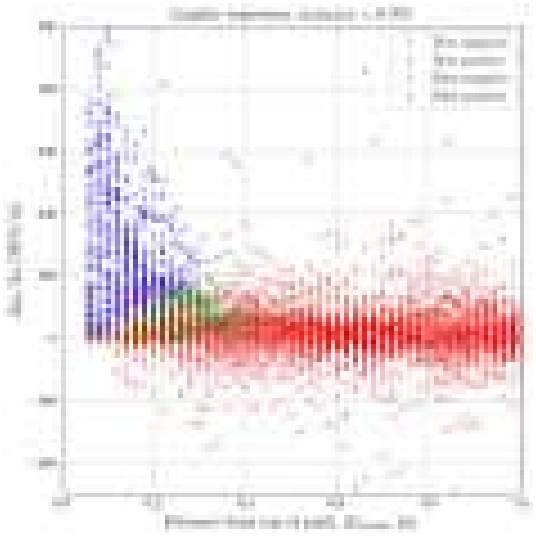

Figure 27. Results of the logistic regression classification.

The region where the model is less certain can be visualised even better by plotting the probabilistic prediction with the logistic regression model. Figure 28 shows which probability the ML model assigns to each sample in terms of belonging to the initial part of the stroke. Each point is color-coded with the calculated probability. A probability of 1 means that the point certainly belongs to the initial part of the stroke, a probability of 0 means that the point certainly does not belong to the initial part of the stroke. The plot clearly shows the transition between the two regions which is seen as a gradual changing of the colors in the region of overlap between the two clusters.

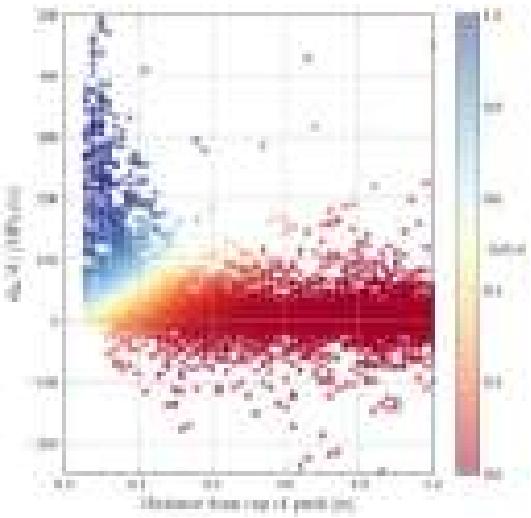

Figure 28. Probabilistic results of the logistic regression classification.

The model can be evaluated on an example CPT trace. Figure 29 overlays the probabilistic prediction of whether a point belongs to the initial part of the stroke with the actual CPT trace. The figure shows that the model performs well, especially when there is a sharp contrast between $\Delta q_c/\Delta z$ in the initial part of the stroke and the remaining part. When the transition is more gradual, the model will assign an intermediate probability to the points. This is clearly visible in the final stroke.

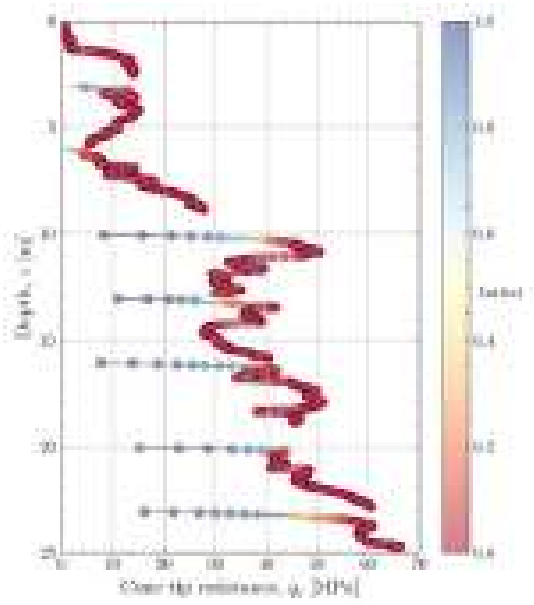

Figure 29. Probabilistic results of the logistic regression classification applied to a single CPT trace.

7.2.2 Decision trees

In a decision tree classifier, a tree is built where the data is split according to a splitting criterion which is aimed at dividing the data into two homogeneous groups. The splitting criterion is determined as the one which maximises the information gain. Information gain is high when the groups of data which remain after splitting are homogeneous, ideally containing only samples of one class. Unless specified otherwise, the splitting will continue until perfectly homogeneous groups are obtained. The user can control a number of hyperparameters of the tree building. For example, the maximum depth of the tree (the number of times splitting is performed) can be specific as a hyperparameter.

A decision tree model with a maximum depth of 3 and another model with a maximum depth of 5 are built for the downhole CPT dataset. The decision tree with maximum depth of 3 is displayed graphically in Figure 30. The first split checks whether the distance from the top of the stroke is less then 0.28m. It partitions the data into two *nodes* which are then split further until the maximum depth of the tree is reached. The nodes at the end of the decision tree are called *leaves*. A class is assigned to each leaf based on which class label forms the majority of samples in the leaf. When looking at the deeper splits, it can be seen that they happen either on the distance from the top of the stroke or on $\Delta q_c/\Delta z$. Points with a distance from the top of the stroke less then 0.28m and $\Delta q_c/\Delta z >$ 37.137MPa/m are most likely to belong to class 1 (part of the initial build-up of resistance in the stroke). The *gini* score shown in the tree in Figure 30 is a measure of the impurity of the node. If all samples at a node are contained in one class, the gini value is low. If the samples are evenly distributed between both classes, the gini score is close to 0.5 and the node is said to be impure. Figure 30 reveals several leaf nodes which still have a high impurity.

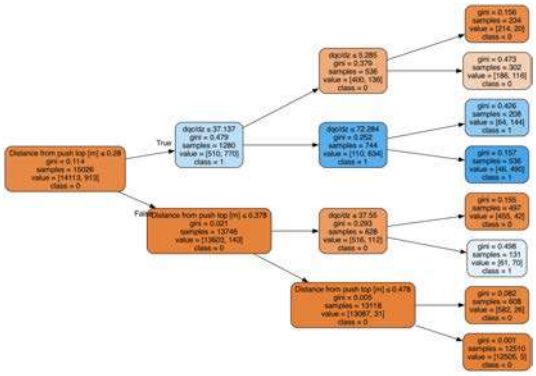

Figure 30. Decision tree with maximum depth of 3.

Figure 31 show the results for the training set for each model. The accuracy for the model with a maximum depth of 3 is equal to that of the logistic regression model. For the model with a maximum depth of 5, the accuracy score is only marginally improved. Comparing Figure 31 and Figure 27 shows that the false negatives are located in a narrower band for the decision tree models. The accuracy scores for the model with maximum depth of 3 are 0.975 and 0.973 for the training and test set respectively. For the model with maximum depth of 5, the accuracy score is equal to 0.982 for the training set and 0.973 for the test set. Although the model with maximum depth of 5 has a higher accuracy on the training set, it does not improve on the test set. It can be concluded that this model slightly overfits the data.

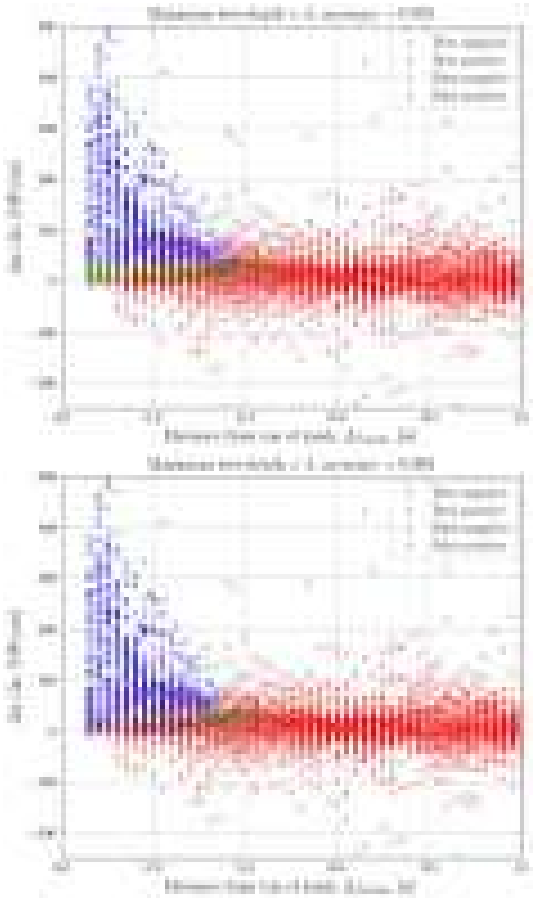

Figure 31. Results of the decision tree classification.

8 Example applications of unsupervised learning

When datasets are not labeled, meaningful discoveries can still be made from the data. In unsupervised learning, patterns will be learned from the data. An example of anomaly detection from CPTs is provided as well as an example of the extraction of meaningful information from the S-PCPT dataset.

8.1 Detection of cone resistance spikes in the PEZ dataset

The stiff clays which form the majority of the foundation subsoil for the Princess Elizabeth Zone are relatively uniform in terms and strength. The CPT traces show relatively uniform q_c profiles (e.g. Figure 9). However, in the geophysical surveys (GeoXYZ and G-tec, 2023), reflections are noticed which could be marker horizons in the Kortrijk Formation. To check whether the reflectors can be correlated to higher resistances in the CPT trace, it is necessary to identify which cone resistance values in the Kortrijk Formation are significantly higher than the cone resistances for the surrounding clay. This is a task of outlier detection for which specific machine learning algorithms exist.

Isolation forests (Liu et al., 2012) are a type of machine learning model which randomly builds a large number decision trees. The trees select random features for the splitting and then select a value between the feature minimum and feature maximum for the split. The splitting continues until individual samples are isolated. The underlying assumption is that it will take a large amount of random splits to separate samples which belong to the normal part of the population. Outliers will however be split off relatively quickly. This is illustrated in Figure 32. In the example in the figure, the outlier is split off with just one partition. Splitting off the normal point marked in red requires four partitions.

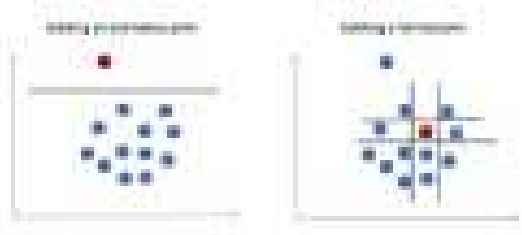

Figure 32. Illustration of the principle of Isolation Forests (Mavuduru, 2021).

Although outlier detection is relatively straightforward in two-dimensional space, identifying outliers in larger dimensional spaces can be challenging. For certain features, the values for a given sample may be within the normal range of values whereas for others, they may deviate significantly. Outlier detection techniques provide a way to identify anomalies in higher dimensional spaces. For the Princess Elizabeth Zone CPT data, outliers are identified in a seven-dimensional space using q_c, f_s, u_2, Q_t, F_r, R_f and I_c as features. It should be noted that depth below mudline is not used as a feature, as the clay formations starts at varying depths due to the varying amounts of sand cover.

Isolation forests can be built for the CPT dataset from the Princess Elizabeth Zone. First, the data in the Kortrijk Formation is identified by assessing at which depth the soil behaviour type index I_c becomes larger than 2.7. The proportion of the CPT profile below this depth is retained for further analysis. Since

the data is downhole CPT data, the initial parts of the stroke are removed using the logistic regression classification algorithm discussed in the previous section.

Next, the isolation forest trees are built. The model has four hyperparameters which can be set to obtain good quality results:

- Contamination: Sets the threshold for anomaly scores. The highest percentage of anomaly scores are retained as outliers. For the PEZ CPT dataset, the contamination was set to 0.005;
- Number of estimators: The number of trees which are built by the algorithm. As the PEZ dataset is quite large with 117732 samples, this hyperparameter is set to 1000;
- Maximum number of samples: Individual trees can be trained on a subset of the dataset. This randomisation avoids certain samples having too much weight in the tree-building process. This hyperparameter is set to 0.9 to retain 90% of the data for each tree;
- Maximum number of features: The number of features which is used for building the trees can also be changed for each tree. Randomising the selected features avoid one particular feature from having too much weight. For the example, four out of seven features are retained for each tree.

The algorithm identifies 589 outlier which can be visualised in q_c-z space (Figure 33). The plot shows that the majority of outliers have q_c values which are higher than the normal population. However, a number of outliers exist which display normal q_c values. These samples would have to be inspected by an engineer to check if they have to be removed from the trace or not. These points are less likely to correspond to a hard layer. There are also a number of outliers which lie below the normal q_c range. These points correspond to samples which are in the initial part of the stroke but which were incorrectly classified by the classification model.

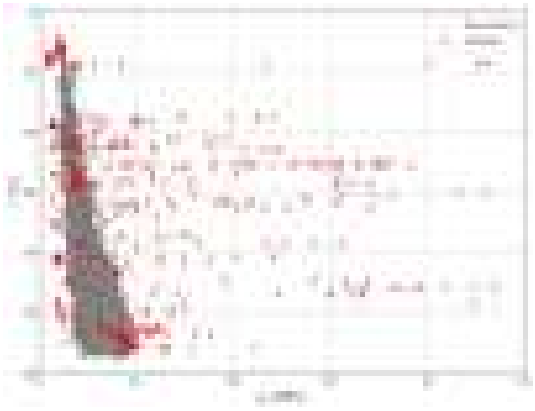

Figure 33. Visualisation of outliers identief in the PEZ CPT dataset.

Location-specific cone resistance traces can be plotted to check the effectiveness of the outlier detection algorithm for identifying hard layers. Figure 34 shows the results for three selected locations. These results show that the algorithm can make a distinction between anomalies of varying amplitude. Larger spikes are detected as outliers whereas smaller spikes are treated as normal data. This shows that Isolation Forests can be used as a technique to rationalise the identification of outliers. When done by a human, the identification of samples as outliers would vary from person to person.

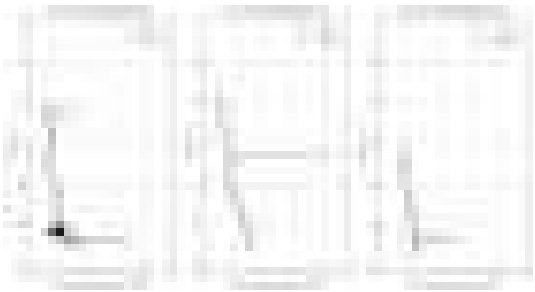

Figure 34. Visualisation of outliers identief in the PEZ CPT dataset.

8.2 Reduction of dimensionality on the S-PCPT dataset

When datasets contain a large number of correlated features, the information contained in the dataset can be expressed in a space of reduced dimensions. Transformations are applied to the features to identify *Principal Components*, a set of orthogonal components which explain the maximum amount of variance. As the dataset which relates V_s with associated CPT features contains several correlated features (depth be-

low mudline, q_c, q_t, f_s, u_2, Q_t, F_r, B_q, I_c, vertical total stress, vertical effective stress and total unit weight), Principal Component Analysis (PCA) can be performed on the data. In this technique the covariance matrix of the data is computed and the eigenvectors of this covariance matrix are the Principal Components. Although the example PCA explained in this section works with the data without transforming it, any non-linear correlations between features will compromise the results. Feature transformation could be considered to linearise the correlations.

For the 11 features of the S-PCPT dataset, 11 principal component are calculated and the variance associated with each component is plotted in Figure 35. The results show that the majority of the variance is contained in the first five components.

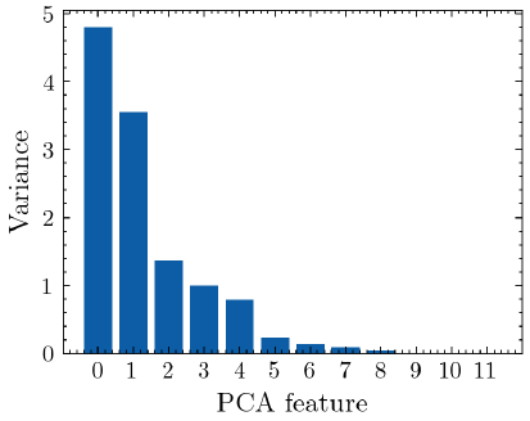

Figure 35. Variance for each of the principal components of the S-PCPT dataset.

The data can then be transformed to the space of the principal components and the first five components can be retained as a representation of the data with reduced dimension which still preserves the majority of variance. A scatterplot of the data in terms of the first two PCA components is shown in Figure 36. When comparing the data in the PCA feature space with the data in the original feature space, reduced scatter can be observed in the PCA feature space. The scatter is however still significant which can be explained by the non-linear correlations between the features.

A PCA transformation can be a useful step when working with large datasets. By using a reduced number of PCA features, building classification or regression models with the transformed data is less computationally expensive.

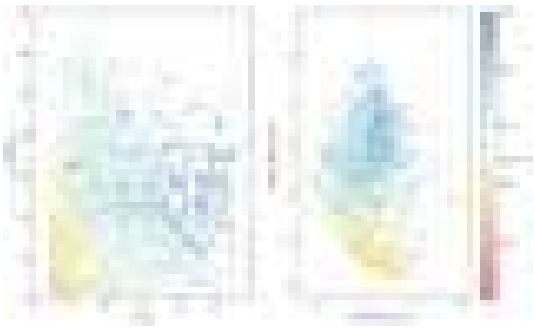

Figure 36. S-PCPT data plotted in the original and PCA feature space.

One of the drawbacks of PCA is that the PCA components are not straightforward to interpret. The physical meaning of PCA features cannot be directly understood from the PCA output. Alternative techniques exist which do allow a physical interpretation of the output. Non-Negative Matrix Factorisation (NMF) is a technique which approximates a non-negative feature matrix with the product of two non-negative matrices. If the feature matrix has n samples and m features (an $n \times m$ matrix), the two non-negative matrices are a $n \times k$ and $k \times m$ matrix, where k is the selected number of NMF components. To allow this technique to be applied, features that have negative values need to be transformed. A min-max transformer assigns a value between 0 and 1 for each feature where 0 corresponds to the minimum of the feature and 1 corresponds to the maximum. The advantage of NMF is that the importance of each original feature in the NMF features is a direct output of the algorithm. Figure 37 shows the importance of the features of the S-PCPT dataset in the four selected NMF features. Not only can the feature importance be observed, the physical meaning of the NMF features is often directly interpretable. In the example, the following physical meaning can be assigned to the NMF features:

- NMF feature 1: High importance of the features depth below mudline, vertical effective and total stress. Cone resistance terms and I_c are represented to a lesser extent. This feature can be associated with the stress conditions for a considered point;

- NMF feature 2: High importance of the cone resistance terms q_c and q_t. Lesser but non-negligible importance is observed for f_s (which is often highly correlated with q_c) and Q_t and B_q, which both contain a cone resistance component

in their formula. This feature is associated with cone resistance;

- NMF feature 3: High importance for the soil behaviour type index and other features which allow the distinction between soil types (e.g. u_2 and B_q allow for a distinction between cohesionless and cohesive soil). This feature is associated with the soil type;

- NMF feature 4: High importance for the total unit weight and lesser importance for Q_t, I_c and B_q. The physical meaning of this feature is less clear. Often the interpretability reduces with increasing NMF component number. Determining a suitable number of components is often an iterative process.

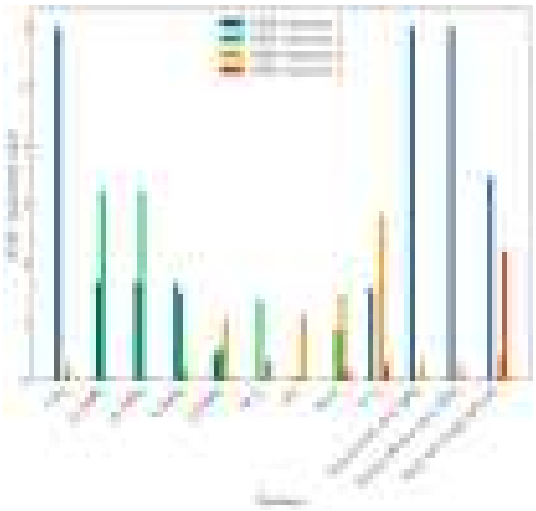

Figure 37. Interpretation of the NMF features in terms of original features.

Similar to PCA, the data can be transformed to NMF feature space to obtain an approximation of the data with reduced feature dimension.

9 Conclusions

As geotechnical site investigations are data-intensive campaigns which provide data of varying quantify and quality, machine learning techniques have a wide range of application. When properly used, these techniques can help the geotechnical engineer to discover patterns in the data or to build predictive models based on site investigation data. In this contribution, an overview of machine learning algorithms for supervised and unsupervised learning is provided. The techniques are illustrated using problems from offshore geotechnical site investigations. Geotechnical engineers can choose between a wide range of available algorithms. The engineer should understand how to formulate a problem involving site investigation data as a machine learning problem and how to select the best available techniques for the task at hand.

When building machine learning models, the engineer should combine insight in the underlying principles of the machine learning model with subject matter expertise and a good understanding of the data. The quality and quantify of available data will determine whether a machine learning modelling exercise will be successful. Careful evaluation of the model quality metrics will also allow the engineer to judge whether a model is fit for purpose or not. Machine learning models may or may not capture the underlying physical behaviour of a problem and the model output should be checked to see if this is the case.

Acknowledgements

The authors would like to acknowledge the support of the Belgian Ministry of Economic Affairs through the ETF project WINDSOIL project. The support of VLAIO through the De Blauwe Cluster SBO SOILTWIN project is also acknowledged. Geotechnical data from the Dutch and German offshore wind farms was used under a Creative Commons License.

References

L. AI. Lit-gpt. `https://github.com/Lightning-AI/lit-gpt`, 2023.

R. D. Andrus, N. P. Mohanan, P. Piratheepan, B. S. Ellis, and T. L. Holzer. Predicting shear-wave velocity from cone penetration resistance. In *Proceedings of the 4th international conference on earthquake geotechnical engineering, Thessaloniki, Greece*, volume 2528, 2007.

Association of Geotechnical and Geoenvironmental Specialists. Electronic Transfer of Geotechnical and Geoenvironmental Data. Technical Report Edition 4.0.4, 2017.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

A. Cadden and B. Keelor. Implementation and Transition of Data Interchange for Geotechnical and Geoenvironmental Specialists (DIGGS v2.0). Technical Report State Job Number 26047, 2017.

M. Cha, J. C. Santamarina, H.-S. Kim, G.-C. Cho, et al. Small-strain stiffness, shear-wave velocity, and soil compressibility. *J. Geotech. Geoenviron. Eng*, 140(10):06014011, 2014.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

J.-P. Chiles and P. Delfiner. *Geostatistics: modeling spatial uncertainty*, volume 713. John Wiley & Sons, 2012.

W. De Vries, A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, and M. Nissim. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*, 2019.

J. Deng and Y. Lin. The benefits and challenges of chatgpt: An overview. *Frontiers in Computing and Intelligent Systems*, 2(2):81–83, 2022.

DIAR.AI. Prompt Engineering Guide, 2024. URL `https://www.promptingguide.ai/`.

R. Fearon and M. Coop. Reconstitution: what makes an appropriate reference material? *Géotechnique*, 50(4):471–477, 2000.

GeoXYZ and G-tec. Geophysical Results Report - FOD Economie - Princess Elisabeth Zone. Technical Report BE4341H-22005_PEZ-RR-01, Nov. 2023.

Y. A. Hegazy and P. W. Mayne. A global statistical correlation between shear wave velocity and cone penetration data. In *Site and geomaterial characterization*, pages 243–248. 2006.

H. Hotz. RAG vs Finetuning — Which Is the Best Tool to Boost Your LLM Application?, 2023. URL `https://towardsdatascience.com/rag-vs-finetuning-which-is-the-best-tool-to-boost-your-llm-application-94654b1eaba7`.

Hugging Face. Fine-tune a pretrained model, 2023. URL `https://huggingface.co/docs/transformers/training`.

M. Jamiolkowski, D. Lo Presti, and M. Manassero. Evaluation of relative density and shear strength of sands from cpt and dmt. In *Soil behavior and soft ground construction*, pages 201–238. 2003.

K. Karkov, E. Dalgaard, A. Diaz, H. Duarte, H. Hansen, S. Hviid, N. H. van Gilse, L. Krogh, S. Kuppens, G. Salaün, et al. Case study: Avo inversion and processing of ultra-high resolution seismic for a windfarm application. In *83rd EAGE Annual Conference & Exhibition*, volume 2022, pages 1–5. European Association of Geoscientists & Engineers, 2022.

Z. Keita. An Introduction to Using Transformers and Hugging Face, 2022. URL `https://www.datacamp.com/tutorial/an-introduction-to-using-transformers-and-hugging-f`

T. King, T. Wuenscher, L. Griffiths, and A. Fraps. A novel approach to quantifying glauconite content in soils using digital image analysis. In *Proceedings of the 9th Int. Conf. on Offshore Site Investigation and Geotechnics*, London, U.K., 2023.

F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(1):1–39, 2012.

T. Lunne, T. Berre, and S. Strandvik. Sample disturbance effects in deep water soil investigations. In *SUT Offshore Site Investigation and Foundation Behaviour New Frontiers: Proceedings of an International Conference*, pages SUT–OSIFB. SUT, 1998.

T. Lunne, J. J. Powell, and P. K. Robertson. *Cone penetration testing in geotechnical practice*. CRC Press, 2002.

A. Mavuduru. How to perform anomaly detection with the Isolation Forest algorithm, 2021. URL `https://towardsdatascience.com/how-to-perform-anomaly-detection-with-the-isolation`

P. Mayne and G. Rix. Gmax-qc relationships for clays. *Geotechnical Testing Journal GTJODJ*, 16 Nr 1:54–60, 1993.

Z. Ouyang and P. W. Mayne. Effective friction angle of clays and silts from piezocone penetration tests. *Canadian Geotechnical Journal*, 55(9):1230–1247, 2018.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

K.-K. Phoon and W. Zhang. Future of machine learning in geotechnics. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, 17(1):7–22, 2023.

S. Raschka. *Python machine learning*. Packt publishing ltd, 2015.

L. C. Reese, W. R. Cox, and F. D. Koop. Analysis of laterally loaded piles in sand. In *Offshore Technology Conference*. OnePetro, 1974.

L. C. Reese, W. R. Cox, and F. D. Koop. Field testing and analysis of laterally loaded piles om stiff clay. In *Offshore technology conference*. OnePetro, 1975.

G. J. Rix and K. H. Stokoe. Correlation of initial tangent modulus and cone penetration resistance. In *Calibration chamber testing. New York: Elsevier*, pages 351–362, 1991.

P. Robertson and K. L. Cabal. Guide to Cone Penetration Testing. Technical report, 2015.

G. Sauvin, M. Vanneste, M. E. Vardy, R. T. Klinkvort, and F. Carl Fredrik. Machine learning and quantitative ground models for improving offshore wind site characterization. In *Offshore Technology Conference*, page D021S016R004. OTC, 2019.

B. Stuyts. Data science applications in geointelligence. In *Proceedings of the Fourth International Symposium Frontiers in Offshore Geotechnics*, Austin, Texas, 2020.

B. Stuyts and S. K. Suryasentana. Applications of data science in offshore geotechnical engineering: State of practice and future perspectives. In *SUT - Proceedings of the 9th International Conference*, London, U.K., 2023.

B. Stuyts, V. Vissers, D. Cathie, C. Jaeck, and S. Dörfeldt. Optimizing site investigations and pile design for wind farms using geostatistical methods: a case study. Perth, WA, 2010. Balkema.

B. Stuyts, W. Weijtjens, and C. Devriendt. Development of a semi-structured database for back-analysis of the foundation stiffness of offshore wind monopiles. *Acta Geotechnica*, 18(1):379–393, Jan. 2023. ISSN 1861-1133. doi: 10.1007/s11440-022-01551-3. URL `https://doi.org/10.1007/s11440-022-01551-3`.

D. Yang. Lecture 4: Learning from Human Feedback, 2023. URL `https://web.stanford.edu/class/cs329x/slides/scribe_human_feedback.pdf`.