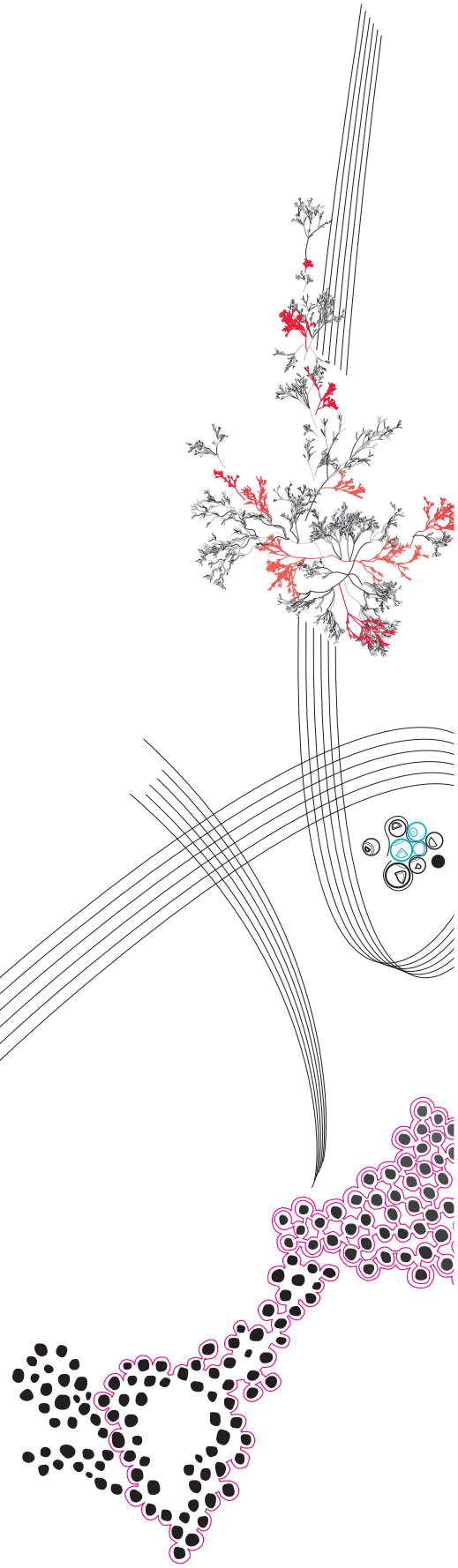

MSc Business Information Technology
Final Project

# Soil organic carbon mapping for farms in the Netherlands

David van de Giessen

Graduation Committee:
Dr. A. Abhishta
Dr. L. Ferreira Pires

April, 2023

Faculty of Electrical Engineering,
Mathematics and Computer Science,
University of Twente

UNIVERSITY OF TWENTE.

# Executive summary

Climate change and global warming is an important and well-discussed topic the past few years. Reducing greenhouse emissions such as carbon dioxide emission is a measure against global warming. The agriculture sector can contribute to the reduction of carbon dioxide emission by storing organic carbon in the soil (carbon sequestration). Farmers can adjust their cultivation plans based on the state of the soil. Therefore it is necessary that farmers gain insights on the state of the soil, specifically the soil organic carbon content. In current practice, the insights on soil organic carbon content are mainly based on soil samples. However, taking soil samples is an expensive practice. There is an increased need for insights into the soil quality without taking soil samples. This thesis presents a method to estimate the soil organic carbon content on farms in the Netherlands.

The study first provides a comprehensive systematic literature review to understand the state of the art and best practices used in similar studies. The results of this literature review include a set of environmental covariates, categorized using the SCORPAN methodology, a set of best performing prediction methods and an analysis of the validation strategies used in similar studies. The results of the literature review are used for designing the artifact (prediction method).

The second part of the study is the design and development of the artifact. During this research, we have followed the steps of Design Science Research Methodology (DSRM). For the design and development of the artifact, we used the method Knowledge Discovery in Databases. The first steps of this method are data selection, pre-processing and transformation. Based on the results of the literature review, we have gathered and used the following environmental covariates as prediction features: climate data (temperature, rainfall, solar radiation), soil property data (underlying soil type), vegetation data (NDVI), land usage (cultivation plan) and an existing soil organic carbon stocks and density map (SoilGrids).

The next step of the KDD method (Data mining) is performed by choosing, implementing, validating and optimizing the prediction methods. Based on the results of the literature review, we have implemented Random Forest, Support Vector Machine and Artificial Neural Networks as prediction algorithms and validated these algorithms using 10-fold cross validation. From these prediction algorithms, Random Forest showed the highest predictive accuracy ($R^2$ = 0.37, MSE = 1.76). An important note to these results is that, due to a lack of measurements, the model is likely not generalizable to more or other farm plots.

There are a few takeaways from this project. First of all, the environmental covariates used as prediction features show predictive potential and can be used in future research to improve the prediction model. The research also resulted in methods and sources that can be used to gather these types of data (i.e. the vegetation index based on satellite imagery), which both contributes to research and practice. Furthermore, the next takeaway is that the Random Forest has predictive potential and performs the best for the soil organic carbon computation. The last takeaway is from this study is the analysis of the potential value of the artifact: farmers receive a more up-to-date view on the state of the soil which can be used for a more tailored advice on how to increase the soil state and reduce carbon emission.

# Preface

This master thesis is the end of my journey at the University of Twente. After finishing the Bachelor's degree of Business Information Technology, I did the Master's degree Data Science and Business. This thesis is the final project of the Master's degree.

The project is performed for Inversable B.V., a small IT company in Deventer. Inversable tries to help customers by enabling them to benefit from standard open-source IT building blocks. These customers are mostly situated in the agriculture, education, governance, energy and compliance sectors. This research is performed for the customers in the agriculture sector. Inversable is co-owner of IntoAgri B.V. together with experts in the agriculture sector. This is a farmers' organization that tries to innovate software in the agricultural sector. The aim of the company is to help dairy farmers optimize the usage of their data.

This led to the initial question of the company: what can we learn from the data gathered by and for farmers? As this is a broad question, I tried to narrow it down to a question that is more specific and has societal value. Based on an interview with the owners of IntoAgri, we determined that there is a need of insights into the quality of soils, specifically for the organic carbon content. This was the basis for our research question.

Before presenting the rest of this thesis, I want to acknowledge several important people that helped during this thesis. First of all, I want to acknowledge the employees of IntoAgri. I specifically want to thank Steven for his guidance, Bart for his availability for and his help during the project and Robin for the daily standups and feedback.

I want express my gratitude to my supervisors, dr. Abhishta and dr. Luís Ferreira Pires, for their support, feedback, guidance and patience during the project. I want to thank dr. Abhishta for the meetings where he was patient with the project and for the way he comforted me and gave me self-esteem needed to complete the project. I want to thank dr. Luís Ferreira Pires for his flexibility after joining the project in a later stadium and immediately creating time to give detailed feedback.

During this research, several external organizations helped me in performing this study. I want to thank Arjan Reijneveld from Eurofins-Agro for the advices he gave me during the meetings and the work he put in to provide me with the carbon content measurements. Furthermore I want to acknowledge Sven Verweij for his advices on the project and the time he had for me when I had agricultural questions. Lastly, I want to acknowledge the help of Gerard van der Zwet from Studielab who provided me with extra guidance during the project.

I now invite you to read the rest of the thesis and I hope you enjoy reading it.

David van de Giessen

# Contents

# List of Figures

# List of Tables

# List of Acronyms

| | |
|---|---|
| **AI** | Artificial Intelligence |
| **ANN** | Artificial Neural Network |
| **BOA** | Bottom Of Atmosphere |
| **BRO** | Basisregistratie Ondergrond |
| **BRP** | Basisregistratie gewaspercelen |
| **$CO_2$** | Carbon Dioxide |
| **DNN** | Deep Neural Network |
| **DSM** | Digital Soil Mapping |
| **DSRM** | Design Science Research Methodology |
| **GSOC** | Global Soil Organic Carbon |
| **KDD** | Knowledge Discovery in Databases |
| **ML** | Machine Learning |
| **MLR** | Multiple Linear Regression |
| **MSE** | Mean Squared Error |
| **NDVI** | Normalized Difference Vegetation Index |
| **NIR** | Near Infrared |
| **NIRS** | Near Infrared Spectroscopy |
| **PDOK** | Publieke Dienstverlening Op de Kaart |
| **RD** | Rijksdriehoek |
| **RF** | Random Forest |
| **RQ** | Research Question |
| **SDG** | Sustainable Development Goals |
| **SOC** | Soil Organic Carbon |
| **SOM** | Soil Organic Matter |
| **SVM** | Support Vector Machine |
| **SVR** | Support Vector Regressor |
| **TOA** | Top Of Atmosphere |
| **WCS** | Web Coverage Service |

# 1. Introduction

Modern agriculture has to cope with several challenges, such as climate changes and the increasing call for food [1]. The storage of soil organic carbon on farm plots is a strategy to reduce the carbon dioxide emission, which is a measure against climate changes. Insights from increased monitoring of the soil carbon stocks at farms can lead to a more active approach to carbon storage that farmers can adopt. However, researchers of the University of Wageningen state that the quality of the soil, including the soil organic carbon content, is sometimes hard to determine without sampling the soil. Soil does not have a 'passport' that indicates whether the soil is of high quality or not [2]. The easiest solution to this is to increase the soil sampling frequency. However, taking soil samples is expensive for farmers. Thereby the call for insights into the soil quality without taking expensive samples arises.

## 1.1 Problem statement

In order to address this call for getting soil insights without sampling, the agriculture sector has to adopt appropriate technologies, such as digital soil mapping (DSM). DSM is seen as an effective method to model soil properties based on the quantitative relationship between soil observations and environmental predictors [3].

The practice of digital soil mapping in the context of soil organic carbon computation has a large potential. The amount of open data and gathered data increases every day and should be utilized as much as possible. However, based on the information we got from a farmers' organization we interviewed, the farmers currently do not use initiatives that predict the soil organic carbon content. Furthermore, we did not find initiatives that reach the level of practical usage. Initiatives such as the Global Soil Organic Carbon map (GSOC map) [4] and SoilGrids [5] provide estimations of the carbon stocks for the entire world. However, these maps include snapshots of the carbon stocks taken in 2020 and do not provide up-to-date information about the current soil state on a local (farm scale) level. To be useful, these carbon maps should be enriched with open environmental data and give a more up-to-date view on the carbon state of the farm plots. Hence, we identified a knowledge gap: how can one accurately predict the current soil organic carbon content on a farm plot scale in the Netherlands based on existing soil maps and environmental covariates?

## 1.2 Research goal

The goal of this research is to design a soil organic carbon prediction method that can be used to map the current soil organic carbon state for farms. Based on the problem typology of Wieringa [6], the stated problem is a **design problem**: the prediction method is the artifact that will interact with the context of farms in the Netherlands and available open data.

### 1.2.1 Solution objectives

The main objective of the artifact is to deliver a prediction on the soil organic carbon content on plots of farms in the Netherlands. However, to accomplish this objective, different research objectives have been established. The literature review we performed (further discussed in Chapter 2) showed that machine learning models outperform standard regression models in recent studies. Therefore, the objective of this research was to evaluate the potential of machine learning models to predict the soil organic carbon content. However, to achieve this objective, several different objectives or questions arise. Assuming the machine learning models show a good prediction potential, we want to identify

relevant environmental variables that give a good indication on the carbon stocks and machine learning models with the highest predictive accuracy.

## 1.3 Research question

As the goal of this research is to develop a soil organic carbon prediction method using machine learning models, we have formulated the following main research question according to the design science template:

*How to develop a method that predicts the soil organic carbon content based on environmental covariates at farms in the Netherlands so that the company can provide actionable insights on the carbon stocks to farmers?*

### 1.3.1 Research sub-questions

In order to answer the main research question, we have defined five research sub-questions.

- **RQ1:** *What covariates can be used as features for the soil organic carbon content prediction?*
- **RQ2:** *What prediction methods can be used to predict the soil organic carbon content and how can they be validated?*
- **RQ3:** *Using covariates identified from the literature, how can an empirical model be developed to predict the soil organic carbon content?*
- **RQ4:** *What is the predictive accuracy of the developed model and how can it be optimized?*
- **RQ5:** *How can the resulting model be used to create potential value for stakeholders?*

## 1.4 Research Methodology

In this research we used the Design Science Research Methodology (DSRM) developed by Peffers et al [7]. It consists of six steps (or activities): *Identify Problem & Motivate, Define Objectives of a Solution, Design & Development, Demonstration, Evaluation and Communication*, as shown in figure 1.

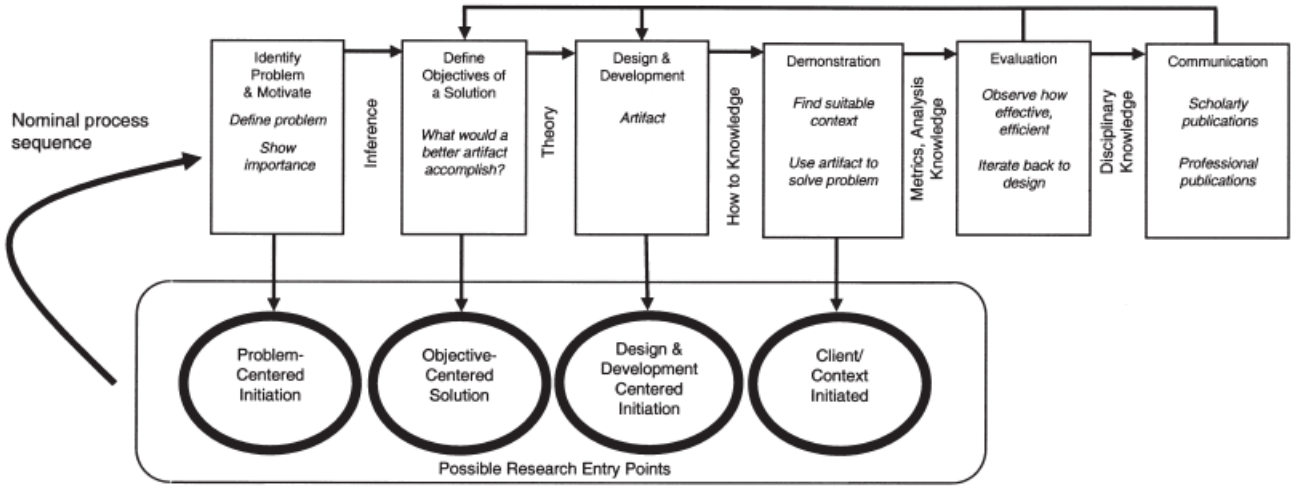


**Figure 1. DSRM process flow**

The activities *Identify Problem & Motivate* and *Define Objectives of a Solution* are performed at the start of the research and are explained earlier in this chapter.

In the *Design and Development* step, we have used the Knowledge Discovery in Databases (KDD) method developed by Fayyad[8]. This method is adequate for this step as its intent is to harvest information by recognizing patterns in raw data [9]. This goal matches the goal of this research: developing a model that recognizes patterns and predicts the soil organic carbon content based on this patterns. The KDD methodology distinguishes five phases that the researcher and thereby the state of the data go through: *Selection, Pre-processing, Transformation, Data Mining and Interpretation/Evaluation.*

In the *Demonstration* step the artifact is used to solve one or more instances of the problem. To do this, the prediction model (artifact) has been tested on other data than the prediction model is trained on, using the concept of cross-validation. The *Evaluation* step automatically follows after the demonstration step. Testing the data in the demonstration phase results in different performance measures. These measures are the $R^2$-value and the Mean Squared Error of the predictions made by the artifact. The *Communication* step is performed via this thesis presenting how we designed the prediction method and discussing the performance of the method.

## 1.5 Master Thesis Structure

**Chapter 2** includes the theoretical background of the research and answers RQ1 and RQ2. **Chapter 3** describes the data collection and soil organic carbon computation. **Chapter 4** discusses the development of the prediction model, which corresponds with RQ3. The validation and optimization of the developed model, needed to answer RQ4, are discussed in **Chapter 5. Chapter 6** discusses the potential value for stakeholders, which is RQ5. The thesis finishes with a conclusion, discussion, limitations and a future work section in **Chapter 7.**

Table 1 shows a mapping of the chapters of this thesis, the related DSRM activities and the research sub-questions.

**Table 1. Mapping DSRM and thesis chapters**

| Chapter | DSRM Activity | Research sub-question |
|---|---|---|
| 1. Introduction | *Problem Definition & Motivation*<br>*Define Objectives of a Solution* | |
| 2. Background | | RQ1, RQ2 |
| 3. Data collection and soil organic carbon computation | *Design and Development* | RQ3 |
| 4. Feature selection and model development | *Design and Development* | RQ3 |
| 5. Model validation and optimization | *Demonstration*<br>*Evaluation* | RQ4 |
| 6. Potential value for stakeholders | | RQ5 |
| 7. Conclusions | | |

# 2. Background

This chapter presents the background information that is relevant to our work. The chapter starts with information about carbon emission and sequestration. Subsequently, information on Digital Soil Mapping is given. The next section gives information about the state of the art regarding environmental covariates, prediction methods and validation strategies used for soil organic carbon mapping. The chapter ends with a section on machine learning.

The state of the art regarding organic carbon mapping was reviewed based on articles found on the repositories Scopus and Web of Science. We have used a systematic literature review methodology following the guidelines from Webster and Watson [10] in order to recognize patterns and concepts in the state of the art. The literature review includes other review papers from before 2019 and new studies performed from 2019 and later.

## 2.1 Carbon emission and sequestration

Reducing greenhouse gas emissions is a key topic within the sustainability discussion around the world. The European Commission has adopted the 2030 agenda for sustainable development from the United Nations, in which is stated that 'Climate action', including mitigation efforts to reduce the greenhouse gas emissions, is one of the 17 Sustainable Development Goals (SDGs) [11]. The Netherlands' goal related to climate action is to reduce the gas emissions by at least 40% before 2030. Carbon dioxide is a main type of greenhouse gases that is emitted and global warming is the most important consequence of the emission of carbon dioxide. The sequestration of carbon in the soil is a mitigation effort to reduce $CO_2$ emission [11], [12]. The European Commission states that land-based solutions should maximize soil carbon sequestration. To maximize the soil carbon storage, active soil quality determination and monitoring is needed.

### 2.1.2 Carbon sequestration

Carbon sequestration consists of capturing and storing of carbon dioxide ($CO_2$) from the earth's atmosphere. The carbon that is sequestrated forms the basis of the organic matter in the soil, which consists of the cells of microorganisms, plants and animal residues at various stages of decomposition [13]. Molecules of organic matter can contain carbon, hydrogen, nitrogen, phosphorus and sulfur and the soil organic carbon content can be increased by increasing the portion of organic matter in the soil. This not only helps in reducing carbon emission, but often also enhances the soils physical, chemical and biological processes and properties [14].

A researcher of the Louis Bolk Instituut discussed different measures that farmers can use to enhance carbon sequestration [15]:

- Leave behind crop residues
- Add extra 'resting crops' such as grains in the rotation
- Use green manures
- Apply non-inversion tillage instead of plowing
- Use additional rough manure or compost
- Keep perennial field margins
- Stop ploughing grassland
- Use herbaceous grasslands

## 2.2 Digital soil mapping

Digital soil mapping is a solution to the lack of measurements in case information on the state or quality of the soil is needed. Minasny and McBratney [16] define digital soil mapping as *the creation and population of spatial soil information systems by the use of field and laboratory observational methods coupled with spatial and non-spatial soil inference systems.* In this context, a soil map is defined as a graphic representation for transmitting information about the spatial distribution of soil attributes. In short, environmental data is used to predict aspects of the soil. According to Minasny and McBratney [16] digital soil mapping has three components:

- Inputs: field observations and laboratory observations
- Process: build a mathematical or statistical model that relates soil observations with their environmental covariates (the so-called SCORPAN factors).
- Output: a soil information system (such as a soil map).

The existing approaches to mapping soil organic carbon do not reach the levels of practical usage. Although the inputs are available as soil samples, there are no appropriate processes. Initiatives such as the Global Soil Organic Carbon map (GSOC map) [4] and SoilGrids [5] provide estimations of the carbon stocks for the entire world. However, these maps are snapshots of the carbon stocks on the entire world and do not provide up-to-date information about the soil state. In order to get a view on the current state of the carbon stocks, the GSOC or SoilGrids maps should be enriched with environmental covariates so that a more detailed and actual status of the soil can be provided to farmers. This research looked for a solution that uses actual soil sample data, SOC maps and up-to-date covariates for an accurate soil organic carbon estimation for farm plots in the Netherlands.

## 2.3 Environmental covariates

In the past, a wide range of potential factors or covariates and several prediction algorithms have been used and tested for the accuracy, suitability and usability of optimizing the digital soil mapping of soil organic carbon. The environmental covariates that can be used to predict the soil organic carbon content can be categorized into logical clusters using the SCORPAN model developed by McBartney et al [17]. According to this model, the soil state at any point in time is a function of seven environmental covariates: Soil properties (S), Climate (C), Organisms (O), Relief (R), Parent material (P), Age (A) and Spatial location (N). This model helps to categorize the covariates used for soil organic carbon content prediction into logical clusters. The different categories used in the SCORPAN typology are described below.

- **Soil properties (S)** – Previously measured properties of the soil at a certain point in time, such as remotely sensed spectral data, existing soil maps and georeferenced point data representing measurements taken in the field or laboratory. Examples of these type of data are bulk density, clay mineralogy, soil depth and the taxonomy class of the soil.
- **Climate (C) –** All climate related covariates such as precipitation, moisture and temperature.
- **Organisms (O) –** Vegetation cover data and land usage data.
- **Relief (R)** – Relief of the soil, including terrain attributes, elevation and topography.
- **Parent material (P) –** Geologic material from which soils form.
- **Age (A) –** Age of the soil.
- **Spatial location (N) –** Location of the soil.

### 2.3.1 Covariate usage

The literature review on the SOC mapping revealed the covariates and corresponding category that are most often used in similar studies. Table 12 in Appendix A shows the resulting concept occurrence table used for this analysis. Figure 2 shows the covariate occurrence of the studies performed after February 2019 that we have examined for this research.

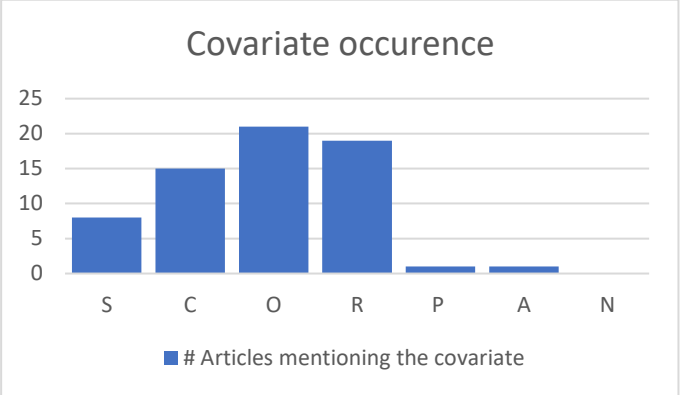


**Figure 2. Covariate occurence**

Covariates categorized as **Organisms** are used the most often in similar studies. The Normalized Difference Vegetation Index (NDVI) score and the cultivation plans used for the farm plot are the highest performing covariates used as features for SOC mapping. The NDVI score is an indication of the amount of vegetation cover. The vegetation cover for a long period of time might be more influential than a single snapshot of the vegetation cover [18], [19].

The **Relief** and **Climate** covariate categories are both used in a big number of studies as well. The Relief category contains covariates such as elevation, slope, curvature, valley depth and landscape position of the area of interest. Several studies have shown that the Digital Elevation Model can be used to predict the soil organic carbon content [20], [21]. The Climate category contains covariates such as amount of precipitation, soil and air temperature and the amount of solar radiation.

Covariates of the **Soil properties** (s) category are used in a few studies. Examples of these covariates are type of soil and soil depth. The data of these covariates is hard to obtain, but can be found in national databases such as the Canadian National Soil Database [22] and the Dutch national soil database (PDOK).

The category **Parent material** contains covariates such as the underlying geological material of the soil. This covariate category is rarely used in soil organic carbon mapping. The same holds for the **Age** and **Spatial location** of the soil.

## 2.4 Machine learning

Machine learning is a branch of Artificial Intelligence (AI) that is based on teaching a machine how to handle data and extract information from the data [23]. It relies on different algorithms that are used to solve data problems. These algorithms use statistics to build mathematical models whose goal is to make inferences on a dataset. There is not a single one-size-fits-all algorithm: the most appropriate algorithm depends on the type of problem the researcher wants to solve and on the data used for the research.

Machine learning algorithms can be split into two types: supervised learning algorithms and unsupervised learning algorithms. Supervised learning involves predetermined output variables and the use of input variables [24] so that the algorithms try to predict the output variables based on the predefined input variables. In contrast, unsupervised learning does not involve a target output variable, but instead the data is labeled during the machine learning process. The accuracy of unsupervised

learning is usually lower than for supervised learning, as the predictions are not validated using the target output variables.

Two different types of machine learning applications can be distinguished: classification and regression. Classification algorithms try to find the decision boundaries for the target value placing it in a category. In contrast, with regression the target value is not a class or category, but a continuous variable.

The literature review we performed on SOC mapping gives us insights in the use of machine learning models and the way these models are validated that are presented in the next subchapters answering RQ2.

### 2.4.1 Prediction methods and SOC mapping

According to the research of Lamichhane [18], Multiple Linear Regression (MLR) is the method that is mostly used in the researches between 2013 and 2019. In general, it is a regression method that estimates the relationship between the dependent variable and two or more independent variables. However, this method is almost always outperformed by other methods in comparative studies. Random Forest is the second most used method to predict the soil organic carbon content in this time frame and outperforms other methods in the majority of the studies. Other popular methods are the Cubist and Regression tree methods.

The literature review performed on the methods used after 2019 confirmed the insights gained in the other review papers. Figure 3 shows the method occurrence found in studies similar to this research performed in 2019 and later. The concept table (Table 13) can be found in Appendix A. The shifting trend from traditional regression methods towards machine learning is confirmed in this research, as almost all studies have used machine learning methods as the prediction method for the SOC content. The **Random Forest** (RF) method is used in the largest amount of researches. This gives a good indication about the relevance of the method. Out of the 17 studies that compare the method to other methods, Random Forest outperformed the other methods in 10 studies [25]–[34].

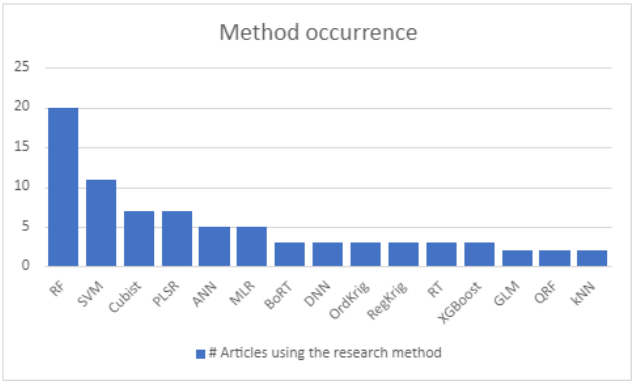


**Figure 3. Prediction method occurrence**

**Support Vector Machine** (SVM) is another popular method. Although it has been used quiet often, it is almost always outperformed by other methods (mostly Random Forest).

The last interesting insight from our literature review is the upcoming trend of **Artificial Neural Networks** (ANN) and **Deep Neural Networks** (DNN). DNN is used in only a few studies, but in these studies it is compared with other methods and it outperformed all other methods [35]–[37].

### *Random Forest*

Random Forest is a tree-based ensemble prediction method that combines the output of multiple decision trees into a single result [24]. Decision trees use a flowchart structure in which tests are performed on attributes or variables. The possible outcomes on the test are branches that lead to new tests and eventually to a resulting value.

When Random Forest is used for regression, the result is determined by averaging the results of the individual trees. When used for classification, the single result is determined based on majority voting: the most frequent categorical variable is chosen as the result. Figure 4 shows an example of three decision trees and the total flow used for random forests.

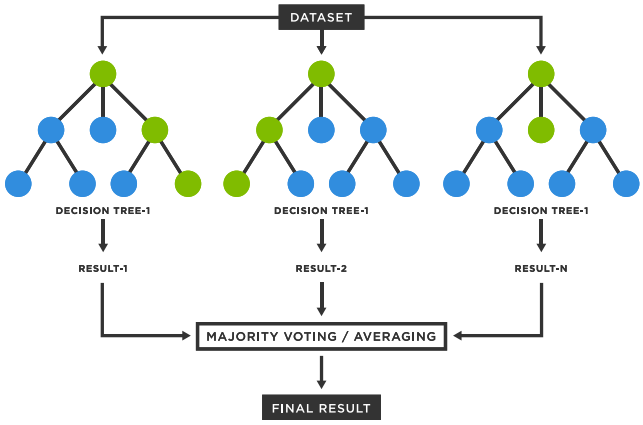


**Figure 4. Random forest flow**

### *Support Vector Machine*

Support Vector Machine works by mapping data points to a multi-dimensional feature space and use lines or hyperplanes to separate classes. It classifies new data points based on the location of the data point in the multi-dimensional space on whether it lies above or below the line or hyperplane [38].

### *Neural Networks*

Neural networks, also known as Artificial Neural Networks, are a subset of machine learning algorithms that uses deep learning. The networks consist of layers of nodes, containing an input layer, one or more hidden layers and an output layer. The nodes are connected to each other and have a weight and a threshold. If the output of an individual node is above the threshold, the node sends data to the next layer of the network [39]. Figure 5 schematically shows a neural network with an input layer, hidden layers and an output layer.

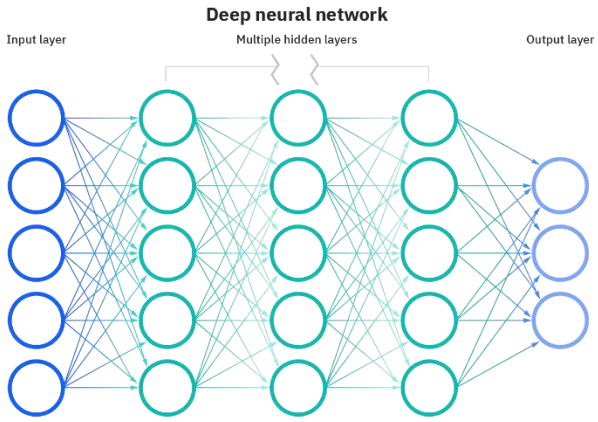


**Figure 5. Neural Network scheme**

The weights of the nodes determine the importance of the variables: larger weights contribute more to the output variable than lower weights. The data is pushed through the layers based on thresholds and eventually the last layer is reached. The output layer processes the outputs of the previous layers and determines the resulting prediction.

Deep learning or deep neural networks are networks with a large amount of hidden layers. It is usually used for a large, potential unstructured dataset.

### 2.4.2 Validation strategies

In our literature review on the SOC mapping we analyzed the validation strategies used in similar studies. The resulting concept table of this literature review can be found in Table 14 in appendix A.

In the early 2000's, a substantial amount of the studies examined in the study of Minasny et al. [17] did not validate the prediction models. This has changed for the studies performed between 2013 and 2019: the studies included in the research of Lamichhane et al [18] have all performed a validation step. The majority of the studies used the data-splitting technique for evaluating the results. Other studies used cross-validation.

Considering studies performed after February 2019 we have found that almost all studies (24 of the 27) have used cross-validation as validation strategy. **Cross-validation** is a widely used data resampling method to prevent overfitting of the machine learning model [40]. Overfitting is the phenomenon that a model is perfectly adapting to the dataset, but is not able to generalize to data that is not in the dataset. When using (k-fold) cross-validation, the dataset is partitioned in subsets (amount = k) of approximately equal sizes without overlap in the subsets [41]. The model is trained for *k-1* subsets and is validated with the remaining subset. This process is iterated until all the subsets are used as validation set.

The majority of the studies using cross-validation in this field have used a 10-fold cross-validation strategy. This means that the data is partitioned in 10 different subsets and all the subsets are used once as testing set. Most of the studies have an $R^2$ value between 0.4 and 0.8.

# 3. Data description and soil organic carbon computation

Figure 6 shows the methodology followed for this research. The literature review on SOC mapping and its conclusions form the theoretical basis that is used for designing and developing the artifact. The methodology followed for the development of the model is Knowledge Discovery from Databases [8]. This chapter covers the first three phases of the methodology: *Selection, Pre-processing and Transformation*.

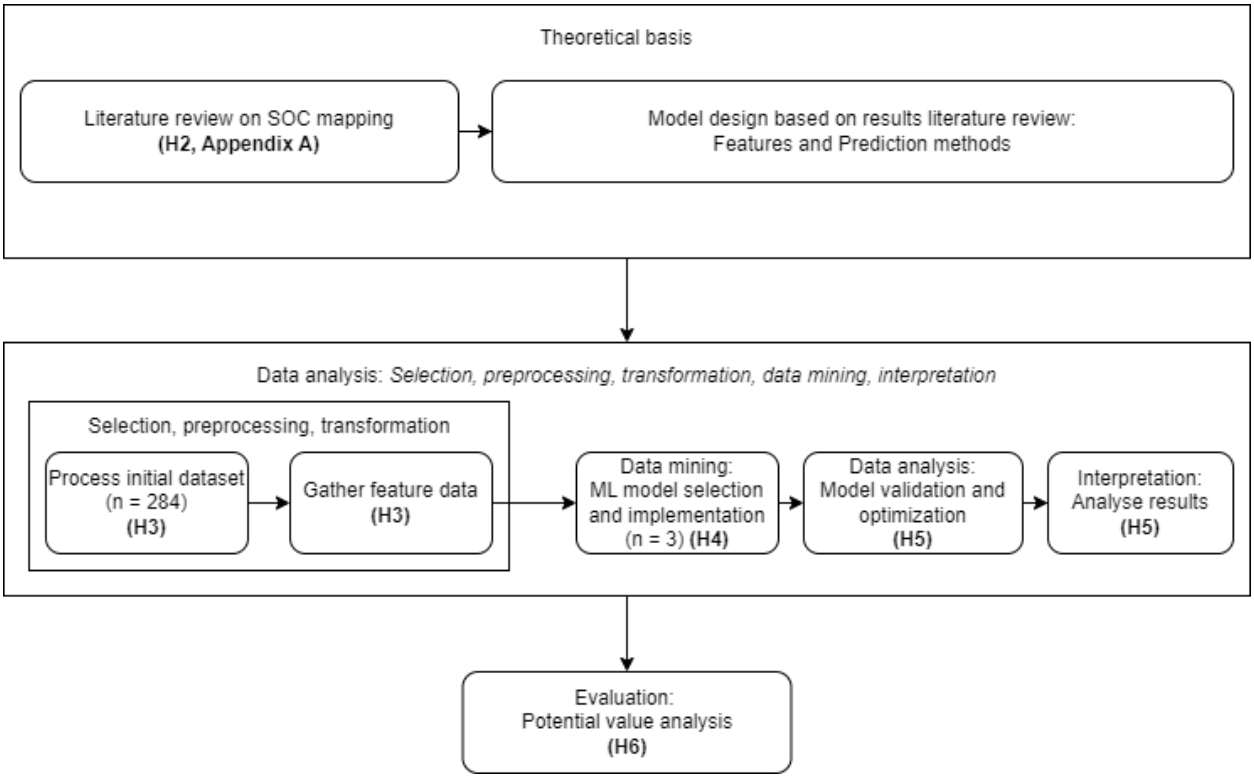


**Figure 6. Methodology used for organic carbon prediction**

## 3.1 Initial dataset

The initial dataset contains observational soil data gathered by taking soil samples. The data is gathered by Eurofins-Agro, a group of laboratories that provide support services to the agriculture industry. This data is gathered by taking multiple soil samples on a farm plot, putting these different samples together on a pile and measuring different aspects of the soil. Although the soil samples are taken (almost) every year on a farm, not every farm plot will be sampled during these visits. On average, a plot farm is sampled once each 4-5 years. The average measurement error on the soil samples is 10-15%.

The dataset contains 284 soil measurements formatted as a comma-separated values file. It consists of measurements taken between 2010 and 2022. Some measurements of the dataset are from the same farm plots, but most of the measurements are from different plots. The measurements taken by Eurofins contain a vast amount of columns (variables), but not all these columns are relevant for this research. The relevant columns for this research are the following:

- *MNNROA:* Unique identifier of the visit/research on which the sample is taken

- *monsternr:* Unique identifier of the soil sample
- *location:* Coordinates corresponding to the border of the farm plot
- *date:* Date on which the measurement is performed
- *os:* Percentage of organic matter in the soil

The Eurofins-Agro measurements have been matched with open farm plot data from the Basisregistratie Gewaspercelen (BRP) dataset of the Nationaal Georegister of the Netherlands in order to identify the farm plots corresponding to the measurement. The location field is replaced by the set of coordinates of the border of the matched BRP farm plot. Thereby, the following columns are added to the dataset:

- *brp_hash:* Unique identifier of the farm plot
- *centroid_distance:* Distance between the centroid of the Eurofins coordinates and the matched BRP farm plot centroid
- *centroid:* Coordinates of the centroid of farm plot

The initial soil measurements do not contain information on the SOC content of the soil. Therefore, different strategies to obtain or compute the SOC content data have been investigated.

### 3.1.1 Soil organic carbon computation

The initial computation strategy used in the international agriculture sector is a rule of thumb that the amount of organic carbon is roughly equal to 50% of the soil organic matter (SOM) content [42]. Research performed by researchers from the University of Wageningen has shown that the SOC:SOM ratio for farms in the Netherlands is in line with this rule of thumb: the datasets investigated show a ratio between 0.38 and 0.60 (0.52 ± 0.08 in two datasets and 0.47 ± 0.09 in a third dataset)[42]. However, we have decided to look further for other computation strategies as a measurement error of 10-15% and a deviation of the carbon content between 0.38 and 0.60 leads to a layer of uncertainty that is too large for this research.

The second SOC computation strategy comes from contact with Eurofins-Agro. Although there are no SOC measurements available in the initial dataset, it turned out that these SOC values can be calculated based on the Near Infrared Spectroscopy (NIRS) values of the samples. Research has shown that the SOC values can be calculated with 'a satisfactory to good calibration performance' ($R^2$ = 0.98, RMSE = 2.98) [43]. The NIRS data of the measurements in the initial dataset are stored by Eurofins-Agro and the SOC values based on this NIRS values are provided to me in two separate datasets. These new datasets contain the SOC and SOM values of measurements. These values are mapped to the initial dataset based on the unique identifier of the sample (*monsternr* in initial dataset).

## 3.2 Prediction covariates

The results on RQ1 are a set of environmental covariates categorized using the SCORPAN typology. I have established a list of environmental covariates that are used as features for the prediction model based on the results of RQ1. The covariates (and corresponding categories) can be found below:

- *Climate*: Annual mean and variance of daily amount of rainfall
- *Climate*: Annual mean and variance of daily temperature
- *Climate*: Annual mean and variance of daily amount of hours solar radiation
- *Soil propierties*: Mostly present type of underlying soil
- *Organisms*: Cultivation plans of the past 5 years

- *Organisms*: Average NDVI score of the past year
- *Other*: SoilGrids organic carbon stock and density estimations

We have not included covariates from the Relief category such as the elevation of the soil as the elevation is less relevant for farms in the Netherlands and close to equal for all farm plots. Figure 7 shows a class diagram that presents an overview of all the data used for this research. The Eurofins measurements are the basis of the target dataset. The five different types of data added to the dataset are presented together with an index (1-5). These indices also correspond to the Jupyter Notebooks that we have created in order to gather and transform the data. Table 2 shows a further explanation of the variables. It includes a description of the variable, the type of data, the amount of missing values and the Jupyter Notebook flow used to gather and further process this data.

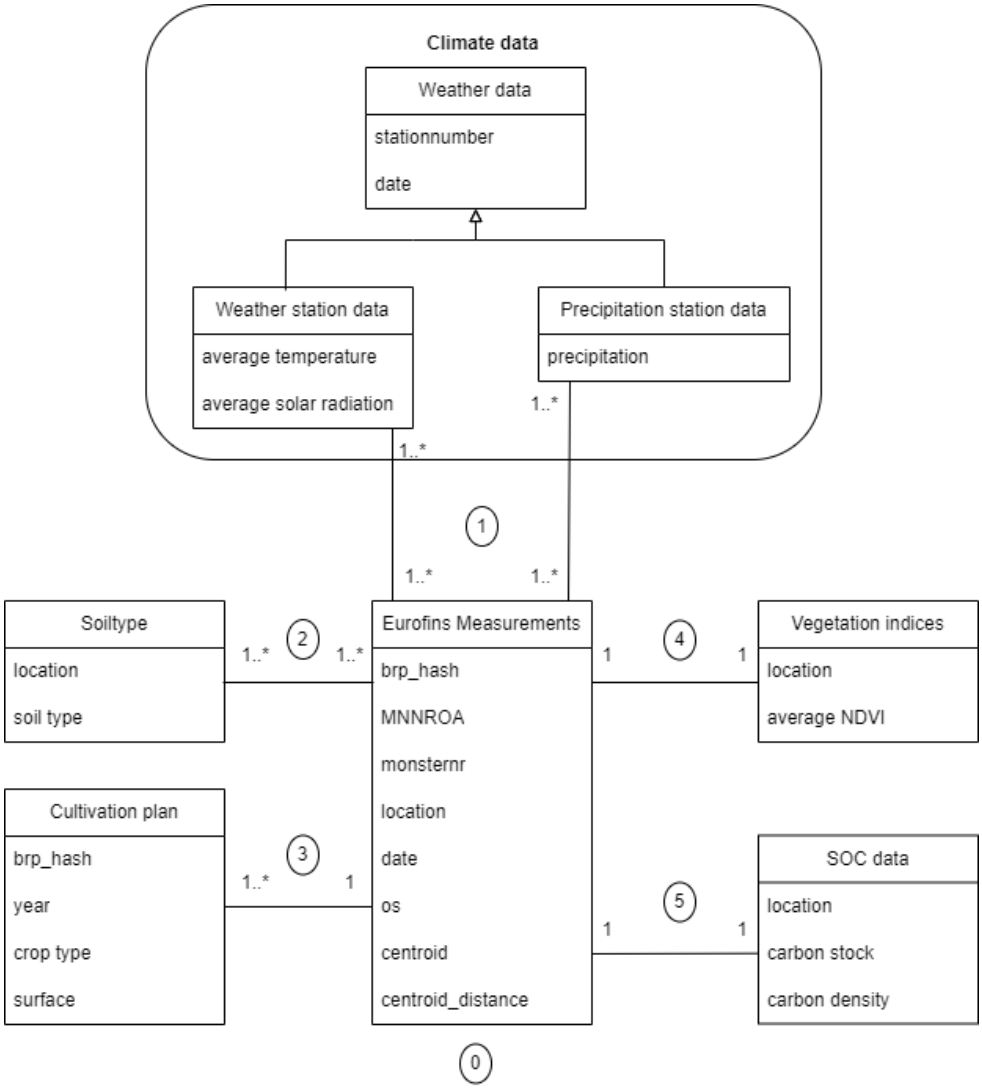


**Figure 7. Class diagram of the data**

**Table 2. Data sources and Jupyter Notebook flows**

| Index | Variable | Unit | Missing | Flow |
|---|---|---|---|---|
| 0 | Carbon content | Percentage | - | [0 Basis] Merge Eurofins Datasets |
| | Location | Polygon | | |
| | Date | Date (DD-MM-YYYY) | | |
| 1 | Station number | Numerical | - | [1 KNMI 1] Create CSV of active precipitation stations<br>[1 KNMI 2] Download Weather Data<br>[1 KNMI 3] Add Weather Data to Merged Dataset |
| | Date | Date (YYYYMMDD) | | |
| | Daily average temperature | Decimal | | |
| | Daily amount of hours of solar radiation | Decimal | | |
| | Amount of daily precipitation | Decimal | | |
| 2 | Soil type | Categorical | - | [2 Soil 1] Add soil texture to Merged Dataset |
| | Location | Polygon | | |
| 3 | Cultivation plan | Categorical | 94 (6.6%) | [3 Cultivation Plan 1] Add Cultivation plans to Merged Dataset |
| | Hash (farm plot identifier) | Hash | | |
| | Surface | Decimal | | |
| | Year | Integer | | |
| 4 | Yearly average NDVI | Decimal | 74 (35.4%) | [4 NDVI 1] Add NDVI Scores to Merged Dataset |
| | Location | Polygon | | |
| 5 | Carbon stock | Decimal (t/ha) | - | [5 SOC 1] Add SOC estimations SoilGrids to Merged Dataset |
| | Carbon density | Decimal (g/dm3) | | |
| | Location | Polygon | | |

### 3.2.1 Climate data

In total, three different types of climate data is gathered and used as input for the SOC prediction: rainfall, temperature and solar radiation. The Koninklijk Nederlands Meteorologisch Instituut (KNMI) has APIs available that can be used to gather daily measures of this data. The rainfall data is gathered via the precipitation stations and the other two types of data are gathered via the weather stations.

The first step in gathering data is to find the active precipitation stations and the active weather stations. We have created a list of precipitation stations that are active within the relevant time frame (between 01-01-2000 and the 01-01-2023) and provide the necessary data. The station list contains the name of the station, the station number and the location of the station.

The next step in gathering climate data is to find the closest station to the centroid of the farm plot. The Jupyter Notebook downloads the weather data that lies within the time frame of the first measurement and last measurement of the dataset for all the unique weather stations and precipitation stations. These data points are daily measurements of the amount of rainfall (in 0.1 millimeters), the average temperature (in 0.1 degrees Celsius) and the amount of solar radiation (0.1 hours).

The last step for gathering the climate data is to transform the downloaded data into relevant measures for all the Eurofins measurements. For each of the covariates, we selected the daily measurements that correspond with the closest weather station and are measured in the past year and calculated the mean and variance for these values. Thereby, the following fields are added to the target dataset:

- PrecipitationAverage
- PrecipitationVariance
- TemperatureAverage
- TemperatureVariance
- RadiationAverage
- RadiationVariance

### 3.2.2 Soil information

The second source of data that is added to the target dataset is the soil information. We have downloaded the BRO Bodemkaart (scale 1:50,000) from the Publieke Dienstverlening Op de Kaart (PDOK). This dataset contains polygons with RD-coordinates (Rijksdriehoekscoördinaten) and a type of soil that corresponds with these coordinates.

We have combined the soil dataset with the target dataset based on the location polygons. The overlap percentage is calculated for each farm plot and the soil type with the highest overlap percentage for the farm plot is selected for the target dataset. The other soil types are dropped from the dataset. Figure 8 shows an example of a farm plot and the underlying soil types. In this example, the blue polygon is the farm plot and the red polygons are the overlapping underlying soil types.

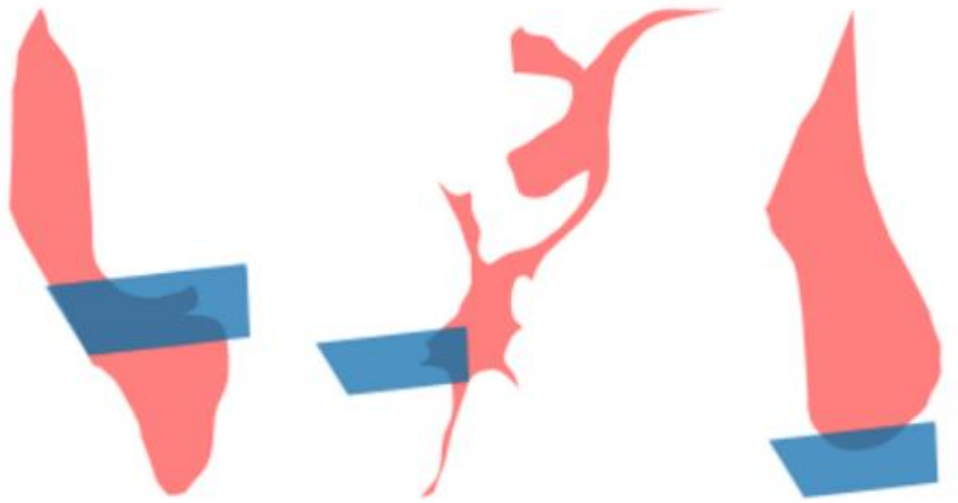

**Figure 8. Example of a farm plot with underlying soil types**

After this addition of the soil type, the target dataset contains one new column that can be used as an input feature: the column *normal_soilprofile_name.* However, these values are not numerical and need to be transformed so that they can be used for regression by the machine learning model. The transformation of these fields will be discussed in section 3.3: *Data pre-processing and transformation*

### 3.2.3 Cultivation plans

The cultivation plans used for this research come from the BRP Gewaspercelen dataset of PDOK**.** The cultivation plans of the last five years are added to the target dataset. However, finding the history of cultivation plans for a farm plot can be challenging as the size of the farm plot can change over time. Sometimes multiple farm plots are merged together into one farm plot or a farm plot is split up into multiple farm plots. The borders of farm plots can change. Therefore, the following logic has been used to match historical farm plots and the corresponding cultivation plans to the current farm plots. In general we have used the rule of at least 75% overlap to match farm plots. Overlapping farm plots can be categorized into different types of matches:

- The plots are matched and identified as identical if the surface difference is a maximum of 20% between the farm plots. The plots are checked if they are not merged or split subsequently.
- The plots are identified as expanded when the surface of the farm plot is at least 20% higher than the surface of the original farm plot. Afterwards, the plots are again checked if they are not merged or split. The same rules apply for reduced plots, but in this case the surface should be at least 20% lower than the original farm plot.
- The plots are identified as split farm plot when the farm plot has at least one other plot of the current year overlapping the old plot and the plot is not merged with another plot.
- The farm plot is categorized as merged and matched to the current year farm plot if the overlap check returns more than one matching farm plot.
- If there is no plot with more than 75% overlap to the plot, the plot is categorized as a new farm plot.

The resulting cultivation plan dataset contains farm plots with the corresponding matched historical plots attached to the farm plot. The cultivation plans of the past five years are added to the target dataset. Whenever there is missing data for specific fields, the fields are left empty in the target dataset. Sometimes, there are more than one historical farm plots and cultivation plans of the same year for a farm plot. Therefore we have used the cultivation plan of the farm plot with the highest surface as input feature for the machine learning model. The cultivation plan data needs to be transformed into numerical values like the soil type data, which will be explained in section 3.3.

### 3.2.4 Vegetation index

The Normalized Difference Vegetation Index (NDVI) is a vegetation quantification measuring the normalized difference between the near infrared and red light bands of satellite images. Vegetation strongly reflects the near infrared (NIR) bands and absorbs the red light (Red) bands. The NDVI values can be calculated using the following formula:

$$NDVI = \frac{(NIR - Red)}{(NIR + Red)}$$

The satellite measures the bands reflected by the area of interest. The NDVI score always ranges from -1 to 1. Negative values usually represent water. A value close to 1 indicates a very green area, usually consisting of a lot of vegetation. When the NDVI score is close to 0, it is likely an urbanized area or an area without vegetation.

The Google Earth Engine has several geospatial datasets available that can be used for research purposes. For this research, we have used the Copernicus dataset of the Sentinel-2 Level-2A satellite. The Sentinel-2A product provides data of Bottom-Of-Atmosphere (BOA) reflectance, which represents the actual reflectance of the areas on the earth's surface. In comparison to data of Sentinel-2 Level-1C, which provides data on Top-Of-Atmosphere level, the BOA values undergo atmospheric correction.

The near infrared band (B8) and the red light band (B4) are both provided with a relatively accurate pixel size (10 meters). The Google Earth Engine also provide datasets that have the NDVI value available without having to calculate it, but these datasets have a substantially higher pixel size (≥ 250 meters). Figure 9 and Figure 10 present examples of a NDVI map of the Netherlands, showing the average NDVI value per pixel. Figure 9 presents the average NDVI score per pixel in april 2020, Figure 10 shows the average NDVI score per pixel in october 2020. In order to create these images I have used the ADM0 boundaries (which refer to country boundaries) as area of interest and used the boundary coordinates as filter when retrieving sattelite data. The difference in color of the two images can be explained by the difference in season for both images: Figure 9 shows the vegetation in the autumn, Figure 10 shows the NDVI scores in the spring season which is usually more green.

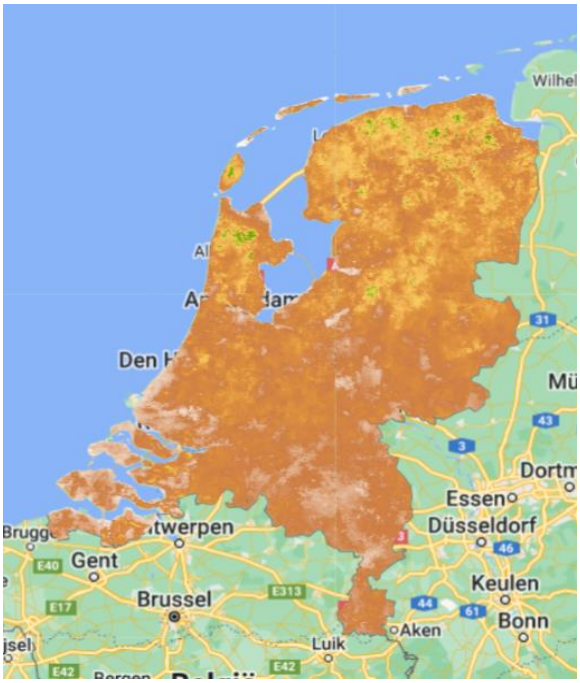


**Figure 9. NDVI image constructed for October 2020**


**Figure 10. NDVI image constructed for April 2020**

The dataset of the Sentinel-2 satellite contains data points measured after march 2017. For all Eurofins samples taken after march 2018, the average NDVI score of the past year for the farm plot is added to the target dataset. For samples taken between march 2017 and march 2018, the average NDVI score from march 2017 until the sampling date is added to the target dataset. For Eurofins samples taken before march 2017, no NDVI score is added to the target dataset.

### 3.2.5 SoilGrids

SoilGrids provides global predictions for standard numeric soil properties such as the organic carbon stock and the organic carbon density. The estimations of Soilgrids may give a good indication on the

organic carbon content on the farm plots. However, the accuracy of SoilGrids layers is still limited and the variation explained by the models is between 30% and 70%. The resolution of the SoilGrids map is 250 meters. The latest release of the SoilGrids map is the version released in 2020.

The SoilGrids map is accessed through the WebCoverageService (WCS). For each Eurofins measurement, a rectangle-shaped box is calculated that fully covers the farm plot. We have retrieved the organic carbon density and organic carbon stock for the box corresponding to the farm plot. The last step is to add the mean values for the density and stock to the target dataset. The organic carbon stock is estimated for a depth of 0-30 cm and the organic carbon density is estimated for a depth of 5-15 cm.

## 3.3 Data pre-processing and transformation

Some outliers in the original dataset have an unrealistic SOC content value. An example of this outlier is a SOC value that is higher than the organic matter percentage. As the organic carbon content is lower (around 50%) than the organic matter percentage, the outliers that have a higher SOC percentage than the organic matter percentage are removed manually from the dataset.

The cultivation plan data and the soil type data need to be transformed in order to be used for regression. The problem with this data is that it is categorical data and regression can only process numerical data. There is no natural ordered relationship between the categories. In order to use the categorical data, the data needs to be transformed into dummy variables [44]. *Dummy variables* can only take the values of 0 and 1 [45]. We have chosen to adopt the One-Hot Encoding method to transform the data. This methodology can be used to remove the categorical variable and transform it into the new binary variables [46].

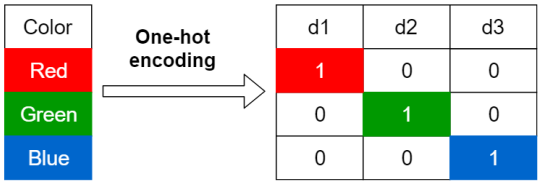


**Figure 11. One-hot encoding example**

All the unique soil types and all unique combinations of the cultivation plan and the amount of years ago are transformed into dummy variables and used as input for the machine learning algorithm. Figure 10 shows an example of how one-hot encoding works. In this example, the three categories for the column 'Color' are red, green and blue. For each category, a new column is created (d1 = red, d2 = green, d3 = blue). Figure 12 shows an example of the soil type and cultivation plan data. Figure 13 shows how the data looks after the transformation of these columns. In this example, the columns (such as Cultplans – year 1 – Maïs, snij-) that have a 0 in it are not shown (as this is a large amount of columns).

| normal_soilprofile_name | year-1 | year-2 | year-3 | year-4 | year-5 |
|---|---|---|---|---|---|
| Veldpodzolgronden; leemarm en zwak lemig fijn zand | Aardappelen, zetmeel | Maïs, snij- | Maïs, snij- | Aardappelen, consumptie | Maïs, snij- |

**Figure 12. Soil type and cultivation plan data**

| Soiltype - Veldpodzolgronden; leemarm en zwak lemig | Cultplans - year 1 - Aardappe | Cultplans - year 2 - Maïs, snij- | Cultplans - year 3 - Maïs, sn | Cultplans - year 4 - Aardappel | C |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | |

**Figure 13. Transformed soil type and cultivation plan data**

For some Eurofins measurements, the NDVI values are missing. The reason for this is that the satellite images from before 2017 are not available in the Sentinel-2 Level-2A dataset. Therefore we had to make a decision on the usage of the NDVI values: either removing the measurements from before 2017 from the dataset, adding NDVI scores with a worse resolution to the dataset or removing the NDVI score from the features list. We have decided to remove the measurements from before 2017 as most of the measurements (± 80%) are still included in the training and testing of the model. The alternative for the NDVI data using other satellite images have a worse resolution (250 meters) and may lead to inaccuracies in the data. We see the last option as the worst option, as the NDVI score was one of the most important features based on the literature review and has a large prediction potential.

Lastly, unnecessary columns are removed from the dataset. These columns already existed in the original dataset and some columns are also added during the process of adding features to the target dataset. The final dataset contained 209 data points.

# 4. Model development

After gathering the data and transforming it into a usable dataset we started analyzing the data. As the goal of the machine learning artifact is to predict a numerical value, the SOC content, we have used regression machine learning algorithms for the prediction. In this chapter, we discuss the machine learning model selection and implementation step depicted in figure 6. We have used the methods found during the literature review as prediction methods for this model.

For the implementation of the machine learning algorithm, we have decided to use the scikit-learn (version 1.0.2) library in Python, which is suitable for performing machine learning regressions in Python. It includes simple and efficient tools for predictive data analysis and also has a vast amount of machine learning algorithms implemented that can be configured to use it optimally for a dataset.

## 4.1 Training and test data

In order to train the machine learning models, we have split the dataset into a training part and a test part for each iteration of model training. This splitting is performed automatically as this is built-in functionality for the cross-validation library we used when training and validating the model. The data set fragments used for training the model are chosen randomly and do not overlap with the test data. We have tested two different split sizes:

- Training the model with 90% of the data and validating the model with 10% of the data. In this scenario, 188 rows are used for training the model and 21 rows are used for testing the model. These split sizes are used when performing *10-fold cross validation*.
- Training the model with 80% of the data and validating the model with 20% of the data. In this scenario, 167 rows of the data are used to train the model and 42 rows are used to validate the model. We have used these split sizes when performing *5-fold cross validation* and when performing the *ShuffleSplit validation*.

The validation strategies are further explained in Chapter 5.

## 4.2 Regressors

Based on the literature review, the Random Forest algorithm is the most promising algorithm and is used the most in similar studies. However, to test the accuracy of this algorithm, we have decided to implement two other machine learning algorithms as well, namely Support Vector Machine and Artificial Neural Network. The machine learning algorithms are tested on different accuracy metrics and are compared to each other based on these measures.

### 4.2.1 Random Forest

When implementing the Random Forest regressor, the part of the data that is labeled as training data is used to construct a set of decision trees. These decision trees all consist of tests that are performed on the data and branches for the potential outcomes on these tests. All branches lead to either new tests and branches or to a result value: the estimation of the target variable. After all decision trees create a prediction for the target value, the average of the prediction is calculated and is taken as the prediction of the random forest.

We used the RandomForestRegressor class imported from the Scikit-learn library, which consists of a forest ensemble method using the DecisionTreeRegressor as sub-estimator implementation. When initializing this regressor, several hyperparameters can be set in order to let the regressor fit the dataset and the problem context. Table 3 shows the hyperparameters that we have used for this regression model. The definitions in Table 3 come from the Scikit-learn documentation. We set most parameters to their default value. The random_state parameters is set to 0 (could take any random integer). This is a fixed integer which will produce the same results across different calls, generating reproducible results while bootstrapping the data. We have kept the default number of estimators (100), which means that the Random Forest consists of 100 decision trees. Increasing the number of trees leads to a smoother model, but also increases the processing time. The criterion we used for measuring the quality of the split is the squared error. Due to the small amount of data points in the dataset, the minimum number of samples required to split an internal node is set to 2 (lowest possible) and the minimum number of samples required to be at a leaf node is 1. These numbers can be increased when a larger dataset is used.

Figure 14 presents a small snapshot of one of the decision trees of a random forest constructed with our data. This figure shows a clear example on the tests (top row of the nodes), the impact on the squared error, the amount of samples per branch and the impact on the prediction value. The snapshot is taken from a tree consisting of 229 nodes. Figure 17 in Appendix B shows the entire tree.

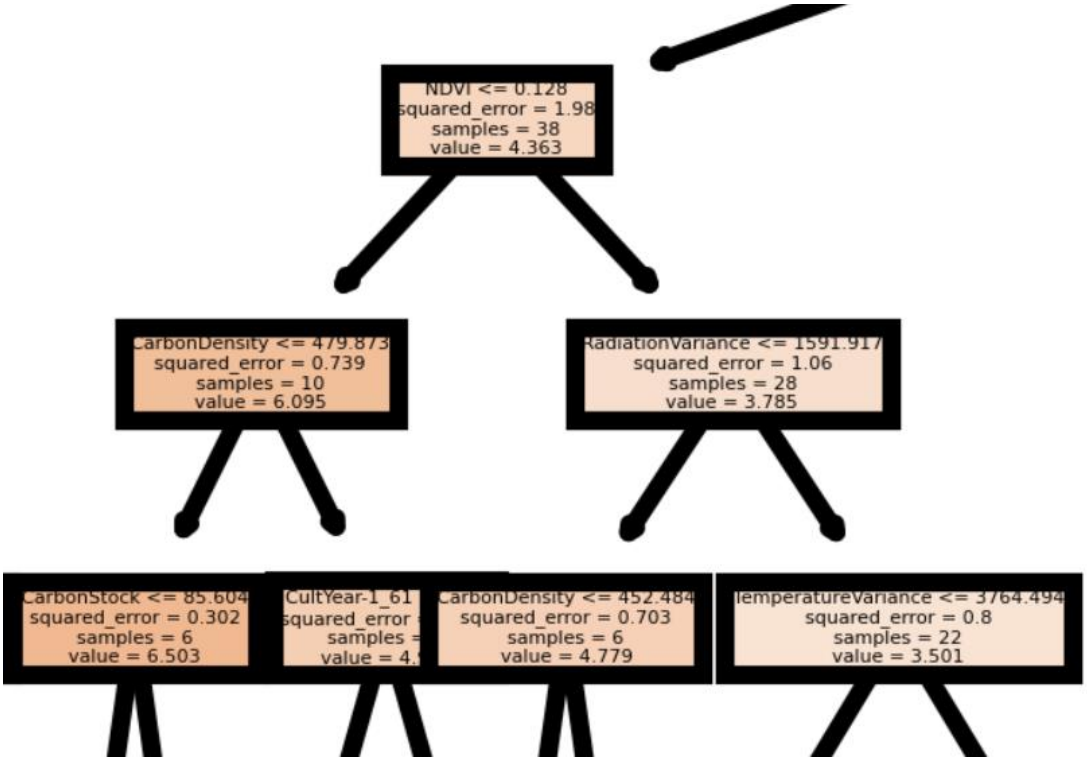


**Figure 14. Constructed decision tree**

**Table 3. Random Forest Regressor hyperparameters**

| Parameter | Meaning | Implementation |
|---|---|---|
| n_estimators | Number of trees in the forest | 100 (default) |
| criterion | Function used to measure the quality of the split | Squared_error (default) |
| max_depth | Maximum depth of the tree | None (default) |
| min_samples_split | Minimum number of samples required to split an internal node | 2 (default) |
| min_samples_leaf | The minimum number of samples required on each side of a split node. | 1 (default) |
| min_weight_fraction_leaf | The minimum weighted fraction of the sum of weights required to be at the leaf node. | 0.0 (default, samples have an equal weight) |
| max_features | The number of features to consider when looking for the best split | 1.0 (default) |
| max_leaf_nodes | Grows trees with nodes 'best-first', nodes with the least impurity. | None (default, unlimited amount of leaf nodes) |
| min_impurity_decrease | Only split a node if the impurity decreases with at least this value | 0.0 (default) |
| bootstrap | Use bootstrap samples are when building trees | True (default) |
| oob_score | Use out-of-bag samples to estimate the generalization score | False (default) |
| n_jobs | Number of jobs that run in parallel | None (default, 1) |
| random_state | Controls the randomness of the bootstrapping | 0 |
| verbose | Controls verbosity (logging) | 0 (default) |
| warm_start | Re-use the solution of previous call (true) or fit a whole new forest (false) | False (default) |
| ccp_alpha | Minimal Cost-Complexity Pruning | 0.0 (default) |
| max_samples | Maximum number of samples drawn to train an estimator | None (default) |

### 4.2.2 Support Vector Machine

In our implementation of the Support Vector Machine (SVM) regressor, the training part of the data is used to determine a multi-dimensional feature space and create a multi-dimensional hyperplane that is used to classify the data points in the test set. This hyperplane is determined in such way that the data points are as close as possible to the hyperplane. Later, the hyperplane function is used to calculate the target value for the test part of the data.

We have implemented the Support Vector Machine algorithm using the SVR class in Scikit-learn, which provides an Epsilon-Support Vector Regressor. Table 4 shows the hyperparameters that can be filled in when initializing the regressor and the values we chose for these hyperparameters. One interesting parameter is the degree of the polynomial function. As we are not working with a large dataset, we have chosen to increase the polynomial degree to 5 (instead of the default value, 3). Increasing the polynomial degree makes the decision boundary more flexible: the constructed hyperplane fits the data

better [47]. However, a risk of increasing this polynomial degree is that the model overfits the dataset. This means that the model can predict the values relatively well, but the model is not generalizable to another (larger) dataset. Therefore we have only slightly increased the degree.

**Table 4. Support Vector Machine Regressor hyperparameters**

| Parameter | Meaning | Implementation |
|---|---|---|
| kernel | Specifies the kernel to be used | None (default, RBF kernel) |
| degree | Degree of polynomial kernel function (represents the similarity of vectors) | 5 |
| gamma | Kernel coefficient | 'scale' (default) |
| coef0 | *Only relevant when kernel = poly or sigmoid* | - |
| tol | Tolerance for stopping criterion | $1e^{-3}$ (default) |
| C | Regularization parameter (inversed) | 1.0 (default) |
| epsilon | Epsilon in the epsilon-SVR model | 0.1 (default) |
| shrinking | Use the shrinking heuristic | True (default) |
| cache_size | Kernel cache (MB) | 200 (default) |
| verbose | Enable verbose (logging) | False (default) |
| max_iter | Limit on iterations | -1 (default) |

### 4.2.3 Artificial Neural Network

In the implementation of the Artificial Neural Network regressor, the training part of the data is used to construct an input layer, multiple hidden layers and an output layer. Figure 15 presents the structure of such a network with 2 hidden layers [48]. During the development of the model, the data is iteratively put through the network in which every feature gets a certain weight (forward propagation), the loss (or accuracy) of the resulting prediction in the output layer is calculated and the weights of the features in the hidden layer nodes are changed based on this loss (backwards propagation). This is done until either conversion to a loss threshold is achieved or the maximum amount of iterations is reached.

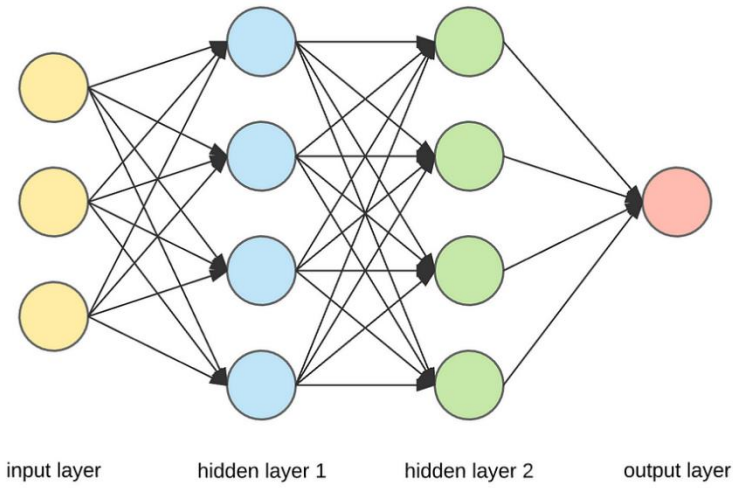


**Figure 15. Artificial Neural Network**

The Artificial Neural Network is implemented using the MLPRegressor class of Scikit-learn, which implements a Multi-layer Perceptron regressor. Table 5 shows the hyperparameters chosen when initializing the regressor. We have chosen to stick with the default values of the regressor for almost all the hyperparameters, as neural networks are usually used for larger datasets and might not fit this data set [49]. Changing the input parameters can lead to model overfitting. However, we advise to critically look at the input parameters when using a different or lager dataset.

**Table 5. Neural Network Regressor hyperparameters**

| **Parameter** | **Meaning** | **Implementation** |
|---|---|---|
| hidden_layer_sizes | Number of neurons in the layer | 100 (default) |
| activation | Activation function for the layer | 'relu' (default, rectified linear unit function) |
| solver | Weight optimization solver | 'adam' (default) |
| alpha | Strength of the L2 regularization term | 0.0001 (default) |
| batch_size | Size of minibatches for stochastic optimizers | 'auto' (default) |
| learning_rate | Learning rate schedule for weight updates | 'constant' (default, constant learning rate given by learning_rate_init) |
| learning_rate_init | Step-size of updating the weights | 0.001 (default) |
| power_t | *Only relevant when solver = sgd* | - |
| max_iter | Maximum number of iterations | 200 (default) |
| shuffle | Shuffle samples in each iteration | True (default) |
| random_state | Controls the randomness of the weights and bias initialization | 0 |
| tol | Optimization tolerance (defines when convergence is reached) | $1\ e^{-4}$ (default) |
| verbose | Enable verbose (logging) | False (default) |
| warm_start | Re-use the solution of previous call (true) or fit a whole new forest (false) | False (default) |
| momentum | *Only relevant when solver = sgd* | - |
| nesterovs_momentum | *Only relevant when solver = sgd* | - |
| early_stopping | Use early stopping when validation score is not improving | False |
| validation_fraction | *Only relevant when early stopping is true* | - |
| beta_1 | Exponential decay rate – first moment | 0.9 (default) |
| beta_2 | Exponential decay rate = second moment | 0.999 (default) |
| epsilon | Numerical stability value | $1\ e^{-8}$ (default) |
| n_iter_no_change | Max number of epochs without meeting tolerance | 10 (default) |
| max_fun | *Only relevant when solver = lbfgs* | - |

# 5. Model validation and optimization

The next step of this research is validating the model, interpreting the results and optimizing the model. The method we used for validating the results is cross-validation. Based on the results of the literature review, we have validated the machine learning models with 10-fold cross validation using a test fraction of 0.1. We compared the results of this validation strategy with two other strategies. As a fraction 0.1 of the data is only 21 data points, we think that it is likely that the results will be highly dependent on the chosen test split. Therefore we used 5-fold cross fold cross validation with an increased test split (test split = 0.2) and we also compared the results with a shuffle split validation strategy (amount of splits = 100, test split = 0.2). We have used the cross-validate class of Scikit-learn for each validation and the ShuffleSplit class of Scikit-learn for splitting the data for the shuffle split validation.

In 10-fold cross validation works is that the data is split into ten equal parts. Each part is used once for testing the model, while the other 9 parts are used for training the model. Thereby, 10% of the data is used to test the performance of the model. The same pattern is used for 5-fold cross validation, but with less splits, less iterations and larger split sizes.

Shuffle split is a method that randomly splits the data into a test and training set with potential overlap of the test sets between the iterations. In each iteration, the data is trained with a fraction (80% of the data in this study) and tested with the remaining parts of the data. The most important difference with cross validation is that more iterations can be performed as the test sets can overlap (in contrast to non-overlapping test sets for iterations of cross validation).

## 5.1 Performance metrics

The first performance metric we used is the $R^2$ value. This value indicates how much variation of the dependent variable can be explained by the independent input variables. It is calculated by dividing the *sum of squares regression* by the *total sum of squares* and subtracting this value from 1 using the following formula [50]:

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

The sum of squares regression is calculated by taking the distance to the regression line to each data point, squaring this distance and summing these values. The sum of squares total is the difference between each data point and the mean of the data points, squared and summed. In general, the higher the $R^2$ is, the better the dependent variable can be explained by the input variables. The value usually ranges from 0 to 1, but can be negative when the chosen model does not follow the trend of the data.

The second performance metric we used is the Mean Squared Error (MSE), which indicates how close the expected values (based on the regression) are to the actual values. It varies from 0 to infinity and the closer the MSE value is to 0, the better the prediction is. It is calculated using the following formula [50]:

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

## 5.2 Results

We have performed 3 types of experiments: training the model with all the features, training the model with all the features except the SoilGrids features and training the model with only the SoilGrids features.

### 5.2.1 10-fold cross validation

Table 6 shows the predictive accuracy of the machine learning models when all the features are used for training the machine learning model. We have performed 5 iterations of the 10-fold cross validation. Table 6 shows the metrics of the iterations and the average between these iterations. All iterations of this experiment show similar results on the $R^2$ and MSE metrics for Random Forest. The five iterations for Support Vector Machine resulted in the exact same results every time. The results of the Artificial Neural Network fluctuate a lot and the optimization does not converge before the maximum of iterations are reached (200) for all iterations. Random Forest shows the highest prediction algorithm ($R^2$ = 0.37, MSE = 1.76). The other algorithms perform poorly and even show negative values for the $R^2$ value. This means that the predictions are worse than a constant function that always predict the mean of the data.

**Table 6. Machine learning performance using all features**

| Iteration | $R^2$ (RF) | MSE (RF) | $R^2$ (SVM) | MSE (SVM) | $R^2$ (ANN) | MSE (ANN) |
|---|---|---|---|---|---|---|
| 1 | 0.38 | 1.74 | -0.06 | 2.80 | -112.98 | 454.91 |
| 2 | 0.37 | 1.75 | | | -729.90 | 634.66 |
| 3 | 0.36 | 1.81 | | | -30.85 | 121.55 |
| 4 | 0.39 | 1.72 | | | -670.36 | 848.78 |
| 5 | 0.35 | 1.77 | | | -408.79 | 904.64 |
| **Average** | **0.37** | **1.76** | **-0.06** | **2.80** | **-408.79** | **592.91** |

Table 7 shows the performance of three iterations of the experiments when not using the SoilGrids features as machine learning features. For this experiment, Random Forest shows the highest predictive accuracy ($R^2$ = 0.47, MSE = 1.53). The other prediction algorithms perform poorly and do not display any predictive value for these features and this dataset. The resulting accuracy in these experiments is higher than the results of the first experiment (when both the SoilGrids data and environmental data are used as features). This raises questions about the value of the SoilGrids estimations, which should be tested and evaluated using a larger data set.

**Table 7. Machine learning performance without using SoilGrids features**

| Iteration | $R^2$ (RF) | MSE (RF) | $R^2$ (SVM) | MSE (SVM) | $R^2$ (ANN) | MSE (ANN) |
|---|---|---|---|---|---|---|
| 1 | 0.47 | 1.51 | -0.10 | 2.89 | -18.50 | 34.15 |
| 2 | 0.46 | 1.54 | | | -67.40 | 168.71 |
| 3 | 0.47 | 1.53 | | | -29.39 | 37.58 |
| **Average** | **0.47** | **1.53** | **-0.10** | **2.89** | **-38.43** | **80.15** |

Table 8 shows the experiment in which we only use the SoilGrids estimations as input features for our prediction algorithm. The predictive accuracy of the machine learning models in this experiment is the lowest of all experiments. In contrast to the previous experiments, Random Forest does not perform the

best with these input features, and only SVM shows some predictive value ($R^2$ = 0.19, MSE = 2.19). However, this still does not display a lot of predictive potential.

**Table 8. Machine learning performance when only using SoilGrids features**

| Iteration | $R^2$ (RF) | MSE (RF) | $R^2$ (SVM) | MSE (SVM) | $R^2$ (ANN) | MSE (ANN) |
|---|---|---|---|---|---|---|
| 1 | 0.01 | 2.31 | 0.19 | 2.19 | -49.97 | 96.98 |
| 2 | -0.01 | 2.40 | | | -69.97 | 98.50 |
| 3 | 0.00 | 2.36 | | | -64.68 | 139.42 |
| **Average** | 0.00 | 2.36 | **0.19** | **2.19** | **-61.54** | 111.63 |

### 5.2.2 5-fold cross validation and shuffle split validation

We have done the same experiments and the same amount of iterations per experiment for the other two validation strategies (5-fold cross validation and shuffle split validation): 5 iterations for experiment 1, 3 iterations for experiment 2 and 3. The results of these validations can be found in Table 9.

**Table 9. Machine learning performance when using alternative validation strategies**

| Experiment | Validation | $R^2$ (RF) | MSE (RF) | $R^2$ (SVM) | MSE (SVM) | $R^2$ (ANN) | MSE (ANN) |
|---|---|---|---|---|---|---|---|
| 1 | 5-fold | 0.36 | 1.74 | -0.07 | 2.87 | -109.81 | 270.96 |
| | Shuffle | 0.52 | 1.36 | -0.01 | 2.78 | -2.17 | 8.41 |
| 2 | 5-fold | 0.40 | 1.61 | -0.09 | 2.93 | -5.31 | 12.09 |
| | Shuffle | 0.48 | 1.48 | -0.03 | 2.89 | -86.97 | 226.53 |
| 3 | 5-fold | 0.09 | 2.44 | 0.20 | 2.22 | -48.89 | 122.33 |
| | Shuffle | 0.39 | 1.68 | 0.21 | 2.24 | -44.92 | 117.21 |

The 5-fold cross validation strategy showed similar results to the 10-fold cross validation strategy. When comparing the different experiments, we can conclude that the added environmental covariates do explain the dependent variable as Experiment 2 (only using added environmental covariates) has the highest predictive accuracy. Experiment 3 (only using SoilGrids) shows the lowest predictive accuracy.

The highest predictive accuracy is used when using shuffle split when cross-validating. The predictive accuracy is the highest when using all input features for the machine learning model. This resulted in a $R^2$ value of 0.52 and a Mean Squared Error of 1.36.

### 5.2.2 Feature importance

The different features used for predicting the soil organic carbon content all have a specific importance. We have looked at the importance of the features when using the highest performing algorithm (Random Forest) with all the input features. Figure 16 shows the importance of the features used in this experiment when using 10-fold cross-validation. The importance of the features is calculated based on the mean decrease in impurity within each tree. The other validation strategies show similar feature importance results.

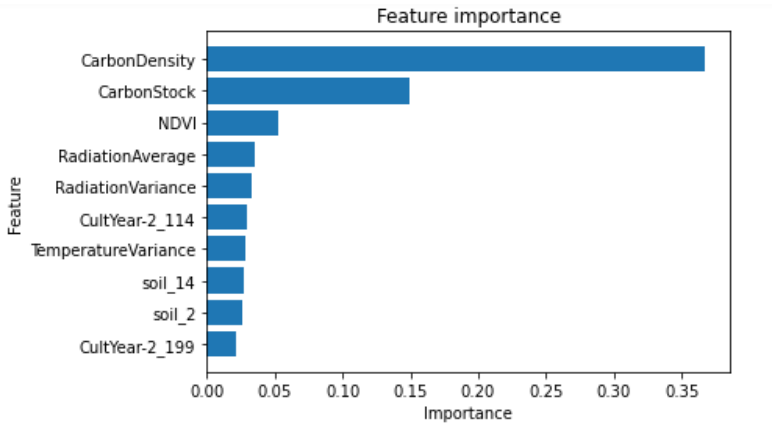


**Figure 16. Feature importance based on mean decrease in impurity**

As expected, the estimations of the carbon density and carbon stock from SoilGrids have the highest impact on the dependent variable. In the different iterations performed for the first experiment, we found that the vegetation cover index (NDVI) and the solar radiation data (average and variance of radiation) are important features. Figure 13 also shows four dummy variable categories as important features. These features show up in the top 10 of important features in multiple iterations. Table 10 shows the meaning of these features. However, it is likely that these features are important for this specific dataset and less important for other datasets due to the low amount of measurements in this dataset. Increasing the amount of data points in the set may lead to totally different values for the relevance of the soil type and cultivation plan usage.

**Table 10. Meaning of important dummy variables**

| **Feature** | **Meaning *(and value in database)*** |
|---|---|
| CultYear-2_114 | Cultivation plan used 2 years ago<br>*(value in database: bieten, suiker-)* |
| Soil_14 | Underlying soil type<br>*(value in database: Moerige eerdgronden met een veenkoloniaal dek en een moerige tussenlaag op zand)* |
| Soil_2 | Underlying soil type<br>*(value in database: Madeveengronden op zand zonder humuspodzol, beginnend ondieper dan 1.2 m)* |
| CultYear-2-199 | Cultivation plan used 2 years ago<br>*(value in database: Maïs, snij-)* |

As the goal of this research was to predict the soil organic carbon content using environmental covariates and existing soil maps, it is interesting to look at the performance of the different experiments. We can clearly see that the environmental covariates show predictive potential, as the results of experiment 1 and 2 are structurally better than experiment 3. This means that the usage of the environmental covariates led to an improvement on the soil organic carbon prediction for this

dataset. We expect that the same holds when using a different, larger, dataset, but this needs to be validated in future research.

### 5.2.3 Reflection on results

The accuracy of the models is not as high as hoped. Whereas Random Forest shows some prediction potential, the other machine learning algorithms do not give meaningful predictions in almost all experiments. However, if we look at the results of the Random Forest algorithm, we can conclude a few things. First of all, the Random Forest outperforms other methods, which is in line with the conclusions we drew from the literature review. Furthermore, we can conclude that the features added to the SoilGrids map improve the carbon content prediction to some extent as the experiment with only using the soil map always shows a substantially lower predictive accuracy.

However, we have worked with a relatively small dataset (209 data points), which increases the risk of overfitting in the prediction model. It is likely that the model tends to predict the carbon content relatively well for the data points in the data set, but is not generalizable to other plots or a larger dataset. In that case the model is not able to predict the SOC content for farm plots in general, but is able to predict the SOC content for the specific farm plots in this dataset.

We expect that increasing the amount of data points will increase the performance of the machine learning models and prevents the models for overfitting better. This expectation is based on the predictive potential the model has shown in this study and based on the similar studies that have shown working predictive models [22], [29], [30], [34], [51].

## 5.3 Model optimization

Machine learning is the process of iteratively improving the machine learning model and its accuracy by tweaking the configurations (also called hyperparameters) of the models. In machine learning, the hyperparameters are set by the researcher and not built by the model. For example, we have chosen to optimize the Random Forest algorithm using the shuffle split validation strategy, as this combination of model and validation strategy show the highest predictive accuracy.

Optimization of the machine learning brings the risk of overfitting the model. By tweaking the hyperparameters we try to reduce the mean squared error of the predictions for data points in this data set. Thereby the generalizability, if existing, can be compromised. During the optimization iterations we have tried to tweak the following hyperparameters:

- *Max depth*: we have tried a maximum depth of the trees of 3, 5, 7, 10 and 20 layers. None of these values of the hyperparameter led to a higher predictive accuracy. The predictive accuracy when using 10 or 20 layers is close to having no maximum depth.
- *Min_samples_split*: Increasing this value varying from 4, 6, 8, 10, 20 did not improve the predictive accuracy. The predictive accuracy was similar to using default settings, as the $R^2$ value was slightly lower for these settings (varying from 0.48 to 0.51) and the mean squared error was similar (varying from 1.34 to 1.39). We also tried to combine the increase of the min_samples_split with an increase of the min_samples_leaf. However, this did not lead to a higher predictive accuracy.
- *Criterion*: we have tried all the different options that can be used to analyze the quality of the splits. Using “absolute_error” made the prediction of organic carbon significantly slower, but did not increase the predictive accuracy. “Friedman_mse” did not lead to better results either, but

was not slower than using the default setting ("squared_error"). The same holds for the "poisson" method of analyzing the quality of the splits.

- *Max_features*: The default value of maximum fraction of features considered when looking for the best split is 1.0. This means that all features are considered. Lowering this fraction led to slightly better results. When we set the fraction to 0.8, the $R^2$ value was 0.55 and the MSE was 1.29. Increasing it to a lower fraction did not improve the results.
- *Min_impurity_decrease*: Increasing the minimum decrease of impurity for a split from 0.0 to 0.1 and 0.2 led to worse results on both the $R^2$ value and MSE.
- *Max_leaf_nodes*: We have tried different values for the maximum amount of leaf nodes. When using values lower than 20, the accuracy of the model decreased. Higher values did not improve the accuracy.

We can conclude that the optimization step does not significantly increase the predictive accuracy of the model. The only change in hyperparameters that improved the model was the reduction of the maximum fraction of features. We do not know if this led to extra overfitting of the model.

# 6. Potential value for the stakeholders

This chapter discusses the potential value of our artifact for the stakeholders. We assume that in future work the predictive accuracy for the SOC content is sufficient and the predictions are visualized into an application that can be used for practice. We start by analyzing the stakeholders for this project and subsequently we discuss the value created by the new solution compared to the solution in the old situation.

## 6.1 Stakeholders

We have identified the stakeholders in this problem context and classified them using the stakeholder taxonomy of Alexander [52]. Table 11 shows the stakeholders and their taxonomy, description and goal.

We have chosen to include the society as a stakeholder for this project. Although this project and the artifact developed do not directly influence society, we see a potential influence in the future once the artifact reaches the level of practical usage and is used to reduce the carbon emission on farm plots.

**Table 11. Stakeholder**

| Stakeholders | Taxonomy | Description | Goal |
|---|---|---|---|
| Supervisors of the University of Twente | Supplier | Supervisors of the University of Twente supply knowledge to the researcher and provide guidance and support to the researcher during the project. | Contribute to the research by supporting the researcher. |
| Master thesis researcher | Developer | The researcher designed and developed the prediction method. | Develop the prediction method. |
| Inversable B.V. | Product champion | Inversible B.V. initiated the development of the artifact based on contact with the other stakeholders. It also supported the master thesis researcher during the development of the project. | Provide farmers with meaningful insights on the state of the soil |
| IntoAgri B.V. | Supplier & product champion | IntoAgri B.V. provided the researcher with agricultural knowledge, explained the relevance of the problem and initiated the project. | Provide farmers with meaningful insights on the state of the soil and give tailored advice to farmers on how to act based on these insights |
| Farmers | Functional beneficiary | The farmers benefit by the insights created by the artifact. | Keep the soil healthy and reduce carbon emission. |
| Society | Functional beneficiary | The society potentially benefits from the carbon emission reduction measures that farmers take. | Live on a healthier planet. |

## 6.2 Potential value

We analyze the potential value of the artifact in four steps: describing the *current situation*, analyzing the *flaws of the current situation*, discussing the *direct impact* of the new situation and *long term*

*implications* of the new situation. Thereby we have taken the stakeholders from Table 11 into account, specifically Inversable B.V., IntoAgri B.V., farmers and the society.

### 6.2.1 Current situation

In the current situation, IntoAgri B.V. provides information about the soil state based on measurements taken averagely once each five years. The advices given by IntoAgri B.V. are mainly based on measurements and not based on estimations of the soil state. The data provided by farmers is not fully utilized.

Inversable B.V. focusses on providing the applications, data management and data analysis that IntoAgri B.V. uses to build tailored advice for farmers. The applications developed by Inversable B.V. contain information about SOC measurements.

If farmers want to have insights into the state of their soil regarding the SOC content, and they have a few options. The first option is to increase the frequency of sampling done by companies like Eurofins-Agro. Thereby the farmers receive information based on laboratory tests, IntoAgri B.V. interprets these measurements and provides advice based on them. The second option is to use snapshots of the soil organic carbon state such as SoilGrids. However, we learnt from the interview with the farmers' organization that the maps of SoilGrids or alternatives are currently not used in practice.

The interview with the farmers' organization learnt us that most farmers do not have clear insights into the state of their soil, specifically regarding the soil organic carbon content. They feel the pressure by the government that encourages them to reduce the carbon emission and increase the carbon sequestration, but they are likely to choose the practices with the highest return or practices based on guidelines of farmers' organization without considering the current SOC state of the soil.

### 6.2.2 Flaws of the current situation

For IntoAgri B.V., the most important flaw of the current situation is that they are not always able to provide up-to-date information and estimations of the current state of the SOC content of the soil on the farm plots. Inversable B.V. is limited to using measurements in the application and can not provide IntoAgri B.V. with information on all the farm plots, since for most farm plots no soil sample data is available.

When Farmers choose to increase the soil sampling frequency, the biggest flaw is the financial impact since it is expensive to increase the frequency of soil sampling. Whenever the farmers choose not to increase the frequency of soil sampling, they will have less insights into the current state of the soil and the organic carbon content of the soil. Generally, the less insights into the state of the soil the farmers have, the less tailored advice the farmers can get on how to improve the soil state and the less the farmers are able to increase the carbon stock and reduce carbon emission.

### 6.2.3 Direct impact of the artifact

The direct impact is the highest for the farmers' organization (IntoAgri B.V.) and the farmers. When the artifact reaches the levels of practical usage, the farmers' organization is able to provide estimations about the current state of the soil. This has two direct implications. The first implication is that they can give a more detailed and tailored advice on how farmers can improve the soil state and increase the soil organic carbon computations. The farmers' organization can give the farmers a data-driven advice on how to do this. The second implication is of financial nature: by being able to provide more up-to-date

information on the soil state, the organization can create more value for their customers and can financially benefit from this.

The direct impact on farmers consists of having a more up-to-date view on the current state of the soil regarding the organic carbon content and being provided with a data-driven plan tailored to their farm plots on how to improve the soil state and increase SOC content.

### 6.2.4 Long term implications

When the artifact reaches the level of practical usage, Inversable B.V. and IntoAgri B.V. are able to create more value for their customers. Thereby new opportunities arise as they can use the same method to provide more insights on the soil quality, soil usage and best practices for farmers. Furthermore, by increasing the value given to customers, it is likely that the amount of customers will increase as well. It also has implications for soil measurement companies, as their role in providing soil information is at stake since alternatives for soil sampling emerge. However, during our research we experimented support and positive reactions from the soil sampling company we contacted (Eurofins-Agro).

For farmers, the long term implications are that a more healthy soil can lead to better performance [14]. Furthermore, initiatives arise that make use of carbon credits, which are certificates earned by companies or projects that they obtain when reducing their carbon dioxide emission [53]. In general, the amount of carbon credits obtained is equal to the amount of tons of sequestrated carbon, so carbon credits can be sold or bought by other companies or projects to compensate their carbon emission. During interviews with the farmers' organization, we learned that municipalities are also investigating the usage of carbon credits and setting a minimum of carbon credits a company should own (either earn or buy).

In the ideal scenario, the society indirectly benefits from this solution as well. If farmers are able to increase the carbon sequestration and reduce the carbon dioxide emission, this has impact on the climate change and reduces the global warming.

# 7. Conclusions

This chapter first presents the general conclusions of the thesis, the methodology used to achieve this goal and the insights gained while performing the research. It also discusses the research questions and answers to the research questions. Subsequently the chapter presents the contribution of this research to research and to practice. Lastly we discuss an overview of the limitations and recommendations for future work.

## 7.1 General conclusions

This thesis presents a method to estimate the soil organic carbon content on farms in the Netherlands. It makes use of a snapshot of existing soil maps (SoilGrids) and enriches this map with up-to-date environmental covariates. This research is performed at Inversable B.V., which provided us with soil samples of farm plots taken by Eurofins-Agro. The methodology used for this research is Design Science Research Methodology [7]. For the third step of the methodology (*Design & Development*) we have used the method Knowledge Discovery in Databases [8]. In order to develop a method that predicts the organic carbon content in the soil, we answered 5 different research sub questions.

**RQ1:** *What covariates can be used as features for the soil organic carbon content prediction?*

In order to answer this research question, we have performed a literature review following the guidelines of Webster and Watson [10] to analyze similar studies which can be found in **Section 2.3**. We have classified the prediction features in the SCORPAN feature categorization. The different categories of the SCORPAN framework are used as inputs for the concept table. We have identified that the Organisms (O) category is the most used. The vegetation index (i.e. NDVI) and the land usage are features that are commonly used as predictors in prediction models. The Relief (R) feature category has become more popular in the past few years (in comparison to the researches performed before 2019). Other promising feature categories are the Climate category (C) and to a lesser extent the Soil properties (S) category.

**RQ2:** *What prediction methods can be used to predict the soil organic carbon content and how can they be validated?*

The same literature review (**Section 2.4**) used to answer RQ1 was used to answer this research question. Regarding the prediction methods, we have identified the following patterns: the Random Forest method is used in the majority of the studies and also outperforms a lot of other methods within those studies. The method has become more popular in comparison to the researches performed before 2019, which is a logical consequence of the research of Lamichhane et al [18], as the Random Forest method already outperformed a large amount of other methods before 2019. The second most used prediction method is Support Vector Machine. Deep learning (Artificial Neural Networks) has been introduced in the last few years and is a promising method as it outperforms other methods in the researches that involve method comparison.

The amount of researches including a validation step has drastically increased in the past few years [18]. An interesting trend shift is that the majority of papers published before 2019 have used the data splitting validation methodology, while papers published from 2019 and later mainly use the cross-validation methodology. The cross-validation methodology, specifically the 10-fold cross-validation, has

become more popular because it makes better usage of data by training your model on multiple train-test splits.

**RQ3:** *Using covariates identified from the literature, how can an empirical model be developed to predict the soil organic carbon content?*

For this research question we have used the method Knowledge Discovery in Databases. We have used the answers of the first research question to develop a prediction model for the soil organic carbon content. Based on the answer to RQ1, we have chosen the following environmental covariates as prediction features (**Chapter 3**):

- *Climate*: Annual mean and variance of daily amount of rainfall
- *Climate*: Annual mean and variance of daily temperature
- *Climate*: Annual mean and variance of daily amount of hours solar radiation
- *Soil propierties*: Mostly present type of underlying soil
- *Organisms*: Cultivation plans of the past 5 years
- *Organisms*: Average NDVI (vegetation index) score of the past year
- *Other*: SoilGrids organic carbon stock and density estimations

The answer to RQ2 consisted of a set of three prediction algorithms: Random Forest, Support Vector Machine and Artificial Neural Networks. We have implemented these three prediction algorithms (**Chapter 4**) and validated the prediction algorithms in three different experiments (**Chapter 5**).

**RQ4:** *What is the predictive accuracy of the developed model and how can it be optimized?*

In order to validate the results of the prediction algorithms, we used the 10-fold cross validation technique based on the results of the literature review performed for RQ2. We have also tried two other validation strategies as we are working with a small dataset, namely 5-fold cross validation and cross validation based on a shuffle split. The results of these strategies can be found in **Chapter 5**.

The Random Forest algorithm resulted in the highest predictive accuracy when using all features as inputs for the prediction: when using 10-fold cross-validation, the $R^2$ value is 0.37 and the Mean Squared Error is 1.76. This is slightly lower than similar researches that show $R^2$ values between 0.4 and 0.8. The other two algorithms showed no prediction potential for this dataset. We reached the highest predictive accuracy when validating the model with shuffle split ($R^2$ = 0.52 and MSE = 1.36). However, these results have a high risk of overfitting as we were working with a relatively small dataset (209 datapoints). Anyway, we can conclude that the environmental covariates add value to the existing soil organic carbon maps for this dataset.

We have performed different optimization iterations by tweaking the hyperparameters of the model. None of the changes in the hyperparameters improved the predictive accuracy of the model significantly.

**RQ5:** *How can the resulting model be used to create potential value for stakeholders?*

We answered this question in **Chapter 6** with the assumption that the model can be further developed and reaches the level of practical usage. For the farmers' organization, the model creates value as the organization can give more tailored advice to farmers which increases the value of their business. The

farmers receive more up-to-date estimations of the state of their soil and tailored advice from the farmers' organization, which enables them to increase their soil sequestration, improve the quality of the soil and reduce carbon dioxide emission. As a long term implication of this project, society can benefit if farmers manage to reduce carbon emission.

**Main RQ***: How to develop a method that predicts the soil organic carbon content based on environmental covariates at farms in the Netherlands so that the company can provide actionable insights on the carbon stocks to farmers?*

This thesis presented the method we developed and used to predict the soil organic carbon content. We gathered different types of environmental covariates (climate covariates, vegetation indices, land usage, underlying soil type and existing soil maps) and tested the performance of different machine learning algorithms. Although the validation of the method showed predictive potential for the Random Forest algorithm, it still remains questionable if these results are generalizable and future research is needed to validate this.

## 7.2 Contributions

We have developed a method that can be used for soil organic carbon predicting for farm plots in the Netherlands. Although the accuracy of the resulting prediction method is relatively low and future research is needed on how to improve this (further discussed in section 7.3), the development of the model already makes contributions to research and practice.

### 7.2.1 Contribution to research

This thesis has the following contributions to research:

1. The literature review we performed on the prediction features, prediction methods and validation strategies. By analyzing the current state of the art in this field, we have gained valuable insights into the various methods employed in similar studies. The results of our literature review offer a clear overview of the existing prediction methods, providing a foundation upon which future research can be built. Moreover, our review can be a valuable resource for other researchers, offering them guidance and insights as they design their own prediction methods. The documentation of the findings of our literature review contributes to the broader knowledge base and to the understanding of the various approaches towards digital soil mapping. This can ultimately lead to the development of more accurate and effective prediction models, with the potential to benefit society in numerous ways.
2. The demonstration of an effective method for collecting the environmental covariates. As a part of our research, we have effectively transformed the existing theoretical knowledge about the identification of environmental prediction covariates into practical, collected data points. The databases, tools, and APIs that we have used for acquiring the data can serve as a guideline for future researchers seeking to collect relevant data points for similar predictions. By documenting and sharing our data collection process, we enable other researchers to conduct similar studies with greater ease and accuracy.
3. The findings regarding the performance of the prediction algorithms. The research has shown that the Random Forest algorithm can provide the most precise and reliable predictions of the soil organic carbon content when compared to other prediction algorithms. The insights on the

performance of the prediction algorithms can guide future researchers in selecting the most appropriate algorithm for their research questions and objectives.

### 7.2.2 Contribution to practice

This thesis is a step towards accurate SOC prediction, but the developed model does not reach the level of practical usage yet. Section 6.2 discusses the potential practical value the artifact has when reaches the level of practical usage. That section also forms the basis for the contributions to practice this thesis has, which are described below.

1. The tailored advice the farmers' organization can give to farmers using this model, which in turn frees them from the dependency on soil sampling frequency of their customers. It enables the organization to play a more proactive role to help farmers achieve success in their agricultural practices. Furthermore, once the full potential of the model has been reached, the farmers' organization is able to deliver more value to the customers. This ultimately leads to the creation of new business opportunities.
2. Through the utilization of this model, farmers are able to receive an up-to-date view of the soil state of their plots, which enables them to make more informed decisions about their agricultural practices. They can use the tailored advice provided by the farmers' organization to increase the organic carbon content in their soil. This is an essential step towards improving the overall quality of the soil and ultimately leads to higher yields and greater profitability.
3. During the data research we have provided the organization with a method to determine the NDVI scores and visualize these scores. The companies that initiated this project wanted stated that they wanted/needed this functionality and are able to reuse this in other contexts as well.
4. If the farmers manage to use the insights from the model and the tailored advice from the farmers 'organization, they are contributing towards the reduction of carbon emission. As the reduction of carbon emission is an important step towards stopping global warming, the development of this model ultimately contributes to achieving these goals.

## 7.3 Limitations and future work

The first limitation of this research is the size of the dataset used as target values. We were not able to evaluate the performance of the different machine learning models due to the limited amount of available data. The goal of this research was to design and develop a prediction method that accurately predict the soil organic carbon content, which has not been accomplished and we are not sure whether this goal could be accomplished with a larger dataset. Therefore we propose that in future work the same method should be evaluated with a larger dataset. This also involves research on how much data is needed to perform an accurate prediction. In this study we have used data gathered by Eurofins-Agro. Future studies could evaluate other data sources and other ways to determine the organic carbon content, if available.

Furthermore, we learned from the interview with the farmers' organization that the organic carbon content does not change very much over years for a farm, and the carbon stock values are close to historical measurements. Due to the lack of measurements, we were not able to use historical carbon measurements as inputs for the organic carbon prediction. We advise to use historical soil measurements as features in order to make the soil organic carbon prediction more accurate. Furthermore, this historical data can be used to analyze the changes in organic carbon and thereby the

following questions can be addressed: What measures of the farmers influence the soil organic carbon content? How well can the change in organic carbon be explained by the measures of the farmers?

The measurements of the organic carbon content from Eurofins-Agro are taken from different depths, either 10 cm or 25 cm. In our research, we assumed that the carbon content is the same (or close) for the different depths of the soil as we did not have enough data points. We advise to test this assumption or split the dataset into groups of samples taken from the same depth.

Another assumption we made was the time span of the NDVI data we used for each measurement. We have chosen to take the NDVI scores of the past year for each pixel lying within the farm plot. This is because the NDVI values of the Sentinel-2 Level 2A dataset are only available from 2017 and later. We propose that in future research a solution for this problem is found, by either using more recent data points or finding an alternative for determining the NDVI values.

In future studies, a closer look on the different features should be taken. In this research we have transformed the soil type and cultivation plans into dummy variables. This resulted in a large number of features. However, a deeper understanding of these variables and a different way of transforming the variables could lead to more valuable insights. For example, one should consider classifying the possible soil types into categories so that similarities and relations between the soil types can be modeled. The same applies to the cultivation plans. We translated each unique combination of a cultivation plan and the amount of 'years ago' into a dummy variable. Other ways of transforming this data (such as counting the amount of times a cultivation plan is used in the past five years or determining the variation of the cultivation plans) could improve the feature importance and give new insights.

Lastly, we propose that the practical usage of the predictions should be investigated. The goal of the developed artifact was to provide insights into the soil organic carbon content on farm plots. Assuming it is possible to give a good prediction on the soil carbon content, the estimations should be presented to potential stakeholders. The knowledge gained on the influence of the cultivation plans should be used as well to provide a tailored advise for farmers when planning the occupation of their farm plots.

# References


[1] L. Benos, A. C. Tagarakis, G. Dolias, R. Berruto, D. Kateris, and D. Bochtis, “Machine Learning in Agriculture: A Comprehensive Updated Review,” *Sensors*, vol. 21, no. 11, p. 3758, May 2021, doi: 10.3390/s21113758.

[2] University of Wageningen, “Boer en bodem: is er een economisch probleem?,” 2015.

[3] T. Zhou *et al.*, “Prediction of soil organic carbon and the C:N ratio on a national scale using machine learning and satellite data: A comparison between Sentinel-2, Sentinel-3 and Landsat-8 images,” *Science of the Total Environment*, vol. 755, 2021, doi: 10.1016/j.scitotenv.2020.142661.

[4] D. Brus *et al.*, “Soil organic carbon mapping: GSOC map cookbook manual.,” 2017.

[5] L. Poggio *et al.*, “SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty,” *SOIL*, vol. 7, no. 1, pp. 217–240, Jun. 2021, doi: 10.5194/soil-7-217-2021.

[6] R. J. Wieringa, *Design Science Methodology for Information Systems and Software Engineering*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014. doi: 10.1007/978-3-662-43839-8.

[7] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, “A Design Science Research Methodology for Information Systems Research,” *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, Dec. 2007, doi: 10.2753/MIS0742-1222240302.

[8] U. Fayyad, “Knowledge discovery in databases: An overview,” 1997, pp. 1–16. doi: 10.1007/3540635149_30.

[9] Nwagu, K. Chikezie, Omankwu, C. Obinnaya, and H. Inyiama, “Knowledge Discovery in Databases (KDD): An Overview,” *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 15, no. 12, 2017.

[10] R. T. Watson and J. Webster, “Analysing the past to prepare for the future: Writing a literature review a roadmap for release 2.0,” *J Decis Syst*, vol. 29, no. 3, pp. 129–147, Jul. 2020, doi: 10.1080/12460125.2020.1798591.

[11] European Commission, “Sustainable Development Goals,” 2015.

[12] European Commission, “How soil organic matter composition affects carbon sequestration,” *EU Science Hub*, Nov. 29, 2019.

[13] D. W. Nelson and L. E. Sommers, “Total Carbon, Organic Carbon, and Organic Matter,” 2018, pp. 961–1010. doi: 10.2136/sssabookser5.3.c34.

[14] H. Blanco-Canqui *et al.*, “Soil organic carbon: The value to soil properties,” *J Soil Water Conserv*, vol. 68, no. 5, pp. 129A-134A, Sep. 2013, doi: 10.2489/jswc.68.5.129A.

[15] C. Koopmans, “Koolstof vastleggen door slim landgebruik,” 2019.

[16] B. Minasny and Alex. B. McBratney, “Digital soil mapping: A brief history and some lessons,” *Geoderma*, vol. 264, pp. 301–311, Feb. 2016, doi: 10.1016/j.geoderma.2015.07.017.

[17] A. B. McBratney, M. L. Mendonça Santos, and B. Minasny, “On digital soil mapping,” *Geoderma*, vol. 117, no. 1–2, pp. 3–52, Nov. 2003, doi: 10.1016/S0016-7061(03)00223-4.

[18] S. Lamichhane, L. Kumar, and B. Wilson, “Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review,” *Geoderma*, vol. 352, pp. 395–413, Oct. 2019, doi: 10.1016/j.geoderma.2019.05.031.

[19] B. R. Wilson and V. E. Lonergan, “Land-use and historical management effects on soil organic carbon in grazing systems on the Northern Tablelands of New South Wales,” *Soil Research*, vol. 51, no. 8, p. 668, 2013, doi: 10.1071/SR12376.

[20] Y. Ma, B. Minasny, and C. Wu, “Mapping key soil properties to support agricultural production in Eastern China,” *Geoderma Regional*, vol. 10, pp. 144–153, Sep. 2017, doi: 10.1016/j.geodrs.2017.06.002.

[21] X. Song, F. Liu, G. Zhang, D. Li, Y. Zhao, and J. Yang, “Mapping Soil Organic Carbon Using Local Terrain Attributes: A Comparison of Different Polynomial Models,” *Pedosphere*, vol. 27, no. 4, pp. 681–693, Aug. 2017, doi: 10.1016/S1002-0160(17)60445-4.

[22] C. Sothe, A. Gonsamo, J. Arabian, and J. Snider, “Large scale mapping of soil organic carbon concentration with 3D machine learning and satellite observations,” *Geoderma*, vol. 405, 2022, doi: 10.1016/j.geoderma.2021.115402.

[23] B. Mahesh, “Machine Learning Algorithms - A Review,” *International Journal of Science and Research (IJSR)*, vol. 9, no. 1, Jan. 2020.

[24] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, “A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science,” 2020, pp. 3–21. doi: 10.1007/978-3-030-22475-2_1.

[25] M. Zeraatpisheh *et al.*, “Improving the spatial prediction of soil organic carbon using environmental covariates selection: A comparison of a group of environmental covariates,” *Catena (Amst)*, vol. 208, 2022, doi: 10.1016/j.catena.2021.105723.

[26] B. Wang *et al.*, “Modelling and mapping soil organic carbon stocks under future climate change in south-eastern Australia,” *Geoderma*, vol. 405, Jan. 2022, doi: 10.1016/j.geoderma.2021.115442.

[27] M. Zeraatpisheh, S. Ayoubi, A. Jafari, S. Tajik, and P. Finke, “Digital mapping of soil properties using multiple machine learning in a semi-arid region, central Iran,” *Geoderma*, vol. 338, pp. 445–452, 2019, doi: 10.1016/j.geoderma.2018.09.006.

[28] L. C. Gomes, R. M. Faria, E. de Souza, G. V. Veloso, C. E. G. R. Schaefer, and E. I. Fernandes Filho, “Modelling and mapping soil organic carbon stocks in Brazil,” *Geoderma*, vol. 340, pp. 337–350, Apr. 2019, doi: 10.1016/j.geoderma.2019.01.007.

[29] H. Keskin, S. Grunwald, and W. G. Harris, “Digital mapping of soil carbon fractions with machine learning,” *Geoderma*, vol. 339, pp. 40–58, Apr. 2019, doi: 10.1016/j.geoderma.2018.12.037.

[30] S. Tajik, S. Ayoubi, and M. Zeraatpisheh, “Digital mapping of soil organic carbon using ensemble learning model in Mollisols of Hyrcanian forests, northern Iran,” *Geoderma Regional*, vol. 20, 2020, doi: 10.1016/j.geodrs.2020.e00256.

[31] K. John, I. Abraham Isong, N. Kebonye, E. Ayito, P. Chapman Agyeman, and S. Marcus Afu, “Using Machine Learning Algorithms to Estimate Soil Organic Carbon Variability with Environmental Variables and Soil Nutrient Indicators in an Alluvial Soil,” *Land (Basel)*, vol. 9, no. 12, p. 487, Dec. 2020, doi: 10.3390/land9120487.

[32] H. Mahmoudzadeh, H. R. Matinfar, R. Taghizadeh-Mehrjardi, and R. Kerry, “Spatial prediction of soil organic carbon using machine learning techniques in western Iran,” *Geoderma Regional*, vol. 21, 2020, doi: 10.1016/j.geodrs.2020.e00260.

[33] Y. Hong *et al.*, “Comparing laboratory and airborne hyperspectral data for the estimation and mapping of topsoil organic carbon: Feature selection coupled with random forest,” *Soil Tillage Res*, vol. 199, 2020, doi: 10.1016/j.still.2020.104589.

[34] K. Wang, Y. Qi, W. Guo, J. Zhang, and Q. Chang, “Retrieval and Mapping of Soil Organic Carbon Using Sentinel-2A Spectral Images from Bare Cropland in Autumn,” *Remote Sens (Basel)*, vol. 13, no. 6, Mar. 2021, doi: 10.3390/rs13061072.

[35] O. Odebiri, O. Mutanga, and J. Odindi, “Deep learning-based national scale soil organic carbon mapping with Sentinel-3 data,” *Geoderma*, vol. 411, 2022, doi: 10.1016/j.geoderma.2022.115695.

[36] M. Emadi, R. Taghizadeh-Mehrjardi, A. Cherati, M. Danesh, A. Mosavi, and T. Scholten, “Predicting and mapping of soil organic carbon using machine learning algorithms in Northern Iran,” *Remote Sens (Basel)*, vol. 12, no. 14, 2020, doi: 10.3390/rs12142234.

[37] R. Taghizadeh-Mehrjardi *et al.*, “Improving the spatial prediction of soil organic carbon content in two contrasting climatic regions by stacking machine learning models and rescanning covariate space,” *Remote Sens (Basel)*, vol. 12, no. 7, 2020, doi: 10.3390/rs12071095.

[38] W. S. Noble, “What is a support vector machine?,” *Nat Biotechnol*, vol. 24, no. 12, pp. 1565–1567, Dec. 2006, doi: 10.1038/nbt1206-1565.

[39] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Muller, “Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications,” *Proceedings of the IEEE*, vol. 109, no. 3, pp. 247–278, Mar. 2021, doi: 10.1109/JPROC.2021.3060483.

[40] M. W. Browne, “Cross-Validation Methods,” *J Math Psychol*, vol. 44, no. 1, pp. 108–132, Mar. 2000, doi: 10.1006/jmps.1999.1279.

[41] D. Berrar, “Cross-Validation,” in *Encyclopedia of Bioinformatics and Computational Biology*, Elsevier, 2019, pp. 542–545. doi: 10.1016/B978-0-12-809633-8.20349-X.

[42] J. A. Reijneveld, M. Knotters, T. P. van Tol-Leenders, J. P. Lesschen, J. W. H. van der Kolk, and P. J. Kuikman, “Het aandeel bodemorganische koolstof (SOC) in bodemorganische stof (SOM): Een eerste verkenning,” Jun. 2021, Accessed: Dec. 14, 2022. [Online]. Available: https://edepot.wur.nl/570424

[43] J. A. Reijneveld, M. J. van Oostrum, K. M. Brolsma, D. Fletcher, and O. Oenema, “Empower Innovations in Routine Soil Testing,” *Agronomy*, vol. 12, no. 1, p. 191, Jan. 2022, doi: 10.3390/agronomy12010191.

[44] S. Jolly and N. Gupta, “Understanding and Implementing Machine Learning Models with Dummy Variables with Low Variance,” 2021, pp. 477–487. doi: 10.1007/978-981-15-5113-0_37.

[45] S. Garavaglia and A. Sharma, “A smart guide to dummy variables: four applications and a macro,” *Proceedings of the northeast SAS users group conference*, vol. 43, 1998.

[46] J. Brownlee, “Why One-Hot Encode Data in Machine Learning?,” 2017. https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/ (accessed Feb. 22, 2023).

[47] A. Ben-Hur and J. Weston, “A User’s Guide to Support Vector Machines,” 2010, pp. 223–239. doi: 10.1007/978-1-60327-241-4_13.

[48] A. Dertat, “Applied Deep Learning - Part 1: Artificial Neural Networks,” *Towards Data Science*, Aug. 08, 2017.

[49] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, “Deep Learning is Robust to Massive Label Noise,” 2018.

[50] Ü. Ağbulut, A. E. Gürel, and Y. Biçen, “Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison,” *Renewable and Sustainable Energy Reviews*, vol. 135, p. 110114, Jan. 2021, doi: 10.1016/j.rser.2020.110114.

[51] J. Padarian, B. Minasny, and A. B. McBratney, “Using deep learning for digital soil mapping,” *SOIL*, vol. 5, no. 1, pp. 79–89, Feb. 2019, doi: 10.5194/soil-5-79-2019.

[52] I. F. Alexander, “A Taxonomy of Stakeholders,” *International Journal of Technology and Human Interaction*, vol. 1, no. 1, pp. 23–59, Jan. 2005, doi: 10.4018/jthi.2005010102.

[53] B. Freedman, G. Stinson, and P. Lacoul, “Carbon credits and the conservation of natural areas,” *Environmental Reviews*, vol. 17, no. NA, pp. 1–19, Dec. 2009, doi: 10.1139/A08-007.

[54] J. Mallick *et al.*, “Spatial stochastic model for predicting soil organic matter using remote sensing data,” *Geocarto Int*, vol. 37, no. 2, pp. 413–444, 2022, doi: 10.1080/10106049.2020.1720314.

[55] F. Castaldi *et al.*, “Evaluating the capability of the Sentinel 2 data for soil organic carbon prediction in croplands,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 147, pp. 267–282, 2019, doi: 10.1016/j.isprsjprs.2018.11.026.

[56] G. B. M. Heuvelink *et al.*, “Machine learning in space and time for modelling soil organic carbon change,” *Eur J Soil Sci*, vol. 72, no. 4, pp. 1607–1623, 2021, doi: 10.1111/ejss.12998.

[57] E. Vaudour, C. Gomez, Y. Fouad, and P. Lagacherie, “Sentinel-2 image capacities to predict common topsoil properties of temperate and Mediterranean agroecosystems,” *Remote Sens Environ*, vol. 223, pp. 21–33, 2019, doi: 10.1016/j.rse.2019.01.006.

[58] F. Veronesi and C. Schillaci, “Comparison between geostatistical and machine learning models as predictors of topsoil organic carbon with a focus on local uncertainty estimation,” *Ecol Indic*, vol. 101, pp. 1032–1044, 2019, doi: 10.1016/j.ecolind.2019.02.026.

[59] K. John, I. I. Abraham, N. M. Kebonye, P. C. Agyeman, E. O. Ayito, and A. S. Kudjo, “Soil organic carbon prediction with terrain derivatives using geostatistics and sequential Gaussian simulation,” *Journal of the Saudi Society of Agricultural Sciences*, vol. 20, no. 6, pp. 379–389, 2021, doi: 10.1016/j.jssas.2021.04.005.

[60] K. Dvorakova, U. Heiden, and B. van Wesemael, “Sentinel-2 Exposed Soil Composite for Soil Organic Carbon Prediction,” *Remote Sens (Basel)*, vol. 13, no. 9, May 2021, doi: 10.3390/rs13091791.

[61] Y. Zhou, A. E. Hartemink, Z. Shi, Z. Liang, and Y. Lu, “Land use and climate change effects on soil organic carbon in North and Northeast China,” *Science of the Total Environment*, vol. 647, pp. 1230–1238, 2019, doi: 10.1016/j.scitotenv.2018.08.016.

[62] S. Chen *et al.*, “Model averaging for mapping topsoil organic carbon in France,” *Geoderma*, vol. 366, May 2020, doi: 10.1016/j.geoderma.2020.114237.

[63] X. Meng *et al.*, “Regional soil organic carbon prediction model based on a discrete wavelet analysis of hyperspectral satellite data,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 89, 2020, doi: 10.1016/j.jag.2020.102111.

[64] M. Zeraatpisheh, E. Bakhshandeh, M. Hosseini, and S. M. Alavi, “Assessing the effects of deforestation and intensive agriculture on the soil quality through digital soil mapping,” *Geoderma*, vol. 363, 2020, doi: 10.1016/j.geoderma.2019.114139.

# Appendix

## Appendix A – Literature review

**Table 12. Concept table Covariate categories**

| | S | C | O | R | P | A | N |
|---|---|---|---|---|---|---|---|
| [3] | | x | | x | | | |
| [22] | x | x | x | x | | | |
| [25] | x | | x | x | | | |
| [54] | | | x | x | | | |
| [26] | | x | x | x | x | x | |
| [27] | | | x | x | | | |
| [28] | x | x | x | x | | | |
| [55] | | | x | | | | |
| [56] | | x | x | x | | | |
| [57] | | | x | | | | |
| [29] | x | x | x | x | | | |
| [51] | | x | | x | | | |
| [30] | x | x | x | x | | | |
| [35] | | | x | | | | |
| [36] | | x | x | x | | | |
| [37] | | x | x | x | | | |
| [31] | x | x | x | x | | | |
| [32] | | x | x | x | | | |
| [58] | | x | x | x | | | |
| [59] | | | | x | | | |
| [60] | | | x | | | | |
| [34] | | | x | | | | |
| [61] | x | x | x | x | | | |
| [62] | x | x | x | x | | | |
| Total: | **8** | **15** | **21** | **19** | **1** | **1** | **0** |

**Table 13. Concept table Prediction methods**

| | RF | Cubist | SVM | PLSR | MLR | NN-RMA | RT | GLM | QRF | BaRT | CaRT | BoRT | RegKr | OrdKr |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| [3] | x | | x | | | | | | | | | x | | |
| [22] | x | | | | | | | | | | | | | |
| [25] | x | x | x | x | | | | | | | | | | |
| [54] | | | | | x | x | | | | | | | | |
| [26] | x | | | | x | | | | | | | | | |
| [27] | x | x | | | x | | x | | | | | | | |
| [28] | x | x | x | | | | | x | | | | | | |
| [55] | x | | | x | | | | | | | | | | |
| [56] | | | | | | | | | x | | | | | |
| [57] | | | | x | | | | | | | | | | |
| [29] | x | | x | x | | | | | | x | x | x | x | x |
| [51] | | | | | | | | | | | | | | |
| [30] | x | | x | x | | | x | x | | | | | | |
| [35] | x | | x | | | | | | | | | | | |
| [36] | x | | x | | | | x | | | | | | | |
| [37] | x | x | | | | | | | | | | | | |
| [31] | x | x | x | | x | | | | | | | | | |
| [32] | x | x | x | | | | | | | | | | | |
| [58] | x | | | | x | | | | x | | | x | x | x |
| [59] | | | | | | | | | | | | | x | x |
| [60] | | | | x | | | | | | | | | | |
| [34] | x | | x | x | | | | | | | | | | |
| [61] | x | | | | | | | | | | | | | |
| [62] | | x | | | | | | | | | | | | |
| [33] | x | | | | | | | | | | | | | |
| [63] | x | | x | | | | | | | | | | | |
| [64] | x | | | | | | | | | | | | | |
| **Total** | **20** | **7** | **11** | **7** | **5** | **1** | **3** | **2** | **2** | **1** | **1** | **2** | **3** | **3** |

**Table 13 (continued)**

| | BPNN | CNN | kNN | ANN | DNN | XGBoost | AvNNet | HK | SGS | GR | BC-VW | BMA | Res-Cub |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| [3] | | | | | | | | | | | | | |
| [22] | | | | | | | | | | | | | |
| [25] | | | | | | | | | | | | | |
| [54] | | | | | | | | | | | | | |
| [26] | | | | | | | | | | | | | |
| [27] | | | | | | | | | | | | | |
| [28] | | | | | | | | | | | | | |
| [55] | | | | | | | | | | | | | |
| [56] | | | | | | | | | | | | | |
| [57] | | | | | | | | | | | | | |
| [29] | | | | | | | | | | | | | |
| [51] | | x | | | | | | | | | | | |
| [30] | | | x | | | | | | | | | | |
| [35] | | | | x | x | | | | | | | | |
| [36] | | | | x | x | x | | | | | | | |
| [37] | | | | x | x | x | x | | | | | | |
| [31] | | | | x | | | | | | | | | |
| [32] | | | x | | | x | | | | | | | |
| [58] | | | | | | | | x | | | | | |
| [59] | | | | | | | | | x | | | | |
| [60] | | | | | | | | | | | | | |
| [34] | | | | x | | | | | | | | | |
| [61] | | | | | | | | | | | | | |
| [62] | | | | | | | | | | x | x | x | x |
| [33] | | | | | | | | | | | | | |
| [63] | x | | | | | | | | | | | | |
| [64] | | | | | | | | | | | | | |
| **Total** | **1** | **1** | **2** | **5** | **3** | **3** | **1** | **1** | **1** | **1** | **1** | **1** | **1** |

**Table 14. Concept table Validation strategies**

| | Cross-validation | | | Data splitting |
|---|---|---|---|---|
| | 5-fold | 10-fold | Other | |
| [3] | | x | | |
| [22] | x | | | x |
| [25] | x | | | |
| [54] | | x | | |
| [26] | | x | | |
| [27] | | x | | |
| [28] | x | | | |
| [55] | | x | | |
| [56] | | x | | |
| [57] | | | x (LOO) | |
| [29] | | | | x |
| [51] | | | | x |
| [30], [35] | | x | | |
| [35] | | x | | |
| [36] | | x | | |
| [37] | | x | | |
| [31] | | x | | |
| [32] | | x | | |
| [58] | | | x (3) | |
| [59] | | | x (2, LOO) | |
| [60] | | x | | |
| [34] | | x | | |
| [61][ | | | x | |
| [62] | | x | | |
| [33] | | x | | |
| [63] | | | | x |
| [64] | | x | | |
| **Subtotal** | **3** | **17** | **4** | **4** |
| **Total** | **24** | | | **4** |

# Appendix B

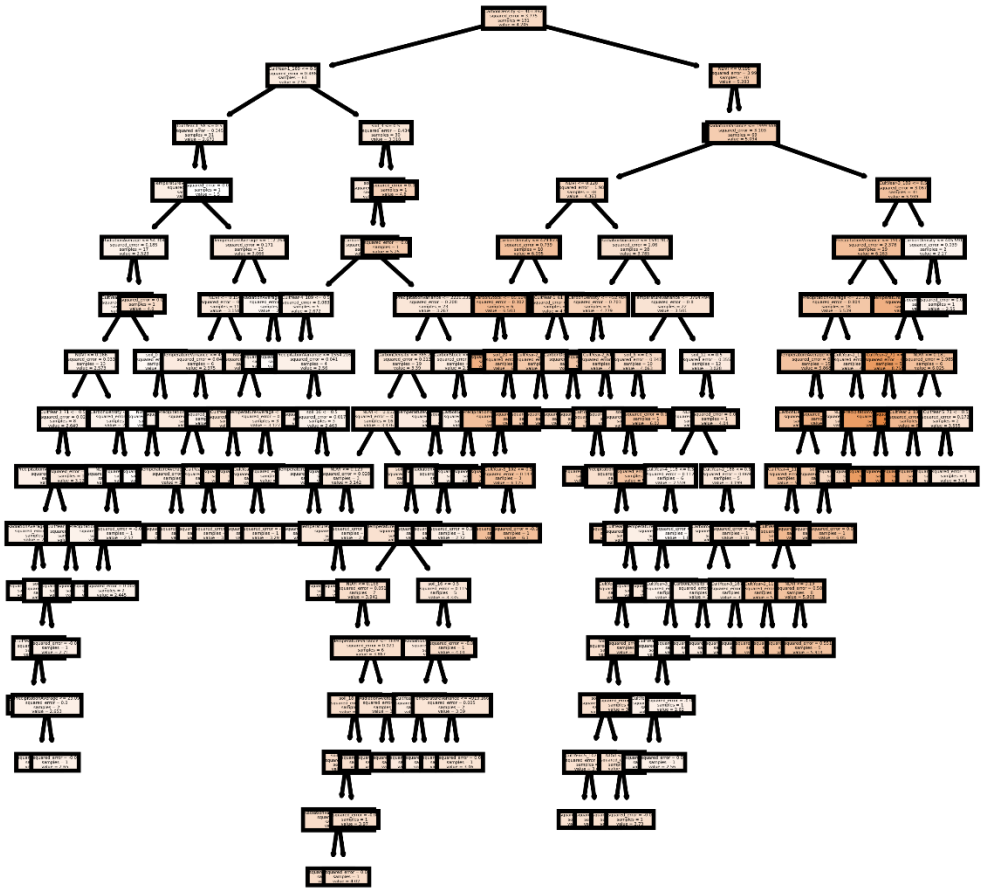

**Figure 17. Example of constructed decision tree**