CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

A machine learning approach for predicting bacteria content in drinking water

A case study for finding a suitable machine learning model including requirements and a recommended implementation

Master's thesis in Computer science and engineering

ERIC JONSSON

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2023

Master's thesis 2023

A machine learning approach for predicting bacteria content in drinking water

A case study for finding a suitable machine learning model including requirements and a recommended implementation

ERIC JONSSON

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2023

A machine learning approach for predicting bacteria content in drinking water
A case study for finding a suitable machine learning model including requirements and a recommended implementation
ERIC JONSSON

© ERIC JONSSON, 2023.

Supervisor: Dana Dannélls, Department of Swedish, multilingualism, language technology
Advisor: Jacob Cahn, Nocoli
Examiner: Marina Axelson-Fisk, Applied Mathematics and Statistics

Master's Thesis 2023
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2023

A machine learning approach for predicting bacteria content in drinking water
A case study for finding a suitable machine learning model including requirements and a recommended implementation

ERIC JONSSON
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

The current method for finding whether drinking water contains bacterial contamination is a very slow process and it can take up to eight days before the results are obtained. During this time, a significant proportion of the population has potentially obtained diseases from contaminated water. As a mitigating action, this thesis aimed to understand if machine learning could be a promising method for forecasting the bacteria level and how such a model could be designed. The project was performed in association with a case company called Nocoli, which is spun out of Chalmers Ventures and desired an examination of the potential implementation. A literature review including eight different case studies of how machine learning was previously applied in the field and three semi-structured interviews with industry-specific stakeholders were conducted. The research methodology originated from the fact that both an overview of the current industry situation as well as machine learning applicability was required. Moreover, by using an extracted theory of machine learning algorithms for different objectives, the case studies were evaluated to find patterns that could meet the case companys demands.

It was found that machine learning is promising and desired in the industry to improve current operations. The Random Forest algorithm was recommended in the initial stage due to its trade-off between accuracy and interpretability. Data on bacterial content and other factors including weather was intended as the data source. The recommendation included a 3:1:1 split between training-, validation-, and test sets as well as using a recursive feature selection algorithm. Additionally, a combination of error measures was recommended including Mean Squared Error with an out-of-bag supplement to reduce overfitting. Furthermore, although no data could be obtained to evaluate the recommended model, it was concluded that machine learning could have a positive impact on today's approach and contribute to improved water management and safety by enabling reliable forecasts.

Keywords: machine learning, forecasting, drinking water quality, contaminated water, drinking water treatment, escherichia coli prediction, HPC method, Random Forest.

Acknowledgements

First and foremost, I would like to express my sincerest appreciation to my supervisor Dana Dannélls at the Department of Swedish, multilingualism, language technology at the University of Gothenburg. Thank you for supporting me with objectivity and helping me understand the important perspectives in this master's thesis. Your expertise and consultation have been invaluable for the project's outcome and your extensive knowledge has provided a guiding foundation for this report.

Secondly, I would like to extend my candid gratitude to the participants of the case company, Jacob Cahn and Jacob Nissén Karlsson, who provided me with this opportunity. Their engagement, ideas, and encouragement contributed to valuable insights which helped me complete this report with conviction and satisfaction.

Eric Jonsson,
Gothenburg, May 2023

Contents

List of Figures

List of Tables

1

Introduction

More than two billion people globally have a main drinking water source contaminated with feces (Worlds Health Organisation, 2022). The contamination results in the transmission of dangerous diseases including diarrhea, cholera, dysentery, typhoid, and polio, especially in developing countries in Asia, Africa, and sub-Saharan. Consequently, the quality of global drinking water is considered one of the main challenges during the 21st century and impacts the potential of societal, economic but also environmental development in the affected areas (UNESCO World Water Assessment Programme, 2021). Moreover, an increased threat has occurred even in countries with more sophisticated water systems (UNICEF, 2022). Aspects such as climate change, increased water scarcity, a growing population, demographic changes, and urbanization put pressure on the existing systems, and by 2025, 50% of the global population is expected to live in water-stressed areas.

To improve the critical situation, actions such as recovering water, nutrients, or energy from wastewater are becoming strategically important, which is enabled by aiding technologies (Zarei, 2020). WHOs water quality guidelines highlight the importance of such technologies to manage risks, both in highly developed areas as well as areas with urgent needs (Worlds Health Organisation, 2022). Especially in developing countries, the bacteria levels in the water sources are important to understand in order to prevent both sickness and death. Prest et al. (2016) describe that it is vital to reach a deeper understanding of bacterial interactions in the distribution system to better manage the bacterial communities when drinking water is produced and distributed. As the demand for bacterial observation has increased during the last decades, different detection methods have been developed (Kumar & Ghosh, 2019). For the new methods to be successful, the authors conclude, they must be complex enough to be accurate for many analyses from different sample matrices, while being fast and cost-effective. As a result, new technology often enables selections of methods based on sample, characteristics, and proficiency in the technology of the users.

Furthermore, besides the technical challenges to detecting bacteria, other aspects also impact the importance and difficulty to provide accurate measurements (Charles et al., 2022). The writers argue that the occurrence of increased waterborne diseases and affected health due to weather-related shocks such as heavy rainfalls have been proven to be associated with several outbreaks of diseases in developing countries.

Moreover, outbreaks are associated both with floods that damage infrastructure systems and loss of water sources in droughts. Additionally, Whitman and Nevers (2003) showed the correlation between E. Coli bacteria levels in water and air as well as water temperatures when the temperature varied between approximately zero and 25 degrees Celsius. A temperature range that is common in Nordic countries. Moreover, LeChevallier et al. (1991) mention temperature and rainfall as two critical factors for bacteria growth in water. Since the two factors often vary within a year, seasons were also suggested as a broader distinguishment of bacterial changes where the regrowth seemed higher during the summer months.

This thesis will be written in collaboration with a Gothenburg-based company called Nocoli and aims to understand if machine learning is an adequate method to predict future bacteria content in drinking water based on current levels and external factors including, among others, rainfall, and temperature. From the companys perspective, besides the purpose above, the company also aims to get insights into how this type of model can be developed to reach optimal results.

1.1 Aim & Research Question

Due to the complexity, importance, and time horizon, the first purpose of this master thesis is to identify the plausibility of using machine learning to predict future bacteria content in drinking water for Nocoli. The second purpose is to design a model that can be applied in this field considering aspects that are required or coveted in terms of features, design choices, and data. With the above as an introduction, the research question can be summarized as;

- **Question:** Is machine learning adequate and how should a machine learning model be designed to predict future bacteria content in drinking water based on today's measures and external factors?
    - How should the model be built?
        - Training procedure
        - Evaluation methods
    - Which model is suitable for the purpose including requirements from the case company?
        - Classification or regression model
        - Trade-off between interpretability and performance
    - Which design choices should be incorporated?
        - Data features
        - Feature selection

1.2 Limitations

First and foremost, a limitation for the readers of this master thesis is the scarce information about the case company. Because the company is driven by profit and since the industry is highly competitive, much of the technology used is confidential. Thus, it is only possible to mention limited information regarding its operations. Although this may diminish the overall understanding of the project, this is a limitation that needs to be made. Secondly, a major limitation of this master's thesis is the lack of data. Generally when it comes to machine learning, one of the considerable risks is the absence of data and if data is acquired, reliability and biases are critical factors. For this master thesis, no data will be obtained which makes the development of a trained model unattainable. As a consequence, this will thesis will be of a more theoretical nature where previous research will be used as the major data source. However, a hypothetical model will be recommended under the assumption that data will be available in the future.

2

Theory

In the section below, relevant backgrounds are presented for a better understanding of the related fields to the subject of this thesis.

2.1 Measuring of bacteria levels in drinking water today

Today at water treatment plants, bacteria analysis often takes days (European Drinking Water, 2017). According to the latest records, the fastest technologies that currently can be used in Scandinavia last for about twelve hours. Nordic microbiologists admit that the waiting time can be highly devastating since a lot of people simultaneously intake contaminated water (Højris et al., 2016). Consequently, there exists a great need for more rapid detection of bacteria.

Svenskt Vatten (2022) says that different methods are used today with different duration. Moreover, all methods are not allowed for public testing since a few are more reliable and can be used for detailed reporting while others only are used in an operating context. Sweden's largest municipal water treatment corporation, Stockholm Vatten och Avfall (2022) mentions daily analyses in various locations of the local drinking water using laboratories. However, no information is specified if the company makes predictions of future levels or if only past levels are considered.

2.1.1 Case company

This thesis was written in collaboration with the Gothenburg-based company Nocoli which is spun out of Chalmers Ventures. Moreover, it is active in bacteria sensing in drinking water. Nocoli's business revolves around a real-time water quality monitoring sensor that provides information regarding the water quality and bacteria levels, more rapidly and cost-efficient than conventional analysis methods. The company's customer sections include waterworks, water service providers, and private wells. Nocoli contributes to value-adding activities by offering the opportunity to prevent negative consequences by acting and managing the water systems proactively instead of taking measures reactively. In Figure 2.1 below, Nocoli has sketched a scheme for the different components of the general water system in Sweden.

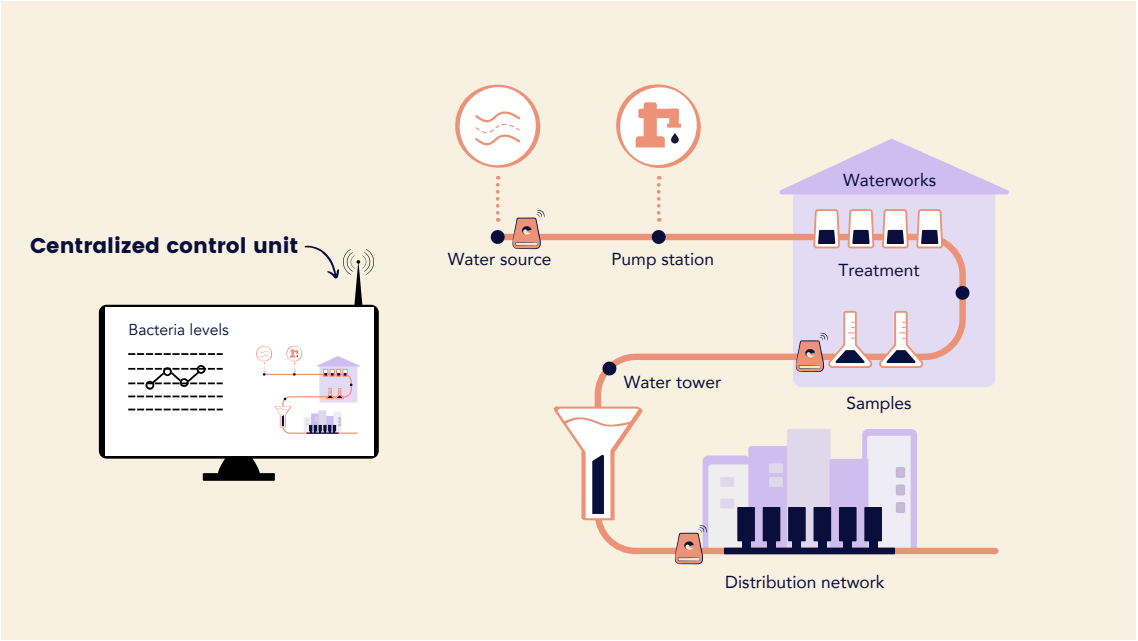

Figure 2.1: The general water system in Sweden (Nocoli, 2023)

The company's sensors can be placed in several locations which can be advantageous to find possible sources of bacteria levels. One challenge that Nocoli is facing is to predict future bacteria content in drinking water, using the current bacteria level, as well as external factors such as rainfall, snowfall, and seasonality. Therefore, this master's thesis will focus on the opportunity of predicting future levels of bacteria using machine learning.

2.1.1.1 Designing a theoretical machine learning model

During the research, a theoretical machine learning model was designed to meet the case company's desires. The literature review in combination with the aim of the case company was primarily used to establish decisions for the architecture.

Important decisions for the theoretical model included:

- Regression model or Classification model (Section 2.3.1)
- Model selection (Section 2.3.2)
- Data (Section 2.3.3)
- Splitting the hypothetical data set for sufficient training (Section 2.3.4)
- Feature selection (Section 2.3.5)
- Error measure (Section 2.3.6)
- A more complex model which is conceivably less interpretable or more interpretability which can result in less accuracy (Section 2.3.7)

2.2 Research methods and data collection

Williams et al. (2007) provide three different approaches to conducting research including quantitative, qualitative, and mixed methods. Moreover, the method chosen should be based on the type of data needed and available to establish an answer to the purpose of the study. Furthermore, Morgan and Harmon (2001) emphasize that it is advisable to use data collection methods that have been shown reliable in previous research similar to the active study. To begin with, the quantitative, qualitative, and mixed approach is presented, followed by a further background on interviews due to their importance in this report.

2.2.1 Quantitative research methodology

The procedure of quantitative research includes gathering data that can quantify the information in order to apply statistical analysis (Creswell & Poth, 2016). Moreover, Leedy and Ormrod (2019) mean that the collected data often is numerical and used in mathematical models to conduct data analysis. In addition, there exist three orientations of quantitative research, namely descriptive, experimental, and causal comparative. Williams et al. (2007) describe descriptive research as using numeric data to examine the current situation by identifying features of events and finding correlations between the features and events. Next, experimental quantitative research is carried out by using an experiment on a particular group and measuring how the group responds to the experiment. Lastly, causal comparative research aims to understand if independent variables impact dependent variables by studying data and analyzing data statistically.

2.2.1.1 Quantitative data analysis

There exist quite a few methodologies to analyze quantitative data that are common in the approaches above (Creswell & Poth, 2016). Examples that commonly are used include procedures called correlational design and developmental design. Leedy and Ormrod (2019) describe correlational research as examining the data to find differences between characteristics of a studied group. An important aspect to consider when performing such an analysis is to find a statistical correlation between the actual characteristics using some kind of specific test. An example of such a test is the Pearson Correlation Coefficient equation which is the most common way of measuring a linear correlation. The equation is provided below where x represents the samples of the x-variable, y corresponds to the samples of the y-variable, $\bar{x}$ and $\bar{y}$ are the means of the samples respectively, and r is the Pearson Correlation Coefficient.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.1}$$

Consequently, validity and reliability are mentioned to be important in correlational research. Additionally, Creswell and Poth (2016) explained correlational research as

trying to prove statistical dependence between two or more variables. For development design, the purpose is to understand how characteristics develop for a group over a defined time horizon (Leedy & Ormrod, 2019). Moreover, the procedure can further be divided into cross-sectional and longitudinal studies where the latter investigates a particular group over a specific period of time while cross-sectional explore corresponding variables for different groups.

2.2.2 Qualitative research methodology

Qualitative research is an iterative process to improve the understanding of a scientific question by obtaining new findings about the phenomenon studied (Aspers & Corte, 2019). Consequently, the authors mention two different criteria for qualitative research including collecting and processing empirical data, iteratively, and improving understanding of the phenomenon. Moreover, Williams et al. (2007) explain qualitative research as more sensitive than quantitative in terms of collecting and interpreting the gathered data. Furthermore, Leedy and Ormrod (2019) mean that qualitative research is less structured since it often is used to formulate new frameworks. Another perspective is that qualitative research occurs in natural circumstances that help the researcher to gain actual experiences within the field (Creswell & Poth, 2016). In addition, qualitative research often builds its premises on inductive, rather than deductive reasoning which means that observations are used to formulate a theory or framework while deductive reasoning often tries to prove a theory using observations. The strong correlation between the observer and the data is a marked difference from quantitative research, where the researcher is strictly outside of the phenomena being investigated (Leedy & Ormrod, 2019). Williams et al. (2007) provide a framework with five different qualitative research areas including case study, ethnography study, phenomenological study, grounded theory study, and content analysis. The framework is also supported by Kyngäs (2020) who also find these categories to be essential. Moreover, Creswell and Poth (2016) state that these areas correspond to different aspects of research.

Creswell and Poth (2016) depict that in a case study, an event, an activity, a process, or one or more individuals is investigated. In addition, Leedy and Ormrod (2019) mention that case studies aim to obtain knowledge about a less researched event and exemplify case studies as including political science research on a presidential campaign or medical research such as diseases. Hence, Williams et al. (2007) states that the data collection for a case study needs to be exhaustive from multiple different data sources such as observations, and interviews with deep knowledge about the explored event.

Ethnography studies differ from case studies because this type aims to understand an entire population following a distinct culture (Leedy & Ormrod, 2019). Creswell and Poth (2016) emphasize that an ethnography study aims to explore transformations in the behavior of a group over time. Consequently, the findings are often not general. Creswell and Poth (2016) highlight that similarly to case studies, the data collection should be extended from multiple different data sources.

The third pillar in the framework consists of grounded theory research and has the purpose to derive a general, abstract theory of an event grounded in the research objects (Williams et al., 2007). Leedy and Ormrod (2019) further describes that grounded theory research regards formulating practical actions into a theory. Often in grounded theory research, data is gathered through interviews or observations, then analyzed in an iterative process until a theory is formed. Additionally, Creswell and Poth (2016) highlight that validation often is crucial in order to reach a credible theory.

Furthermore, a phenomenological study tries to understand a phenomenon by exploring the experiences from a research objects point of view (Williams et al., 2007). Creswell and Poth (2016) mean that the participants opinions of the activity are essential to understand the event. To obtain the data needed for such a study, long interviews with the participants are required in order to understand the conscious, experience, and event in order to interpret the situation properly (Leedy & Ormrod, 2019).

The last version is content analysis study and Leedy and Ormrod (2019) defines this as a thorough investigation of a particular body of materials in order to find the contents and attributes. Creswell and Poth (2016) state that content analysis research is a method that examines human communication such as articles, books, movies, and audio material to understand patterns, themes, or biases.

2.2.2.1 Qualitative data analysis

An important aspect of qualitative data analysis is to interpret the data correctly (Bell et al., 2022). Moreover, the authors mean that qualitative data analysis often is complex due to the large amount of data and thereby a lot of existing details, but also the difference in structure and language. Two different approaches that often are considered when analyzing data are content analysis or thematic analysis (Vaismoradi et al., 2013). The first technique is based on exploring the presence of certain elements, such as words, numbers, or concepts within the data to answer the research questions. Thematic analysis searches for patterns in the data rather than elements which makes it potentially more biased and difficult to interpret than content analysis. An example of such an analysis is given by Castleberry and Nolen (2018) where interview objects' answers are analyzed based on their relationships for finding reasonable patterns. Bell et al. (2022), suggests thematic analysis as a proper analysis method when dealing with qualitative data since general themes and patterns are looked for to be able to distinguish answers to the research questions. On the other hand, content analysis is regarded as better suited for quantitative research.

2.2.3 Mixed research methodology

A third research methodology is the mixed approach (Johnson & Onwuegbuzie, 2004). This uses a combination of the ones above to obtain the advantages of each

of the quantitative and qualitative research methodologies. However, Williams et al. (2007) highlight that the strengths and weaknesses that come with each approach are not general and need to be evaluated for each respective research study. Johnson and Onwuegbuzie (2004) point out that research studies can test and build theories by using the advantageous parts in terms of data collection and data analysis methods from both quantitative and qualitative approaches.

2.2.4 Data collection using interviews and observations

There exist two ways of collecting data for a research study without being given a complete data set (Morgan & Harmon, 2001). Observation studies are defined as the procedure where the researchers observe and record actions in order to collect useful data. In contrast, interviews or questionnaires are a type of survey-based research method where a sample of participants is chosen to answer questions orally or in writing. While questionnaires are based on questions that the respondents answer in writing, interviews are performed orally. The population sample should be chosen in order to represent the entire society. Sapsford and Jupp (1996) mean that observations have some crucial advantages compared to interviews. Firstly, the information is received directly by the observer instead of receiving facts through an interview object which may increase credibility. Secondly, the perspective becomes broader since the observer can understand perspectives that the respondent can not. Thirdly, there exist individuals, such as babies and animals, that can not be a part of an interview since they are not able to speak for themselves. Additionally, a few limitations are mentioned including the impossibility of observing due to environmental circumstances. An example of this can be to observe an elite army. Other limitations are the aspect of behavioral changes from the observed population due to the unusual situation and the risk of bias. Something that the observer needs to consider before evaluating the findings.

A few types of interviews exist in terms of structure where the three categories unstructured, semi-structured, and structures are used (Bell et al., 2022). The difference considers how the questions are prepared. For structured interviews, the questions are predetermined and every interview object is asked the same question which makes comparisons simple. On the contrary, unstructured means that no questions are decided in advance which often gives more open and unprepared answers. Certain types of interviews are semi-structured which is described as efficient since rich and detailed answers can be gathered in combination with the flexibility for the interviewee to elaborate on different questions depending on his or her preference. Semi-structured interviews can be conducted using a list of questions that make the interview progress effectively without affecting the interview objects answers. The authors also emphasize the flexibility of this approach since follow-up questions easily can be used without broadly leaving the list.

Moreover, Opdenakker et al. (2006) describe questionnaires as advantageous since the researcher can compose questions as well as the respondent can answer them conveniently without noise disturbance due to the independence of place and time.

On the other hand, lack of social cues is mentioned as a disadvantage since the questioner can not obtain more insights through social reactions. For both interviews and questionnaires, using remote versions is also suggested as an accurate method (Bell et al., 2022). The reason is mainly due to the geographical differences between the author and interviewees which potentially can save time and result in a larger amount of data sources.

2.3 Machine Learning

In recent years, machine learning has been introduced as a comprehensive approach, enabling complex data analysis and future predictions in a short amount of time (Batanlar & Özuysal, 2014). Furthermore, a distinction can be made between three types of machine learning applications, namely supervised, semi-supervised and unsupervised learning. The first category, supervised learning, is based on predicting which class the input belongs to after the model has been trained on a labeled data set. A method that often is considered logical and straightforward. On the other hand, when unsupervised learning is used, the aim is often not to obtain a distinct label on the input but to identify different patterns that the data contains (Mahesh, 2020) Lastly, semi-structures learning corresponds to a combination of the two where a small part of the data is labeled and a larger part is unlabeled (Van Engelen & Hoos, 2020).

2.3.1 Classification and Regression

Besides supervised, semi-supervised, and unsupervised machine learning, a common distinguishment within supervised machine learning is classification models and regression models. The section below discusses the aims and differences between the two groups.

2.3.1.1 Classification

Osisanwo et al. (2017) describe machine learning classification as a method for making a model take data-driven decisions to divide the data into different distinct groups based on linear combinations of their feature values. In other words, this means using input vectors including features for hyperplane decisions to classify the input. Generally, classification is favorable for tasks where the data points have many variable properties including similarities and differences but still a fundamental quality that identifies them (El Naqa & Murphy, 2015). Hence, classification interprets those properties and classifies the new data point with the proper label. Examples, where classification is used, are Jajodia and Garg (2019) who made a model that classifies whether images contain dogs or cats, and Li et al. (2020) who identified heart diseases using machine learning classification in e-healthcare. Moreover, Ayodele (2010) states that linear classifiers often are beneficial where fast decisions are required. Osisanwo et al. (2017) also highlight that quite a few models are adequate for classification which among others include Logistic Regression, Support Vector Machine, Random Forest, and Neural Networks.

2.3.1.2 Regression

In contrast to classification, Ahmad and Chen (2020) describe regression as machine learning algorithms that estimate a specific value to a task such as future energy load using other information including sunlight and wind. Furthermore, Maulud and Abdulazeez (2020) mean that regression can be used for two specific cases. Firstly, for forecasting and prediction where future values are predicted based on dependent variables. Secondly, regression is used to determine correlations between independent and dependent variables such as air temperature and water temperature which clearly are dependent (Webb & Nobilis, 1997). Consequently, regression is adequate for finding correlations between a specific variable and a data set containing points with other features. As a result, machine learning regression has been used in successful studies previously. Examples, which aim for different research fields are Zhang and Hong (2021), who proposed an approach for electric forecasting with support vector regression, Pereira and Cerqueira (2022) who used machine learning regression methods to forecast hotel demand for revenue management, and lastly Chou and Nguyen (2018) who predicted stock prices using sliding-window metaheuristic-optimized machine learning regression.

2.3.2 Machine Learning algorithms

Ray (2019) provides a review of several different machine learning algorithms that are used for solving various kinds of analysis. Depending on the available data and the task for the model, different models are used. In the sections below, some of the most common approaches that currently are used are presented.

2.3.2.1 Gradient Descent

The first algorithm that Ray (2019) mentions is Gradient Descent. The approach is aiming to minimize a cost function through iterations using the partial derivative (Jurafsky & Martin, 2023). The method proceeds until it converges at the local minimum of the cost function using coefficients that are updated in each iteration in combination with the derivative and a predetermined learning rate that determines the size of the steps taken to reach the minimum (Ruder, 2016). There are a few different versions of which Stochastic Gradient Descent is the most widely used (Ray, 2019). The others include Batch Gradient Decent and Mini-Batch Gradient Decent (Ruder, 2016). Jurafsky and Martin (2023) means that one of the main advantages of Gradient Descent is the improvement rate by each iteration. However, a few disadvantages include large computational requirements, a risk of missing the local minimums if the learning rate is too high as well as a risk of slow convergence if the learning rate is too low. A diminishing action is a changing learning rate that becomes lower when the model starts to approach the minimum.

2.3.2.2 Linear Regression algorithm

The next algorithm that is depicted is Linear Regression which tries to fit the data to a straight hyperplane (Kavitha et al., 2016). Maulud and Abdulazeez (2020) mean that linear regression often can be described using the following equation:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{2.2}$$

where y is the dependent variable, x is the independent variable, β are coefficients and ϵ corresponds to the residual. Jurafsky and Martin (2023) describe that the method is efficient when the correlation of parameters in the data is linear which makes the method easy to implement and interpret. Maulud and Abdulazeez (2020) highlight two different application areas for Linear Regression including forecasting and prediction but also to determine causal relations between the independent and dependent variables. However, a significant disadvantage of the method is that the method only can manage linear problems which leaves the model simplifying more complex problems (Ray, 2019). Since the majority of real-world problems are non-linear, the model can only be used as a guide and not as a credible alternative in these cases.

2.3.2.3 Multivariate Regression algorithm

In contrast to the linear regression algorithm mentioned above, Maulud and Abdulazeez (2020) state that the Multivariate Regression algorithm finds many-to-one, non-linear, relationships between input and output variables. Consequently, Multivariate Regression can mathematically be explained as:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m + \epsilon \tag{2.3}$$

where again y is the dependent variable, x is the independent variable, β are coefficients and ϵ corresponds to the residual. Accordingly, a major advantage of this algorithm is its easier applicability to real-world problems (Sarker, 2021). A few demerits that are mentioned include its high complexity which yields an obstacle to understanding alongside a crucial demand for high statistical knowledge (Ray, 2019).

2.3.2.4 Logistic Regression

Logistic Regression is a regression approach to deal with classification problems that aim to find the probability of a specific outcome (Schober & Vetter, 2021). The approach can manage both multinomial outcomes which means the probability of different classes as well as ordinal outcomes including rankings (Ray, 2019). Generally, the author means that Logistic Regression is advantageous in terms of computability, easiness of use, and training efficiency. Another advantage, stated by Maalouf (2011) is that the algorithm is adaptable to manage critical data mining challenges, such as missing data, redundant attributes, and non-linear separability. Hence, the author establishes Logistic Regression as a powerful and resilient data mining method. Furthermore, due to its scaling benefits, it is applicable to a large amount of data as well as in businesses since the probability of outcomes is common

in real-world tasks (Ray, 2019). Lastly, Subramanian and Simon (2013) emphasize that the method can not solve non-linear problems and is easily overfitted to the training data.

2.3.2.5 Decision Tree and Random Forest

In Decision Trees, the data is iteratively split based on dividers that enable predictions depending on the task (Ray, 2019). Ali et al. (2012) explain that the architecture is tree-like where the result of the split is represented by the leaves of the decision node. There is a distinction between classification and regression where in regression, the thresholds are represented by a continuous number while in classification are based on if the data point has the dividing characteristic or not (Ray, 2019). Moreover, a Random Forest is a set of Decision Trees that work interactively to make the predictions (Ali et al., 2012). The Random Forest makes the predictions by evaluating the answer from each separate Decision Tree which often contributes to generalization (Horning et al., 2010). Pickles et al. (2020) have in Figure 2.2 illustrated the correlation between decision trees and Random Forests and how the Random Forests use the predictions from each tree to draw conclusions.

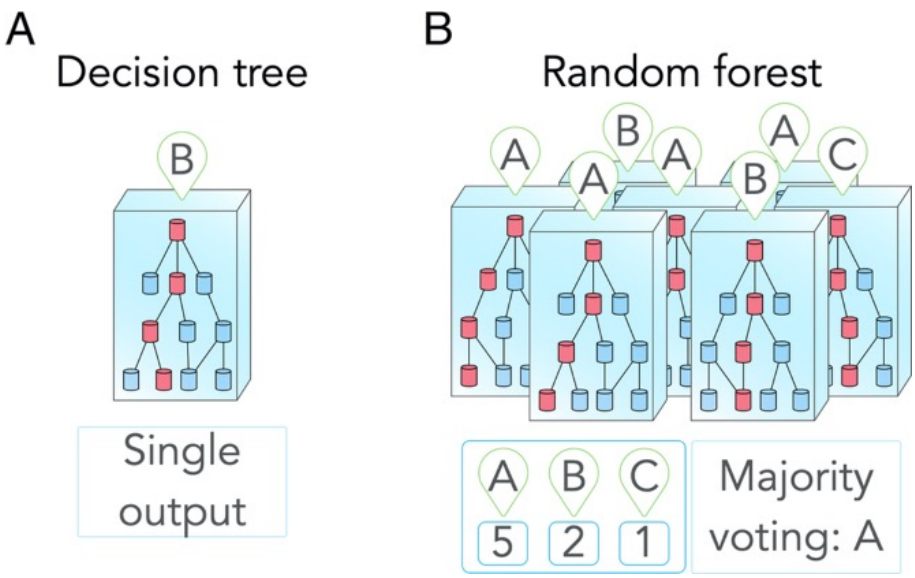

Figure 2.2: Illustration of decision trees and Random Forests

A potential disadvantage of Decision Trees is overfitting, while Random Forest can be used as a mitigating action to this limitation (Fratello & Tagliaferri, 2018). The advantages of Decision Trees and Random Forests are interpretability, broad applicability, and established reliability in terms of performance. Additionally, Decision Trees can be unstable, problematic to tune, and sensitive to sampling errors.

2.3.2.6 Support Vector Machine

Support Vector Machine is an algorithm based on separating the training data by maximizing the distance to the closest point from each data label which is then used

to classify the test data (Bishop & Nasrabadi, 2006). Ray (2019) mentions Support Vector Machines as a method that is efficient for both classification and regression. The advantages of the method include flexibility in terms of data structure, and the complexity of functions that divide the data labels as well as proneness of overfitting (Pisner & Schnyer, 2020). Suthaharan and Suthaharan (2016) distinguish between two different kinds of Support Vector Machines, namely linear and nonlinear models. If the data can be split linearly to separate the classes, the model is called Linear Support Vector Machine. Consequently, if the data cannot be split linearly, but can be transformed into another dimension where a linear split can be performed, it is called Nonlinear Support Vector Machine. Moreover, Pisner and Schnyer (2020) mean that the algorithm distinctively provides balanced predictive performance, even in studies where sample sizes may be limited. Ray (2019) highlights the disadvantages of reduced accuracy with noisy and large data sets, as well as low interpretability.

2.3.2.7 Naïve Bayes

The Naïve Bayes method is an algorithm based on conditional probability using a constantly updated probability table based on the data features in the training data (Bishop & Nasrabadi, 2006). Jurafsky and Martin (2023) explain that the model uses the probability table by predicting the most common alternative. Due to the fact that conditional independence of features is assumed, the method is called naive since it simplifies reality (Ray, 2019). As a consequence, alternative properly trained and tuned models often become superior to the Naïve Bayes (Bishop & Nasrabadi, 2006). Additionally, the method requires a lot of memory and if one of the data features is continuous such as time, Naïve Bayes works poorly (Ray, 2019). Kaur and Oberai (2014) also depict that the algorithm can not handle previously unseen labels which also contributes to weak performance. However, the algorithm is simple to develop, often performs tolerable, and is useful in several circumstances including continuous and discrete data, as well as multi-class classification. Another advantage is that it only requires a small amount of training data to estimate the parameters necessary for predictions (Kaur & Oberai, 2014).

2.3.2.8 K Nearest Neighbor algorithm

Ray (2019) describes the K Nearest Neighbor algorithm as a simple reliable algorithm that is flexible for both binary and multi-class classification. Bishop and Nasrabadi (2006) describes that the architecture builds on classifying a new data point based on matching the data point's features with the different classes' general features in order to optimize the similarities. An important phase of the model development is the decision on which number K should have (Kramer & Kramer, 2013). The K corresponds to the number of neighbors that are considered when making the prediction. If K is small, then only the closest neighbors are considered and this often results in smaller neighborhoods. On the contrary, if K becomes larger, more neighbors are included which results in larger neighborhoods. Alaliyat (2008) shows the procedure of the algorithm in Figure 2.3 where the new data point would be red if $k = 3$ and blue if $k = 5$.

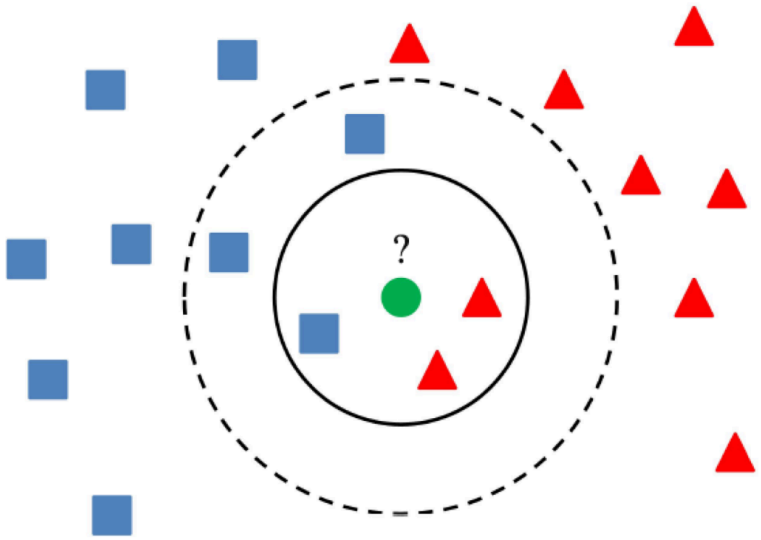

Figure 2.3: Illustration of K Nearest Neighbor algorithm

Furthermore, although the model is simple to implement, the model faces obstacles when treating unknown data sets since it requires distance computation (Kolahdouzan & Shahabi, 2004). In addition, when the size of the data set increases, the computational requirements increase heavily which is connected to the fact of decreased accuracy if the data becomes noisy (Ray, 2019).

2.3.2.9 Neural Networks and the Backpropagation algorithm

Ray (2019) means that the Backpropagation algorithm is a common algorithm that is used when developing Neural Networks. A Neural Network is an architecture with different layers where the layers consist of nodes that obtain information from an input node, using a mathematical function to process the information and then send it to the next node (Bishop & Nasrabadi, 2006). Jurafsky and Martin (2023) describe that the architecture of a Neural Network often consists of an input layer and hidden layers where input nodes are multiplied with trained weights using an activation function. The magnitude of the weights signifies their impact on the result. The activation function provides the output of the specific layer which then is sent to the next layer. After all layers are activated, the information comes to the output layer which gives the result of the model (Bishop & Nasrabadi, 2006).

Ray (2019) suggests Backpropagation as a good method for Deep Learning Neural Networks since it offers an efficient way to compute the gradient. Furthermore, da Silva et al. (2017) explains that the algorithm is used to train the network since the weights are adjusted as a result of the established input and output in the training data. Moreover, Ray (2019) highlights that Neural Networks are very applicable when the prediction patterns are very complex. Consequently, the approach is useful when the tasks are complex and the solutions are based on higher dimensions. Due to the complexity, Liang et al. (2021) describes the approach as a black box where

the interpretability of answers is low. Additionally, Ray (2019) mentions a drawback in terms of slow and computationally heavy Backpropagation learning when the number of hidden layers increases which potentially can reduce the useability in obscure circumstances.

2.3.3 Data collection and structure

One of the most important perspectives to train a machine learning model is to have a reliable data set (Alzubi et al., 2018). Jain et al. (2020) depict that the maximal performance of a model is bounded by the data quality. Additionally, Baryannis et al. (2019) mean that researchers focus on improving model accuracy by tuning hyperparameters and algorithms instead of optimizing the data. Alongside architecture, the data needs to be understood and validated to avoid inaccurate analysis and unreliable decisions (Jain et al., 2020). An instance of low-quality data, that needs to be examined when collecting a data set, is the risk of biased data (Paullada et al., 2021). The authors exemplify biased data by mentioning human annotators which often is used today and can unconsciously contribute to incorrect data. Jo (2019) meant that by investigating the data quality and handling such errors in the data effectively, the model can be improved both in terms of accuracy and reliability. However, Paullada et al. (2021) highlight the difficulty in handling the errors but recommend that the model developer needs to be attentive to reduce the risk of misleading models. Furthermore, Jain et al. (2020) mention several perspectives that are important to consider when evaluating the data. First, label noise is depicted which means that the data points have an incorrect label. A fact that is common on larger data sets. Secondly, the class imbalance is recognized which highlights the importance of a balanced amount of input data in order for the model to perform unbiasedly. Thirdly, data homogeneity is described as a risk which means that the data needs to have different characteristics for proper predictions.

Moreover, Brownlee (2020) explains that the data consists of different features which is a representation of an aspect in the raw data. This representation can be numeric, including binary or continuous numbers, or strings including words. Smitha and Bharath (2020) highlight that the latter often is transformed into integers in order to be efficient for model training. An example can be that *Red* equals *1* and *Blue* corresponds to *2*. Further, Brownlee (2020) states that the data often needs to be structured before it can be applied in the model since it often is collected in unfavorable shapes. The most adequate way of structuring the data is by arranging it as tabular. The rows correspond to an instance, which can be an instance of time or an image, and the columns contain the additional features of the instance. Also, one of these columns corresponds to the output variable, which is to be predicted by the model, and the rest are the input features. An example of such data is the well-known data set, *Iris Species* which consists of 3 types of iris plants (Kaggle, 2023). A fragment of the data set is presented in Table 2.1 below.

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
5.1	3.5	1.4	0.2	Setosa
4.9	3.0	1.4	0.2	Setosa
7.0	3.2	4.7	1.4	Versicolor
6.4	3.2	4.5	1.5	Versicolor
6.1	2.6	5.6	1.4	Virginica
6.3	3.4	5.6	2.4	Virginica

Table 2.1: A fragment of the Iris Species data set

2.3.4 Data preprocessing

Data preprocessing includes preparing the raw data to work in the model (Chu et al., 2016). One of the first aspects to consider regarding data preprocessing is dealing with incomplete data rows. Brownlee (2020) describes that two different approaches exist to deal with such situations, where the first corresponds to removing the rows with the missing values and the second consists of data imputation. The latter is explained as using a method to estimate the missing values. Which method is suitable depends on the situation but common approaches include using mean, median, or applying a K Nearest Neighbor model to find reasonable values. Osman et al. (2018) highlight that both approaches can be efficient where removing missing values is described as a simple technique that often is applicable. However, the method has o few disadvantages. If the data set is too small, removing rows can contribute to a data set that is insufficient for the model. Additionally, the data set can be biased. On the other hand, is data imputation complex in its nature and if the estimation method is inaccurate, the data can be worse than before the preprocessing. Nevertheless, if applied properly, one can obtain a larger reliable data set which can improve the conditions for the model. Additionally, Chu et al. (2016) mean that the main objective of removing or generating missing values is to obtain a data set that has complete coverage for each time point.

Furthermore, a consideration when developing a model is to split the data into three distinguished sets, namely the training set, the validation set, and the test set (da Silva et al., 2017). The authors suggest this since it often reduces overfitting and enhances accuracy because the model is keener to be generalized where the procedure includes pivoting if the model performs heavily worse on the test set than on the training set. Lindholm et al. (2022) describe the training set as the largest set, from which the model finds patterns to make the final predictions. Furthermore, the validation set is used for tuning and smaller improvements. Lastly, the test set is often the smallest set and this is used to obtain the models current accuracy (da Silva et al., 2017). The authors highlight that each set holds a critical role in the development and that the ratio and method to split the sets must be contemplated before the final implementation.

2.3.5 Feature selection

Feature selection is a process where some of the features, which correspond to external parameters that work as the base for the predictions, are chosen out of the original set based on a specific principle (Cai et al., 2018). According to the authors, this plays a critical role in compressing data and removing insignificant information. Lindholm et al. (2022) mention feature selections as an important part of machine learning design which aims to select the most important features for the machine learning models predictions. Moreover, this is described as a critical step of the development process because an excessive number of features otherwise risks being used (Muthukrishnan & Rohini, 2016). A weakness that in the long run can contribute to an increase in training time, dilution of important features, and overfitting. If feature selection is applied properly, redundant information will be removed in combination with a low level of information dilution. Lindholm et al. (2022) mean that this can lead to a higher level of interpretability alongside lower computational necessity. Cai et al. (2018) claim that a good feature selection enables improved performance alongside reduced learning time, and simplified learning results. Dhal and Azad (2022) highlights the difficulty in feature selection since an analysis of the importance of distinguished features needs to be conducted. However, the authors mention advantageous techniques to do so which among others include statistical measure-based, probability measure-based, and sparse learning measure-based feature selection. Additionally, Rong et al. (2019) elaborate on the drawbacks and emphasize that selection methods sometimes can eliminate important factors, which concludes that feature selection needs to be managed carefully.

2.3.6 Error measures

Error measures are tools that are used to evaluate a models performance and are in the majority of models the same in evaluation, validation, and testing (Lindholm et al., 2022). There exist quite a few different metrics to evaluate the performance of machine learning. Vandeput (2021) mentions that commonly used measures in regression models include:

- **Mean Squared Error (MSE)** - Mean of the squared difference between predicted and actual output which punishes large errors more than small ones. The mathematical formula of the equation is given below where D is the number of predictions, y_i is the *i:th* predicted value and x_i is the corresponding actual value.

$$MSE = \sum_{i=1}^{D}(x_i - y_i)^2 \tag{2.4}$$

- **Mean Absolute Error (MAE)** - Mean of the absolute difference between predicted and actual output which punishes large differences less than MSE. The mathematical formula of the equation is given below where D is the number of predictions, y_i is the *i:th* predicted value and x_i is the corresponding actual value.

$$MAE = \sum_{i=1}^{D} |x_i - y_i| \tag{2.5}$$

- **Mean Absolute Percentage Error (MAPE)** - Mean of the absolute difference between predicted and actual output as a percentage. The mathematical formula of the equation is given below where D is the number of predictions, y_i is the *i:th* predicted value and x_i is the corresponding actual value.

$$MAPE = \frac{1}{D} \sum_{i=1}^{D} \frac{|x_i - y_i|}{x_i} \tag{2.6}$$

- **Root Mean Squared Error (RMSE)** - The root of the mean of the squared difference between predicted and actual output which is advantageous since it can be compared to the real data. The mathematical formula of the equation is given below where D is the number of predictions, y_i is the *i:th* predicted value and x_i is the corresponding actual value.

$$RMSE = \sqrt{\frac{1}{D} \sum_{i=1}^{D} (x_i - y_i)^2} \tag{2.7}$$

- **Coefficient of Determination (R^2)** - The ratio of the predicted variation subtracted by the mean of actual values and actual variation subtracted by the mean of actual values. The mathematical formula of the equation is given below where D is the number of predictions, y_i is the *i:th* predicted value and x_i is the corresponding actual value.

$$R^2 = \frac{\sum_{i=1}^{D} (y_i - \frac{1}{D}(\sum_{i=1}^{D} x_i))}{\sum_{i=1}^{D} (x_i - \frac{1}{D}(\sum_{i=1}^{D} x_i))} \tag{2.8}$$

Jierula et al. (2021) recommend using a combination of different types of error measures since it results in a more complex evaluation due to their different intrinsic characteristics.

2.3.7 Interpretability and Performance

Furthermore, a common trade-off in machine learning applications is the one between interpretability and model performance (Baryannis et al., 2019). Freitas (2019) mentions that machine learning algorithms are used by an increasingly larger and more diverse set of users which includes individuals with no machine learning experience. In addition, due to the amount of research, new, more complex models have been developed which increases the difficulty for these new inexperienced users even more. As a result, the author establishes a growing interest in developing interpretable models since individuals need to understand the results in order to make recommendations. Additionally, legal requirements force more interpretable models. Moreover, Baryannis et al. (2019) mean that prediction performance often is prioritized while

interpretability is neglected. Although performance is important, if the models are to be used in practice, the interpretability aspect needs to be enhanced as well since practitioners need to understand the reason in order to simplify decisions and prevent risks. However, Baryannis et al. (2019) show that enhancing interpretability can have a significant negative impact on performance score. A case that highlights the considerations that need to be made around this trade-off (Freitas, 2019).

2.3.8 Machine Learning and Forecasting

In machine learning, models developed for time-series forecasting predict future values $y_{i,t}$ at time t and entity i (Masini et al., 2023). Moreover, for the model to be able to make predictions, a set of time-wised information, such as different statistics or measurements is needed. The forecasting can last for a different duration where the simplest occasion corresponds to one-step-ahead predictions. Kamalov et al. (2021) describe forecasting as one of the major areas within machine learning and the wide range of forecasting applications for which the approach is used. Examples are, among others, Cifuentes et al. (2020) who reviewed several different machine learning approaches to forecast air temperature, Onyema et al. (2022), who focused on the usage of machine learning to forecast academic planning routines, and Kamalov et al. (2021) who explored how machine learning forecasting can be used for stock price returns. Due to a large amount of research within machine learning forecasting, Cifuentes et al. (2020) argue about built-in challenges that need to be considered in the procedure of forecasting development to reach an accurate model.

Onyema et al. (2022) mention several limitations of a machine learning approach for forecasting such as proneness to errors, data acquisition, and other time-consuming issues. Furthermore, a challenge that reduces the correctness of machine learning approaches is evaluating the models objectively (Makridakis et al., 2018). The authors insist on the importance of understanding that new models, based on Artificial Intelligence, are not by default more accurate than old methods and need to be carefully considered when developed. However, Onyema et al. (2022) suggest that machine learning is an efficient method for forecasting since it often takes better and more data-driven decisions. Moreover, Makridakis et al. (2018) depict that machine learning uses similar techniques as older approaches where a loss function, such as the sum of squared errors, is minimized. The difference lies in the technical complexity of doing this minimization which often contributes to better results from the machine learning approach. Kamalov et al. (2021) also suggest machine learning as a promising method for forecasting since it is able to analyze and interpret a large amount of information and correlations in an efficient manner.

2.3.8.1 Uncertainty using forecasts

All kinds of forecasts, from weather to stock prices, include some kind of uncertainty (Nadav-Greenberg & Joslyn, 2009). Furthermore, the authors claim that when information about uncertainty is given properly, it can improve decision-making. Consequently, it is crucial that the inherent uncertainty from forecasts is noted and evaluated. An example is SMHI's forecast on precipitation in Figure 2.4 below

which shows the accuracy of the forecasts from each month (SMHI, 2023). A forecast is recognized to be *true* if the residual between the predicted amount and the actual amount is less than three millimeters per three hours, or in a mathematical formula: $residual \leq 3mm/3h$.

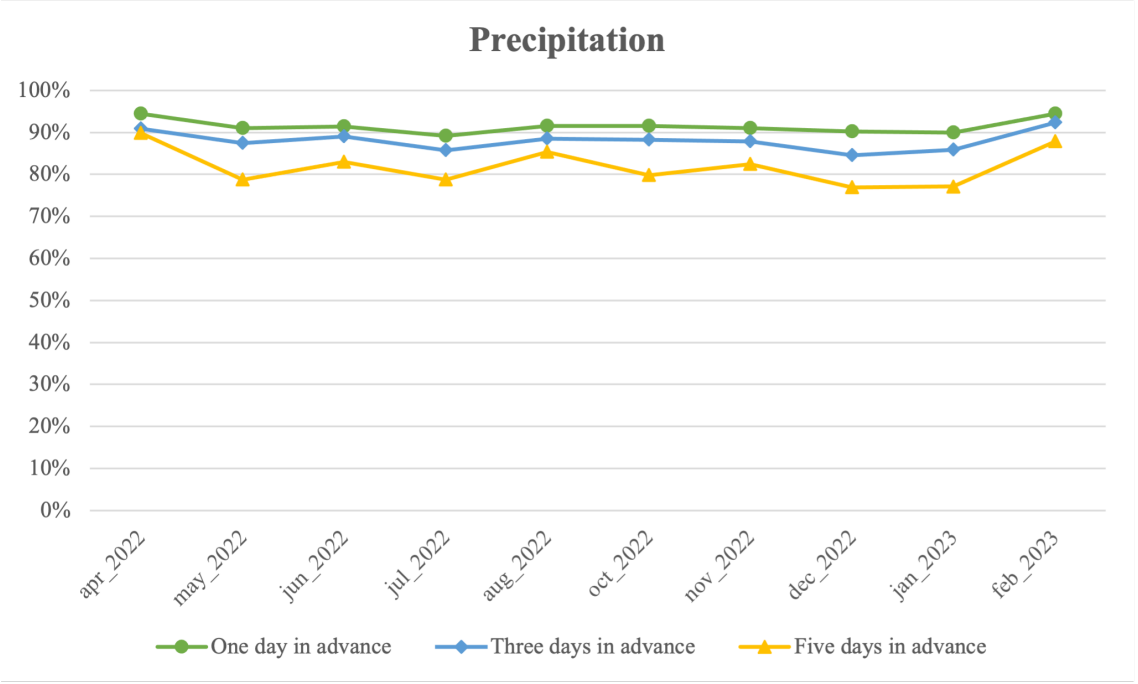

Figure 2.4: Forecast accuracy in Sweden for precipitation by month delivered by the Swedish Meteorological and Hydrological Institute

As the figure illustrated, the precision of precipitation is around 90%. As stated above, this needs to be considered when decisions are taken based on a model's predictions.

3

Methods

The chapter below aims to describe the methodology used in this master thesis, starting with the research strategy and design. Additionally, data collection and analysis are explained for a broader understanding. Lastly, the chosen case studies are depicted to provide findings from previous research related to the subject.

3.1 Research strategy

This study used previous research to understand and draw further conclusions which made it out of inductive nature. Furthermore, this master thesis mainly included qualitative strategies since literature studies about previous machine learning applications alongside complementing interviews were used as the data source for the findings.

3.1.1 Research design

A research design is a template for how the data collection and analysis will be carried out in order to provide a more detailed picture of the research method (Bell et al., 2022). This study was partly of correlational research character since a correlation between bacteria levels and external factors was sought after, and partly of case study character since previous studies were used to study applicable methods for this specific case. The latter included using machine learning to predict bacteria levels in drinking water for a startup company. Additionally, this study was divided into three main steps which are explained in the sections below. Table 3.1 below aims to link data collection, required information, and research questions together.

Research question	Required information	Method
• How should the model be built?	• Training procedure • Evaluation methods	• Literature review
• Which model is suitable for the purpose including requirements from the case company?	• Classification or regression model • Trade-off between interpretability and performance	• Applicability analysis through literature research • Interview with the case company
• Which design choices should be incorporated?	• Data features • Feature selection	• Literature review • Interviews with industry experts

Table 3.1: The research design of the study in terms of research questions

3.1.2 Research steps

This research study was mainly divided into three steps which both reflect the work that was conducted and the existing time frame. Although the steps partly overlapped, the three steps were thought of as a map to evaluate the ongoing work as well as minor deadlines during the course of the project.

1. **Literature Review and Analysis -** In this step, a thorough literature review was conducted in order to establish what kind of models in previous research that had been evaluated. This was also analyzed in order to find successful as well as less successful ones, which provided important considerations when designing the theoretical model. In addition, the literature review worked as an introduction to what questions could be asked during the interviews in the second step.

2. **Interviews -** In this second step, the findings from the first step were used in interviews with three industry-related stakeholders. For instance, an interview was held with the case company to get an understanding of what primary requirements the model needed to be focusing on. Alongside the findings from the first step, the interviews focused on obtaining general industry knowledge for understanding the strengths and challenges the industry possessed.

3. **Development of theoretical machine learning model -** In the third stage, the conclusions from the first and second steps were adopted and implemented into a theoretical model. The aim of this step was to design a template of a model that should be working if it was implemented in future research.

3.2 Data collection

The following sub-section will provide information about the data collection, which consisted of qualitative data in the form of a literature review and interviews. Moreover, the data analysis procedure will be presented which was of thematic analysis character. Hence, during the analysis, the obtained data was structured and interpreted using similarities and differences in the findings which enabled pattern exploration that improved the credibility of the thesis. The methodology used for the data collection and analysis is based on Section 2.2 above which describes the orientations and characteristics of the different forms of data collection and analysis.

3.2.1 Literature review

One of the main data collection sources of this master thesis was a thorough literature review of existing research in the field of supervised machine learning for the prediction of bacteria levels in drinking water. The review was conducted by exploring case studies that were considered to be related to the subject of this thesis. The decisive choices were made by searching through different databases using important keywords such as *Prediction, Bacteria level, Drinking water,* and *E. Coli*. The studies that were identified were then examined to understand the applicability of the studies in this context. The appropriate ones were selected to be a part of the literature review of this study. Although the literature review was arranged to be conducted at the beginning of the study, the literature study was ongoing during the whole project where additional studies were considered.

3.2.1.1 Analysis of literature review

In the literature review, patterns in the case studies were explored by comparing the distinguished aspects from the different studies. The study results highlighted the adequacy of machine learning by finding what previous studies had concluded about the applicability of machine learning in bacteria prediction. In addition, what purposes the studies focused on were considered in order to validate that they were applicable to the case company's aim with this master thesis. Initially, these case studies were essentially described from four perspectives:

- The task the study aimed to investigate
- The machine learning models used to investigate the task
- The conclusion of the study, generally in terms of if the outcome was favorable
- Key predicting factors that were most important for each study result

Moreover, in the result section below, a general evaluation of the accumulated findings from all of the case studies is presented. This part elaborated on several angles to understand the reason behind choices and how well these choices turned out. By comparing and analyzing the different studies, a foundation for the theoretical model

was obtained. As described above, each case study was chosen with respect to the purpose of this master thesis which mitigated the risk of low compatibility with the aim of the case company.

3.2.2 Interviews

This master thesis consisted of three interviews to obtain insights into the current practices and theories in the industry. Two were verbal and held remotely. The last was conducted in writing due to the demands of the interview object. Although written interviews have some crucial limitations, the interview mostly included straightforward developing-related questions. Consequently, the limitations regarding the lack of social cues described in Section 2.2.4 was not considered to be problematic in this specific case. Out of the verbal interviews, the first one was conducted with two representatives from the case company who had good knowledge about the industry, but also which demands the company had on the study. The second interview was conducted with a project leader from Swedish Water Research. Swedish Water Research is a research company, owned by NSVA, Sydvatten, and VA SYD that researches water to develop new efficient solutions to meet the future challenges of the water service industry. Additionally, the interviewee worked within the focal area of biological sustainable drinking water where the focus is on drinking water quality including sensor technology. Previously, the interview object had been working in laboratories that perform tests of water content. However, although the respondent had been exposed to machine learning, it should be mentioned that the person had no expertise in the field. The last interview was conducted through a questionnaire with a research engineer within data science who had six and a half years of experience in applying methods from the data science field to research problems in various areas. More specifically, the interview object had among other activities been the model development expert in projects that specifically investigated the prediction of bacteria in water. Accordingly, the respondent shared a few opinions and experiences from both the actual project and other related subjects.

Moreover, the verbal interviews were of semi-structured character, using a list of questions that made the interview progress effectively without affecting the interview objects answers. Furthermore, as mentioned above, the interviews were held with different industry-specific stakeholders in order to receive a varied and broad perspective of the methods and challenges within the field. As a result, the interview questions were customized for each separate interview object. For example, the case company received questions about what it wanted to accomplish with the model while the industry experts provided answers about the current methods and challenges, see Appendix A. Additionally, the list of questions was also sent to the interview objects in advance to make him or her well-prepared and more confident. During the interviews, the answers were stored electronically and were also sent to the interview objects afterward for inspection and clarification.

The verbal interviews were held remotely using well-established software such as Zoom or Microsoft Teams while the written were held using a questionnaire that

was filled out. The reason was mainly due to the geographical differences between the author and interviewees. The virtual procedure was advantageous both in terms of time management and the ability to record the interviews efficiently.

3.2.2.1 Analysis of interviews

The analysis of the interviews was of thematic character where general themes and patterns were distinguished to find industry aspects and current solutions. Specific and unique answers were considered as less ensured which was emphasized in the result and discussion part of the report. The analysis was conducted by analyzing the answers in terms of the general attitude towards machine learning in the field and how the experts perceived the methods' current implementations as well as their future potential. Lastly, the interviews were also used and analyzed in order to obtain a general intuition of the industry, including significant challenges and advantages that should be incorporated into future solutions.

3.3 Case studies

This section presents case studies related to the subject of machine learning for predicting bacteria in water. The section will mainly consist of a description and summary of the chosen case studies. Moreover, in the result chapter below, the information will be further described and recognized in order to understand how a theoretical model can be developed.

3.3.1 Data-driven approach to predict microbial water quality in drinking water

Sokolova et al. (2022) used different data-driven models, including machine learning, with varying complexity to predict the level of E. Coli bacteria in drinking water. Input variables applied in the model were laboratory measurements of E. Coli and other coliforms, as well as external factors including water temperature, turbidity, precipitation, and water flow in the source. To compare different models, several ordinary benchmarking models were provided such as Exponential Smoothing, ARIMA, and a naive model that predicted the next value to equal the current value. The study found that the more complex models, such as Random Forest and Lasso Regression which included multiple predictors, performed better compared to the benchmark models. Moreover, the most complex models used in this study were the ones that obtained the highest accuracy on the test data although they tended to overfit against the training data. Consequently, the models with moderate complexity, including the Lasso Regression, provided the most generality among the tested models. A result that can be explained by efficient feature selection and interpretable weights. Lastly, the researchers highlighted the impact of external factors on forecasting, where water temperature and precipitation were two of the most prominent ones.

3.3.2 Machine Learning algorithms for the prediction of E. Coli level in agricultural pond water

Stocker et al. (2022) evaluated different machine learning models for predicting E. Coli in agricultural pond waters in order to obtain credible models for the non-linear relationships between water quality and physiochemical parameters. Models that were deployed included Multiple Linear Regression, Stochastic Gradient Boosting Machines, Random Forest, Support Vector Machines, and K Nearest Neighbor algorithms. Additionally, the models were implemented using recursive feature elimination which is a method that removes the least impacting features recursively until the required number of features are used. The review identified that all of the chosen models obtained good results when predicting the level of E. Coli in the pond waters where the Random Forest obtained the best results. However, the authors described that the difference was neglectable. Out of all the models, the Multiple Linear Regression performed worst which indicated the non-linear relationship between E. Coli and the predictors. Furthermore, the authors depicted that the recursive feature elimination was successful and that the most important set of features was similar for all the models. These features included specific conductance, turbidity, temperature, concentrations of chlorophyll, and fluorescent dissolved organic matter.

3.3.3 Prediction of water quality using LSTM deep neural networks

Liu et al. (2019) aimed to understand if a Long Short Term Memory Deep Neural Network was a plausible method for predicting the water quality of drinking water. The certain choice of model was based on the fact that water quality and its input parameters are in the form of a time series. Often, other types of Artificial Neural Networks are considered for forecasting but the authors argue that they are less appropriate for this type of time series prediction since they have no conception of time for the inputs. The implementation offered a basic approach where only single-dimensional inputs were considered. The authors mentioned that they will continue to improve the model using multi-dimensional inputs. However, the procedure established that the developed Long Short Term Memory-model was an advantageous approach to predicting future water quality. It was also found that the predictions were reasonable for a prediction timeline of up to six months. Among the tried input parameters, it was shown that pH and turbidity[1] had a strong positive correlation with the water quality prediction.

3.3.4 Random Forest to predict abundances of bacterial groups in drinking water distribution system

Brester et al. (2020) wanted to understand if a data-driven approach was facilitating when trying to predict the future bacteria content in drinking water distribution systems. To evaluate the idea, a Random Forest model was developed which was trained and tested using different drinking water pipes to obtain a natural split

[1]Turbidity is a measure of relative clarity of a liquid, which was measured at the water source.

between the training and testing data sets. The specific model choice originated from previous studies where the potential of Random Forest in similar forecasting tasks had been established. The study focused on two distinguished groups of predictors, where the first consisted of the material of the pipes and the type of disinfectant used in the pipes. The second group included more general parameters such as temperature, pH, and electric conductivity. Moreover, the forecasting horizon was stated to be seven days in advance. For this time horizon, the model turned out to be inaccurate which was explained by that the current input parameters, mainly from group two, can change heavily during the six days between the current and the forecasted day. Accordingly, the authors were hopeful that the models would be accurate in next-day predictions. Regarding the input factors, the study obtained that pH, temperature, and the level of copper in the water were the most impacting predictors for the study.

3.3.5 Water quality and operation parameters to predict water production by artificial neural network

Zhang et al. (2019) examined water production performance by connecting water quality and operational parameters using a Hybrid Artificial Neural Network (HANN) as well as multi-layer Artificial Neural Network models. Eleven different input variables were used for the study such as turbidity, concentration of ammonium, pH, and energy consumption. Additionally, the authors depicted that the data was carefully scaled and divided into training-, validation- and test sets with a ratio of 3:1:1 in order to validate the model properly. The study concluded that the HANN model performed well for the purpose and utilized the combination of water quality and operational parameters to obtain a credible result. When examining the water quality parameters, it was found that differences in turbidity, differences in ammonium, pH, chemical oxygen demand, as well as residual chlorine of treated water, were the most impacting factors. On the other hand, it was also shown that temperature had only a limited influence on the outcome.

3.3.6 Fecal indicator bacteria prediction in a Norwegian drinking water

Mohammed et al. (2018) used a Zero-Inflated Regression model, a Random Forest and an Adaptive Neuro-Fuzzy Inference System (ANFIS) to predict the level of fecal indicator bacteria in the raw water of a treatment plant in Norway. To examine the bacteria level properly, E. Coli was used as the main bacteria type investigated. The authors clearly mentioned that different Artificial Neural Networks have been used to provide forecasts earlier. However, since an Artificial Neural Network is similar to a black box that can not explain the influence of separate predictor variables, the models in this study were considered as an alternative. As input variables, physicochemical parameters including pH, temperature, electrical conductivity, turbidity, color, and alkalinity were examined. After evaluation, the ANFIS model turned out to be superior in terms of predicting the variations of bacteria level and E. Coli particularly. However, the model also predicted negative values which do not ex-

ist in reality which means that the credibility was diminished. Consequently, the authors suggested that the Zero-Inflated Regression model and the Random Forest were most consistent on the testing data while performing well in the forecasting. If the models were improved, these were proposed as the ones with the highest potential for real-life predictions. Additionally, the study revealed that pH, temperature, turbidity, and electrical conductivity were the most influential factors for the predictions. Lastly, it was found that precipitation was not significantly impacting the forecasting.

3.3.7 A review for machine learning analysis in drinking water treatment

Li et al. (2021) reviewed studies about the use of AI and machine learning in several drinking water treatment processes. Furthermore, the authors compared traditional mathematical models with newer machine learning approaches where the mathematical models often used assumptions or real-life measurements that were not applicable in real-life situations. Additionally, mathematical models did not utilize macro dynamics such as seasonality or pollution as well as non-linear relations. Aspects that frequently contribute to drinking water predictions. In comparison, machine learning models have a strong ability to find non-linear relationships and other dynamics which often contributes to superiority. The authors also meant that AI technology has become even better since it previously has been similar to a black box with low interpretability. An aspect that is being continuously improved with general technical improvements. Moreover, it was mentioned that control systems and costs can be heavily upgraded with the use of machine learning. Consequently, the researchers stated that machine learning is a very efficient tool for predictions regarding drinking water. As mentioned above, this is explained by the models capability of finding nonlinear relationships that mathematical models can not fit. As a result, the authors meant that AI should be implemented in drinking water treatment facilities to help the management make good decisions. Accordingly, pH, temperature, and turbidity were mentioned as important general predictors of water quality. However, machine learning has some ongoing challenges. Firstly, the interpretability needs to be even better in order to reduce the risk of using such systems. Secondly, the approach is relevant for supervised learning but in order to make it more useful, the unsupervised learning implementations need to be improved. Third, the authors mentioned a significant change in the water systems and how the technology can mitigate the risk of failure. With further development within all challenges, the authors were certain that the approach will be useful in the future.

3.3.8 Interpretability versus Accuracy: a comparison of machine learning models to predict E. Coli levels in agricultural water

Weller et al. (2021) aimed to provide a framework that can be used for future studies within the research area. For this purpose, several algorithms, among others Random Forest, Gradient Boosting, and K Nearest Neighbors were evaluated

with and without feature selection, using different error measures such as RMSE and R^2. Furthermore, two distinguished objectives were chosen for the study to be successful. Firstly, the authors wanted to make exhaustive evaluations of the different algorithms and combinations of features existing in the field by developing, assessing, and comparing model types. Secondly, an analysis of the recognized trade-off between interpretation and performance was conducted. A factor that often is important when choosing the most suitable machine learning model. The authors concluded that the choices and configurations that should be chosen depend on the specific situation and area of examination. However, they established that machine learning is helpful for predicting what level, but also when, and where, fecal contamination exists in water sources. The research also highlighted that machine learning is applicable both in terms of relative and absolute measurements of E. Coli. The analysis showed that more complex, less interpretable, models were unable in performing absolute predictions while the more interpretable ones, such as easier Decision Trees and Random Forest, were more accurate. However, the more complex models were although accurate when predicting if the level was above or below a specified baseline. Finally, the authors mentioned that precipitation, air temperature, and turbidity were in general the most important features and meant that these should be included when one tries to develop an accurate model.

3.3.9 Summary of case studies

This subsection includes a brief summary of the case studies above. Important characteristics from each study are presented in Table 3.1 below.

Case-study authors	Section	Task	Applied models	Application conclusion	Specified influential predictors
Sokolova, Ivarsson, Lilliesträm, Speicher, Rydberg Bondelind	3.3.1	To use different data-driven models with varying complexity to predict the level of E. Coli bacteria in drinking water.	Exponential Smoothing, ARIMA, Random Forrest, Lasso Regression, TPOT, and Vector Autoregression	The most complex models obtained the highest accuracy although they tended to overfit against the training data. Consequently, the models with moderate complexity, including the Lasso Regression, provided the most generality among the tested models.	Temperature, Water Intake, Precipitation
Stocker, Pachepsky, Hill	3.3.2	To evaluate different machine learning models for predicting E. Coli in agricultural pond waters.	Multiple Linear Regression, Stochastic Gradient Boosting Machines, Random Forest, Support Vector Machines, and K Nearest Neighbor	All of the chosen models obtained good results when predicting the level of E. Coli in the pond waters where the Random Forest obtained the best results.	Specific Conductance, Turbidity, Temperature, Concentrations of Chlorophyll, Fluorescent Dissolved Organic Matter
Liu, Wang, Sangaiah, Xie, Yin	3.3.3	To understand if a Long Short Term Memory deep neural network is a plausible method for predicting the water quality of drinking water.	Long Short-Term Memory	Long Short-Term Memory-model was an advantageous approach to predicting future water quality for a prediction timeline of up to six months.	Not specified.
Brester, Ryzhikov, Siponen, Jayaprakas, Ikonen, Pitkänen, ... Kolehmainen	3.3.4	To understand if a data-driven approach is facilitating when trying to predict the future bacteria content in drinking water distribution systems.	Random Forest	The model turned out to be inaccurate seven days in advance, mainly explained by the potential change during the six days between the current and the forecasted day. Accordingly, next-day predictions are promising.	pH, Temperature, the level of Copper
Zhang, Gao, Smith, Inial, Liu, Conil, Pan	3.3.5	To examine water production performance by connecting water quality and operational parameters using a machine learning model.	Hybrid Artificial Neural Network (HANN), multilayer Artificial Neural Network models.	The HANN model performed well for the purpose and utilized the combination of water quality and operational parameters to obtain a credible result.	Turbidity, Ammonium, pH, Chemical Oxygen Demand, Residual Chlorine of treated water
Mohammed, Hameed Seidu	3.3.6	To use machine learning models to predict the level of fecal indicator bacteria in the raw water of a treatment plant in Norway.	Zero-Inflated Regression, Random Forest Regression, and Adaptive Neuro-Fuzzy Inference System	Zero-Inflated Regression and Random Forest Regression were obtaining the most reliable results since the ANFIS model wrongly predicted negative values.	pH, Temperature, Turbidity, Electrical Conductivity
Li, Rong, Wang Yu	3.3.7	To review studies about the use of AI and machine learning in several drinking water treatment processes.	Not specified.	Machine learning is a very efficient tool for predictions regarding drinking water due to the models capability of finding nonlinear relationships.	pH, Temperature, and Turbidity
Weller, Love Wiedmann	3.3.8	To provide a framework that can be used for future studies within the research area alongside aspects are important to consider.	A broad range of models with different complexity.	Machine learning is helpful for predicting the level of relative and absolute measurements of E. Coli. Less complex models were accurate in performing regression analysis while the more complex ones could perform classification.	Precipitation, Air Temperature, Turbidity

Table 3.2: Summary of characteristics from relevant case studies within the area of machine learning and drinking water

4

Results

The sections below include the case findings from the interviews with the case company, an industry expert, and a research engineer as well as findings from the literature review.

4.1 Interview results

In this section, three interviews are presented regarding the subject of current procedures in the industry. The first deals with the case company's point of view, the second outlines an industry expert's thoughts and the third depicts the research engineer's experiences on machine learning implementation.

4.1.1 The outcome from the interview with the case company

During the interview, the company representatives from Nocoli depicted that the current industry standard for drinking water bacteria detection in Sweden is called Heterotrophic Plate Counts (HPC method) which also is compulsory due to EU directives. The method is based on taking samples manually with a specific measuring stick which is then sent to a laboratory that provides the kind of bacteria that the water contains. Usually, the measurement is taken on E. Coli bacteria since this bacteria is an indicator of fecal contamination in the water although the bacteria itself is not remarkably dangerous. Often this takes two to seven days depending on if the measuring actors have their own laboratories. However, within real-time detection which is the collective name for measuring without time lags, tools exist that can measure the total amount of bacteria but not the bacterial composition. The interview objects highlighted that this is critical since the level of dangerous bacteria can not be determined.

The representatives outlined several problems with the current situation. Firstly, the duration of measures was mentioned as a result of the HPC method taking several days for manual collection and bacteria cultivation. The water comes to the populations water taps a few hours after it has reached the treatment plant which means that the water that the population drinks has not been approved until several days after consumption. Consequently, if the test indicates fecal bacteria, its difficult to do anything about it since the water already has been consumed. The

interviewees meant that a reason for these lagging measures is due to the old and unmodern infrastructure. To reconstruct the infrastructure, the government needs to invest a huge amount of money and this has been postponed since the Nordics has had good water quality until recently when the quality has become worse. Furthermore, the representatives meant that very complex technology is required to find a new solution which makes it very problematic and time-consuming. To choose a new strategy, all the requirements in terms of time, accuracy, and simplicity must be developed. In combination with tardy public procurement processes, new technology will most likely be developed in the meantime which makes the new strategy even more complex to determine.

Moreover, another problem that was mentioned is finding the source of the bacteria outbreak. Since the test answer is delivered after several days, it is often challenging to find where the source is and thereby how the water should be managed. Additionally, the representatives understood that this will be difficult with a new standard as well. However, it is at this intersection where the interviewees mentioned that their sensors are highly useful. With the sensors, one can say in real time that the outbreak is close to *Sensor 1* which enables better water management. In addition, the representatives explained that water treatment actors are interested in the new technology due to these benefits.

4.1.2 The outcome from the case company's requirements for a machine learning model

During the interview with the case company, the representatives discussed a few aspects of their experience with real-time prediction and the applicability of machine learning in the field. Firstly, the representatives highlighted that the industry is very immature in terms of prediction. Furthermore, it was mentioned that according to their experience, no water treatment plant had implemented machine learning for prediction. The representatives believed that some actors have started developing real-time methods although those actors are very confident about their technology. However, machine learning seems like the most promising approach due to its capability to find complex patterns. The interview objects meant that machine learning is attractive for the case company since it has the potential to improve the analysis method and enables a competing advantage by offering qualitative predictions. An offer that potentially can help the water treatment plants to be more proactive. Since the employees at the case company do not have experience in machine learning development, no strict specifications for the model were demanded. However, some requirements and desires that the company had can be summarized in four categories:

General requirements
The representatives had no demand for the prediction horizon of the forecasting since one day in advance would be very advantageous. However, a percentage for the probability of fecal contamination one or a few days in advance would be sufficient. For example, *Tomorrow it is 80% that the water will be fecally contaminated.*

Trade Off between interpretability and complexity
The interview object stated that interpretability is very important even if the accuracy is reduced. It was mentioned that their customers need to have a clue of why a specific measure is predicted which means that some level of interpretability is required.

Level of detail in the predictions
The interviewees depicted that the water treatment plants use a threshold today for fecal contamination. As a result, an interval rather than a specific number is sufficient. Since the predictions often are uncertain, an interval should be more useful.

Factors to consider as data features
The representatives highlighted a few features that were not mentioned as important in the literature review which still were believed to be interesting. These are:

- The population around the sensor
- Agriculture, fertilizer
- Bedrock
- Watersource
    - Surface water - Higher probability of contamination
    - Groundwater - Lower probability of contamination

4.1.3 The outcome from the interview with the industry expert

The industry expert established that drinking water quality is an extremely important field in todays society that has to overcome a few challenges. These were mentioned to include a changed climate which leads to new versions of contamination as well as a low intrinsic incline to change which can be explained by regulations and security. For example, the companies within the industry have IT infrastructure that is focused on security which means that modern solutions such as machine learning can be too complex for the current systems. Hence, to implement new methods the IT infrastructure needs to be updated. Moreover, the interview object stated that todays competence among several actors is most likely too low for newer implementations. Additionally, it was admitted that the general infrastructure is problematic since the distribution network is lagging in terms of quality measures. Due to the long lead times, the respondent meant that the industry is very reactive when proactive work is required to protect the customers more efficiently.

Regarding the detection methods used today, it was established that every actor needs to conduct the HPC method and this is regulated by the Swedish department Livsmedelsverket. Moreover, the interview object described the HPC method as a very old and inefficient procedure due to low incubation, low accuracy, great varia-

tion, and a lead time that often reaches seven to eight days. In addition, the expert stated that only a small fraction, around a hundredth of a percent, of the bacteria that one can identify in our waters, can be grown using the HPC method. An issue that provides a poor basis for assessing quality. Instead, the interview object highlighted newer methods, such as flow cytometry, ATP measurement, and enzymatic measurement that do not need manual testing and laboratories, as more promising. These were outlined to be faster, the lead time is about 15-30 minutes, but also significantly more accurate than the HPC method. However, it was described that they are not better in all circumstances. For instance, they are often installed in a specific location which reduces flexibility and large-scale measuring. They are also demanding very high maintenance and can be extremely expensive. Additionally, although these methods still can be interesting, the interviewee explained that the HPC tests still are required by regulators which diminishes the newer methods' potential applicability.

Furthermore, the interview objects described that prediction is not used at all in today's systems and that alarms often center around experienced people that signal if the bacteria level curve is following a specific pattern. As a result, today's systems are mentioned to be dependent on special individuals. Consequently, the expert meant that cost-efficient, highly accurate methods are needed in the industry. For example, the companies within the industry are expected to constantly develop new processes to improve their systems although this happens at a low pace. To manage the issue regarding predictions, the industry expert believed that there is a great demand for machine learning. Foremost to be able to manage data structuring and data analysis including new sensor technology and other methods which incorporates real-time measuring with high frequency. Another perspective mentioned that could enable efficient machine learning was a large amount of historical data in the companies. However, this data was declared to be stored in less flexible ways which has resulted in an increased focus on storing it in sufficient and useful locations. Consequently, the interview object meant that machine learning can be very useful for finding correlations in this unexplored data, as long as the data becomes more structured. Hence, new data-driven tools have appeared around visualizing data to get an efficient IT structure although the expert claimed that no current solution that incorporates AI in bacteria-level prediction exists on the market. On the other hand, the interview object highlighted several actors in the start-up stage who focus on machine learning to predict water quality which shows the demand.

In terms of model development, it was stated that more or less everyone needs to be able to manage new systems since this is the current situation. Therefore, interpretability is needed in the models for the customers to be able to improve their operations. However, without enough accuracy, the customers may act based on false information which accordingly can be very expensive. Such a situation means that the customers most likely abandon the tools. Hence, the respondent meant that a suitable trade-off is necessary. Regarding data features, the industry expert mentioned that water temperature at the raw water source, seasons, and events that impact the water production, such as chlorine to clean the water, could be useful.

4.1.4 The outcome from the interview with the research engineer

As mentioned in Section 3.2.2, the research engineer had previously been involved in a related study. The interviewee meant that the project was performed since the pollution of water is an increasing problem globally and securing drinking water is of high importance. Additionally, it was clear that data-driven models had been applied successfully before in different scenarios involving microbial water quality and that the setting that existed for the specific study seemed to fit a similar approach. Although the interview object meant that the project turned out to be successful since it showcased what type of models one can apply as well as what type of predictors could be important, some issues still existed. The respondent described that it seemed like there still is a gap between the performance of the models and human expertise. On the contrary, it was clear that the models could be useful as a supporting tool for decision making which was considered to be a promising start for the research.

For the different implementations of the particular project, the interviewee explained that if one model would be implemented in real-life practice, it would probably be one of the simpler ones. This since real-life circumstances often require good explainability and the ability to display how the predictions were made. Additionally, the person highlighted that more complex models did not improve performance in the study and that these models were more prone to overfit. Another described reason for using less complex models was the lack of data. If it was another scenario with more data available, the respondent meant that more complex models could increase performance. In such a case, the developer needs to consider if it is worth the extra performance at the cost of less explainability. A model that the interview object highlighted as a valid middle path was Random Forest since the model consists of something between linear models and black box-like deep learning. Moreover, it was clarified that if a model does not generalize to the test data, one should be careful about interpreting the feature importance since it is based on the training data when fitting the model. However, depending on the data, more complex models than Random Forest were also mentioned to have potential. Examples could be Gradient Boosting and Neural Networks although the respondent once again was clear about the lack of interpretability in these models.

Regarding data features, the research engineer stated that a lot of different features could be interesting. The broad range was also discussed within the development team. However, some were rejected because no useful data sets could be obtained and others were dismissed based on expertise about water systems. Furthermore, the interviewee concluded with a recommendation to start with a simple model and gradually add complexity. For instance, to start with a baseline model based on only a few features including a good evaluation strategy that later could be used for improved complexity. Additionally, since it was difficult to obtain useful data sets, it was also recommended to only start by considering features where the data set was clearly available.

4.2 Machine learning findings for predicting bacteria level with machine learning

In Section 3.3 above, eight different case studies are presented that were recognized as similar to the task of this master thesis. In the subsections below, similarities and differences between the previous work are summarized from distinct perspectives. This information was in the next step critical for developing a theoretical machine learning model.

4.2.1 Model selection

In the case studies, the researchers chose different models depending on their purposes. The studies varied between using a lot of models in order to try how different complexity impacted the performance and choosing one specific algorithm which was carefully tuned to examine how well the predictions were. A common denominator in the case studies was that different Neural Networks and Deep Learning algorithms were acknowledged to be accurate for predictions. For example, Zhang et al. (2019) used an improved Artificial Neural Network due to the fact that these are efficient when a large amount of data is available. Something that the drinking water treatment station that was active in the study possessed. Moreover, the study showed that an Artificial Neural Network with a generic algorithm improvement outperformed a Support Vector Machine which highlights the potential in the Neural Networks for predictions of bacteria levels. Another example of a Neural Network was Li et al. (2021) who selected a Long Short-Term Memory algorithm to evaluate the prediction performance. The authors described the algorithm as efficient when dealing with long-term dependencies since it efficiently disregards useless information. In addition, since the bacteria level was a continuous time series, the author emphasized the algorithms previous performances as the most prominent factor for the model choice.

However, due to the Neural Networks' characteristics regarding the black box-like functionality, it was common to examine other models as well in order to obtain more interpretability. For instance, Sokolova et al. (2022) aimed to investigate how complexity impacted performance. For this, the authors chose univariate approaches such as Exponential Smoothing and the statistical-based Autoregressive Integrated Moving Average as basis models. These models were compared to multivariate models including Lasso Regression which is an extension of Linear Regression using regularisation and Random Forest to find differences in performance. The authors highlighted the Random Forests ability to understand the importance of data features which distinguishes itself from other similarly complex models. Furthermore, Stocker et al. (2022) also aimed to investigate complexity by comparing the Stochastic Gradient Boosting algorithm which in this case used an ensemble of Decision Trees, with a K Nearest Neighbor, a Support Vector Machine, and a Random Forest. A different approach that also involved quite a few models was Weller et al. (2021) who intended to evaluate how interpretability and performance related to each other. To do so, several models with different complexity were compared.

Firstly, as a baseline, the authors developed an eight-log-linear and a featureless regression model. Additionally, a tree-based approach using Decision Trees and Random Forests, in addition to instance-based algorithms like K Nearest Neighbor and Support Vector Machine was developed. Lastly, the authors tried Neural Networks, penalized algorithms as well as Multivariate Adaptive Regression Splines that utilized old features to construct new ones used for the forecasting.

A few studies had other approaches such as Brester et al. (2020) who wanted to understand if the less complex Decision Trees could make accurate predictions. However, the authors highlighted that Decision Trees often overfit the training data which made them consider Random Forest as a promising substitute. Additionally, the authors mentioned that Random Forests had been successful in similar predictions previously. Another study was Mohammed et al. (2018) who meant that Artificial Neural Networks were too much of a black box and wanted to find models that were able to describe the relation and importance of input data and output data. As a result, Random Forest and Adaptive Neuro-fuzzy Inference System (ANFIS) were used. A last approach was Li et al. (2021) who reviewed several previous studies to get an understanding of which models that had performed best. The authors concluded that Artificial Neural Networks, Deep Learning, Support Vector Machines, and Random Forests were commonly used depending on the approach. If the aim was to reach maximal accuracy, Artificial Neural Networks, Deep Learning, and Support Vector Machines were often exploited while Random Forests were developed if interpretability was essential due to their capability of disclosing important features.

4.2.2 Data

When it comes to data, the case studies exhibited the importance of finding reliable data. All of the studies used water data obtained from different third parties where sample data had been collected. Depending on the subject of the study, the data originated from different sources although the data had similar patterns in terms of duration and input factors. For example Sokolova et al. (2022), Liu et al. (2019) and Mohammed et al. (2018) used water data provided by the local municipality, taken from a river that provided the area with drinking water. Other examples are Stocker et al. (2022) who used two separate ponds as sample location and Weller et al. (2021) who used sixty-eight streams that were sampled more rarely. The water quality consisted in all studies of measures of E. Coli since the bacteria is an indicator of fecal contamination. Moreover, the frequency of sampling varied between daily (Sokolova et al., 2022) and biweekly tests (Stocker et al., 2022), with the exception of Weller et al. (2021) who only used each stream two or three times each. Additionally, Weller et al. (2021) and Stocker et al. (2022) solely had data points from the summer while the others used sets from all of the annual seasons which enhanced findings of data irregularities due to seasonality. The duration of the data sets was between a few months and several years where Sokolova et al. (2022) and Mohammed et al. (2018) used data from seven years while Brester et al. (2020) only exploited data from four months. Usually, the additional data mainly consisted of weather-related factors such as turbidity, precipitation, and temperature which were delivered by

the local weather station. Furthermore, the separate studies occasionally had other inputs which among others included pH, electric conductivity, and residual chlorine obtained from the same or other actors as the E. Coli data.

4.2.3 Data preprocessing

In essentially all of the studies, the authors highlighted how missing data was managed. In general, two different approaches existed. The first consisted of removing the data rows including missing values. An example of this was Stocker et al. (2022). The second and more common approach was data imputation by using some kind of mathematical model that was based on previous data to simulate the missing values. Brester et al. (2020) took advantage of a monotone piece-wise cubic interpolation method that produced continuous functions which enabled the generation of additional sample points. A different approach was Liu et al. (2019) who utilized linear imputation and mean imputation depending on the data to fill the gaps.

Additionally, all the developers also split the data into training- and test sets of which a few studies also used a separate validation set. The split ratio between the different sets varied between the studies but the size of test sets was between 20-40 % where Sokolova et al. (2022) used 40% and Zhang et al. (2019) chose 20%. A reason for the latter relatively small size could be described by the fact that an additional 20% was earmarked as a validation set. A distinction between the procedures was whether the splits were made randomly or by time period. For example, Liu et al. (2019) used a periodic division while Mohammed et al. (2018) split randomly where the latter wanted to avoid structural changes in the data. In addition, several studies standardized the variables between 0 and 1 since it often improves accuracy (Liu et al., 2019).

4.2.4 Feature selection

Depending on the purpose of different studies, feature selection was incorporated in the development of the models applied. As an example where feature selection was not conducted, Weller et al. (2021) did not use the approach since the aim was to compare algorithms and from this, decide whether feature selection was needed. Others, passively conducted feature selection by having algorithms that automatically incorporate some kind of feature selection. Examples were Sokolova et al. (2022) and Brester et al. (2020) who respectively developed a Lasso Regression as well as Random Forest which had feature selection as an integrated part.

Another approach that was used in different versions was to make a feature analysis before the training had begun to understand which parameters that were important. An example that both Liu et al. (2019) and Mohammed et al. (2018) used was a Pearson correlation coefficient matrix. The method is based on quantifying the linear relations amongst the data variables and using this to find importance between the input variables (Mohammed et al., 2018). Mohammed et al. (2018) also added an out-of-bag supplement that was suitable for the Random Forest that the

researchers had implemented. The authors described out-of-bag as convenient for a Random Forest with smaller data sets in particular. This is since only a subset of the Decision Trees in the Random Forest is used to determine the result at each time which helps prevent the model from overfitting. Zhang et al. (2019) have another procedure where active feature selection was applied before model implementation. In this case, a principal component analysis was carried out to avoid redundancy by managing several closely correlated parameters with a similarly important impact on the output. From this, the most important variables were chosen and then incorporated into the final model implementation. Furthermore, a separate instance was Stocker et al. (2022) who developed a recursive feature elimination algorithm imposed on a Random Forest. The algorithm calculated the feature importance after each iteration by understanding the features share of the residual error and removed the least important feature accordingly in order to obtain a useful final configuration.

4.2.5 Error measure

In the theory chapter above, Jierula et al. (2021) described that it was advantageous to use several error measures since the analysis will be more comprehensive due to the characteristics of each measure. However, depending on the applied models in each study, different measures were more or less appropriate. For studies where different models were compared, such as Sokolova et al. (2022) and Stocker et al. (2022), several error measures were chosen due to their intrinsic attributes. Sokolova et al. (2022) used MAE, RMSE, R^2 and a kind of MAPE called Symmetrical Mean Absolute Percentage Error which allows observations that are zero. Equally, Stocker et al. (2022) included MAE, RSME, and R^2, in standard and normalized versions, to obtain errors both in terms of percentage and absolute values. In the remaining studies, some kind of standard MSE or RMSE was applied.

4.2.6 Performance

In terms of performance, all of the case studies found that machine learning in general was an adequate tool for making predictions regarding water quality which includes the prediction of fecal bacteria in drinking water. An example of this is Liu et al. (2019) who found that the implemented Long Short-Term Memory Deep Neural Network was very promising for predictions of drinking-water quality for up to six months and recommended the future adoption of the algorithm. Another instance of a well-performing Neural Network was the Hybrid Artificial Neural Network that used water quality as an input parameter to predict future water production (Zhang et al., 2019). Moreover, Weller et al. (2021) highlighted that machine learning models were efficient for predicting both relative and absolute levels of fecal contamination in terms of when, where, and at what level. However, the authors also found that Neural Networks performed worse than other models in the prediction of actual values while Neural Networks were very competent for relative predictions of E. Coli.

Among the models that were not Neural Networks, Random Forests stood out as a prominent model. For example, Sokolova et al. (2022) who compared several models found that Random Forest achieved the highest performance score even if it showed a tendency of overfitting. The authors also showed that all models were superior to the naive baseline model which highlighted machine learning's applicability. This was further emphasized by Li et al. (2021) who stated that machine learning is efficient for complex problems including drinking water treatment which makes several algorithms suitable for predictions. Although machine learning is efficient, Brester et al. (2020) obtained that a prediction horizon of seven days is the maximum when using Random Forests.

Mohammed et al. (2018) observed that Random Forest was the most applicable among the models for the purpose of fecal contamination prediction in drinking water. Additionally, Stocker et al. (2022) discovered that Random Forest as well as Stochastic Gradient Boosting performed better than Support Vector Machine and K Nearest Neighbor even if the advantage was not too significant. The authors also highlighted that different error measures can contribute to the result where they found that the relative error was not the same for all models. Consequently, the study also suggested using and analyzing multiple error measures.

4.2.7 Data feature importance

In all of the case studies, the most important data features were pointed out. A common denominator from all the studies was that temperature had a significant impact on the result. Although some studies exploited water temperature and other air temperatures, the feature turned out to have a great impact which highlighted the consistent importance of temperature for bacteria level predictions. Additionally, several studies, such as Zhang et al. (2019), Li et al. (2021), and Mohammed et al. (2018), also obtained that pH and turbidity were particularly decisive. Moreover, depending on what different data features were incorporated in each study respectively, different factors turned out to be conclusive. One example was Sokolova et al. (2022) who found microbial concentrations upstream and water intake to be important. In addition, Brester et al. (2020) received concentrations of copper when Stocker et al. (2022) obtained concentrations of chlorophyll and specific conductance as especially impacting. Furthermore, Liu et al. (2019) described that several of the data features had internal correlations which meant that some could be redundant, such as turbidity and precipitation. An aspect that efficient feature selection methods could diminish.

5

Discussion and Recommendation

This chapter discusses the theory and results from the data sources analytically to properly answer the research question. The first part answers whether machine learning is adequate and analyses the result section to understand and nuance the current situation, both in terms of current prediction challenges and machine learning's applicability in the field. Next, data requirements to implement the model properly are considered, including water data, data features, and other perspectives. Moreover, the other parts of the research question are answered by a recommendation including a theoretical model which evaluates model selection, and implementation decisions related to the suggested model. Lastly, the used research method is discussed in terms of ethics and quality to improve the credibility of this study.

5.1 Machine learning's potential in bacteria prediction levels in drinking water

One of the main questions in the report is to evaluate if machine learning can be useful for predictions of bacteria levels in drinking water. Although it is difficult to establish if the approach can be implemented successfully, one can through this report identify if machine learning models are promising for implementation. To start off, it can be purposeful to compare its potential with today's most common but also required method. In comparison to the HPC method, machine learning is advantageous from the time perspective since the approach enables insights beforehand. Using the current method, the results are distributed two to eight days after the test according to actors in the industry. Additionally, the industry expert outlined that the method has low incubation, low accuracy, great variation, and that only a small fraction of the bacteria in our waters, can be grown using the HPC method. Consequently, machine learning should be able to enhance proactive management of water using forecasts which would improve control and safety since today's approach is based on reactive handling after the water has been consumed.

The literature study reveals the belief in machine learning in the field. The result was based on the models' ability on finding high-level patterns as well as their capability on beating easier naive methods. However, the studies from the literature review generally aimed at finding how well the performance of the models was. Another, possibly even more complex issue is to understand if the approach can be

implemented properly and which potential obstacles that can arise. This is problematic since the general machine learning knowledge within the industry is low, which means that the sector might not be mature enough. Additionally, the practical implementation aspect is less evaluated and can be equally problematic as the initial evaluation of machine learning adequacy. Aspects that also were emphasized by the industry expert who highlighted both the lack of knowledge but also the old IT infrastructure which can bring a certain problem when new systems are to be implemented. As a result, although the literature review highlighted the potential of machine learning, it can not be determined that it is an adequate method to use. Nevertheless, since practical implementability is not the scope of this master thesis, mainly due to time limitations and lack of data, the literature review was concludingly highlighting the potential for machine learning in this field.

Moreover, as mentioned in the result section, governments including public procurement play a critical role in the adoption of new technology for measuring bacteria levels. Reasonably are the decision-makers more open to machine learning and other data-driven approaches now than a few years ago. However, because water is a crucial element in the population's well-being, alongside the amount of money needed for a new system to be incorporated, a lot of uncertainty needs to be reduced. To do so, besides continuous conversations, it seems very important to test the models properly as well as develop them accordingly to obtain credibility. In addition, the industry expert highlighted that large amounts of data that can be used for machine learning development already exist in companies within the industry. However, since the data has not been used due to simple human analysis in the companies, it is often too hidden and unstructured to use. Due to the new application areas, new tools that structure data have occurred although no tools for prediction are offered. For a machine learning implementation to be successful, it would be beneficial if these data structuring tools could work together with prediction models in order to optimize the process. Accordingly, different actors have stated that the industry is enthusiastic about machine learning-based prediction models. As a result, the scope of this report appears to be interesting since a theoretical model is a good start, which makes it easier to implement properly when adequate data is available.

5.2 Data requirements for a successful implementation

By using the information provided in the result section, both from the case studies as well as the interviews, it is clear that a few elements are needed to enable an efficient system. These include different kind of data that needs to be available and sufficient in order to implement a model properly. This section has been divided into main data, weather data, and other data features since the two first kinds are compulsory and straightforward while the third has less distinct implications.

5.2.1 Main data

As the main data source, the model needs measurements of the historical bacteria level at specific points in the water distribution system. In the initial stage, samples from a few measurement points can be enough. However, in the long term, to make the model totally optimized, measurements from each specific location might be required since different locations probably have different characteristics. The data sets used in the case studies were most common from local laboratories that provided the study with data. The measurements were varying but typically, the data contained information on how much E.Coli existed per 100 milliliters. Depending on the data, it seemed to be different how the models were trained regarding if a threshold was used as *Low* and *High* or if the data was used in its original shape. For example, Mohammed et al. (2018) had that $200 \leq$ colonies per 100 milliliters were considered contaminated. A potential problem with using a threshold in the training phase is that the model will be trained using such labels, and consequently will conduct predictions accordingly. Hence, since a water treatment company most likely acts differently depending on the magnitude of the contamination, such a model can be defective in some circumstances. On the other hand, the aim of Mohammed et al. (2018) was a comparative assessment of different models to predict bacteria level which made the used labeling reasonable. However, for the theoretical model in this study, the actual values seem rational in the initial stage.

Another influential consideration is whether primarily E.Coli should be investigated or if other forms of contamination also should be included. Firstly, this is highly dependent on if other data sets are available since it often is crucial to find applicable data. Secondly, since E.Coli is a credible indicator for other forms of contamination the bacteria seems reasonable to primarily consider although it by itself is not necessarily dangerous. Additionally, since this is one of the forms that can be grown in the laboratories during the HPC method, the data should be possible to obtain in several different locations. A perspective that should be positively recognized due to scaleability although only a few locations might be utilized in the beginning.

5.2.2 Weather data

The case studies highlighted that different weather factors were the features that impacted the machine learning models most which means that such data are vital for predictions in this field. Foremost, temperature, precipitation, and turbidity were mentioned as important factors which thereby are crucial to receive for the model development. Moreover, the weather data needed for implementation should be divided into two distinct categories where each is important in different parts of the prediction procedure. The first category consists of historical data of the measurement points above which consequently will be used to train the model. Such data is often available in the open data repository from the local weather stations, in this case, SMHI. For example, Sokolova et al. (2022) used this data source in their development which emphasizes adequacy.

The second category involves data that will be used in the prediction part of the model. To be able to make predictions of the bacteria levels using this approach, forecasts of other data points such as weather factors are needed to be used as input. Conveniently enough, there are plenty of these types of forecasts from the local weather stations. For the case company, SMHI also delivers access to their forecasts using their API. These forecasts contain the measures above as well as quite a few other measures that possibly can be interesting for the result. Consequently, by implementing this kind of forecast, a model would conceivably be able to make proper predictions. However, one perspective that carefully needs to be considered if implementing this kind of data is the inherent risk that comes from using such forecasts. Figure 5.1 shows SMHI's precision for precipitation represented by the average of the forecasts from each month, also illustrated in Figure 2.4 (SMHI, 2023). A forecast is considered to be *true* if the residual between the predicted amount and the actual amount is less than three millimeters per three hours, or in a mathematical formula: $residual \leq 3mm/3h$.

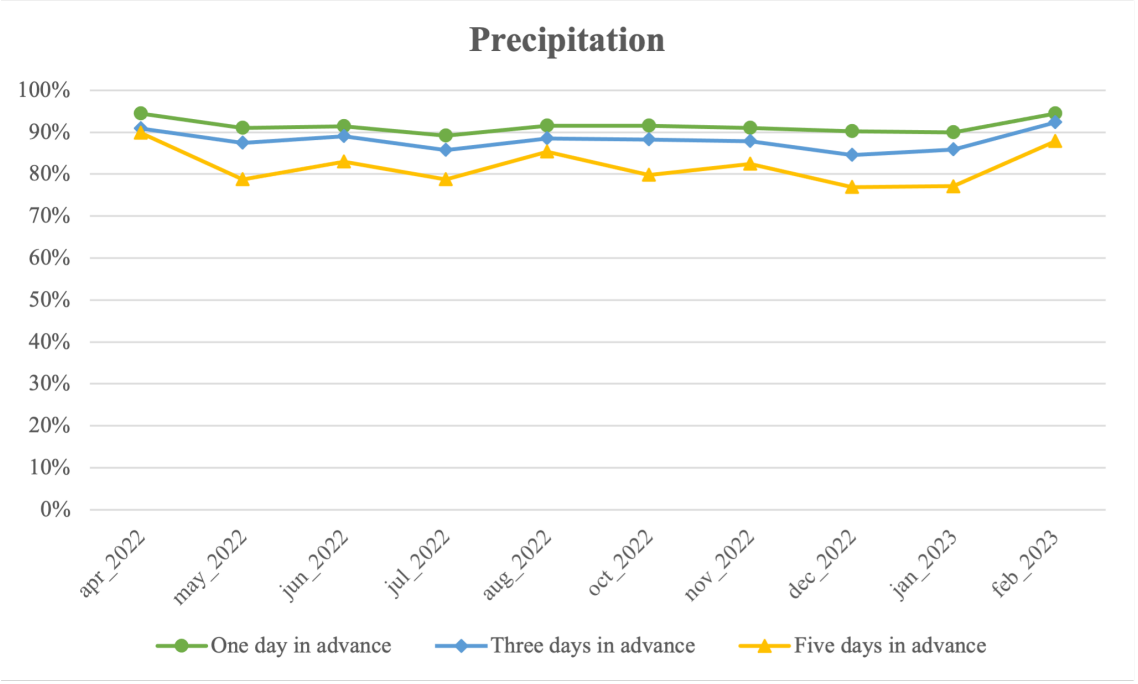

Figure 5.1: Forecast accuracy in Sweden for precipitation by month delivered by the Swedish Meteorological and Hydrological Institute

Similarly, a chart regarding the institution's accuracy for air temperature is shown in Figure 5.2 (SMHI, 2023). As above, the graph demonstrates the average of the forecasts from each month. The definition of *true* for this measure corresponds to if the residual is less than two degrees Celsius, or in a mathematical formula: $residual \leq \pm 2°C$.

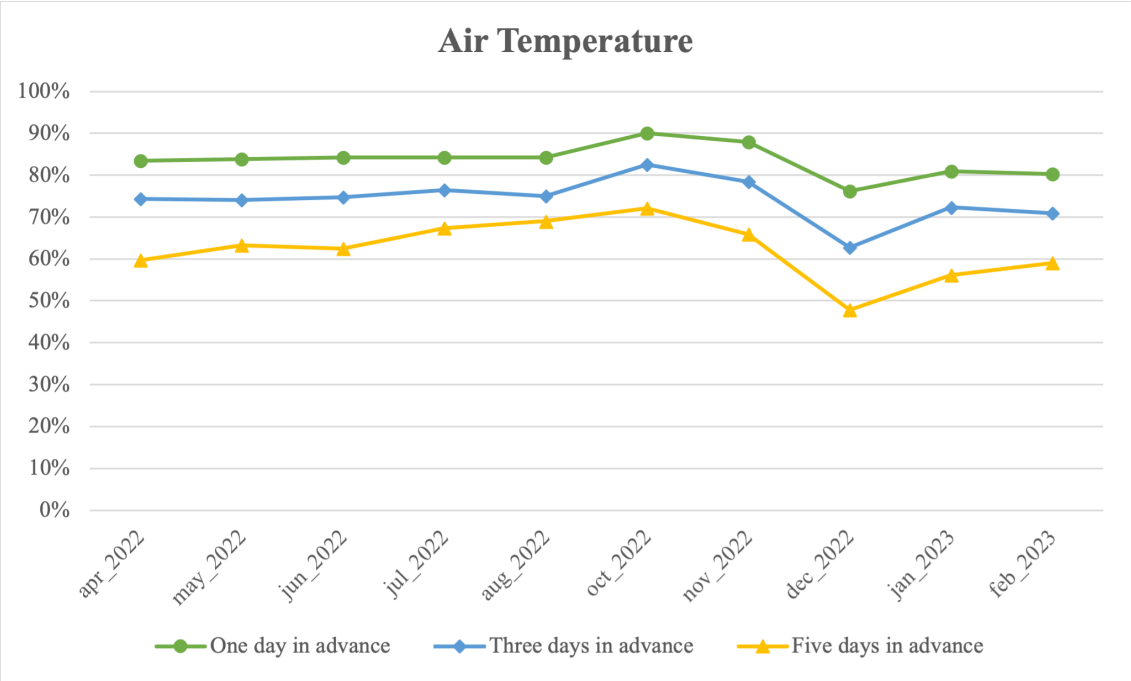

Figure 5.2: Forecast accuracy in Sweden for air temperature by month delivered by the Swedish Meteorological and Hydrological Institute

The charts clarify that the forecasts regarding precipitation are in general more accurate than the forecasts for air temperature. Additionally, the forecasts for the next day are logically more precise than the forecasts for several days ahead. Firstly, that precipitation is more accurate than air temperature could be better than the opposite for this kind of modeling since small changes in precipitation reasonably impact bacteria levels more than small errors in air temperature. Secondly, the case company currently desired predictions for only one day in advance since this would be a proper first step. For instance, Brester et al. (2020) found the difficulty of making predictions for seven days ahead, and in combination with the information provided by the charts, one-day ahead forecasts seem reasonable.

5.2.3 Other Data Features

When evaluating the opinions provided in the result section, it is clear that other factors than weather should be considered for the models to be optimized. Examples that were mentioned in the case studies included concentrations of copper, concentrations of chlorophyll, and specific conductance. Moreover, the industry expert introduced that events impacting water production, such as the usage of chlorine to clean the water, could be useful if considered in the model. In addition, during the interview with the case company, the representatives took a broader perspective and discussed factors that exist around each sensor, where population, fertilizers, bedrock, and type of water source were mentioned. Although sensors might be within a specific area, such as Stockholm, every sensor location will have different characteristics which make the suggestions from the representatives plausible.

When inspecting the features above, a distinction should be made between highly changeable features and less changeable ones. Examples could be bedrock and population, which do not change at all or at a slow pace, compared to concentrations of chlorophyll as well as chlorine level for cleaning the water which should be inclined to change fast. In a totally optimized model, these should be handled differently since the more varying parameters should need to be updated more often while the more fixed values only need to be modified rarely. For the more fixed parameters, it could be appropriate to create a database with the most prominent aspects for each location that could be used for the model to train on. A simplified illustration of such a database is displayed in Table 5.1.

Location	**Depth**	**Bedrock**	**Water source**	**Population Category**	**Agriculture area**
Stockholm	32	Limestone	Ground Water	100 000	Yes
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...

Table 5.1: Template for a database to train the machine learning models

Additionally, a simplification that could be tolerable, especially in the initial phase of the model development, is to add a few columns with more changing parameters. An example could be to have a column with concentrations of chlorophyll and then use a threshold value with the specification *High*, *Moderate* or *Low* depending on the most common value of concentrations of chlorophyll at that specific location. Although this might be a simplification, it could be useful in an initial case which makes it reasonable in this context.

To handle the events that occur by human interaction, such as chlorine, the model should have an input parameter that could be changed manually or automatically when the handlers act in a certain way. The most efficient way of doing this is probably dependent on the customer since different actors use different operators. Consequently, it will not be considered for the theoretical model in the next section.

5.3 Theoretical machine learning model

In the section below, a recommendation about how a theoretical model could be developed is presented based on the findings from the previous parts. Consequently, the sections below answer the research questions regarding how the model should be built, which model that is suitable for the purpose including requirements from the case company, and which design choices that should be incorporated. Additionally, it should be mentioned that this recommendation assumes that data will easily be available in the future and that hypothetical machine learning engineers can retrieve the data conveniently. To begin with, model selection will be conducted where several models are considered and one is proposed. In the next phase, different architectural choices are recommended including the splitting of data sets, feature selection, and error metrics. Finally, in Appendix B a possible implementation is

sketched which includes a few different possible models with a common algorithm. In addition, code blocks and sketches will be in the programming language Python since it is the language that the author of this report has practiced most.

5.3.1 Model selection

To select the most suitable model for these circumstances where the aim is to predict bacteria levels in the water, several factors can be considered. Although some research found Neural Networks very accurate, these types of models seem not to be appreciated by the case company or the industry expert since high interpretability among the models was considered important. Therefore, models that are found adequate in terms of interpretability are more interesting. However, the industry expert also highlighted that without enough accuracy, the models would not be used by customers. This means that a model that belongs between easier linear models and complex models is the best choice.

Accordingly, the algorithm that is recommended is Random Forest since the model often obtains a solid performance alongside comparably interpretable results. Additionally, the algorithm was often applied in case studies where it generally showed reliable predictions. Another perspective is that the research engineer with experience in the field recommended the algorithm as useful and meant that it could have great potential. Consequently, it seems like a reasonable choice to begin with, which can then be extended if an implementation turns out well.

5.3.2 Data

The data for the model needs to be trustworthy and follow the data requirements in Section 5.2 above. Generally, the more data is available, the better can the model be since it will have more data to train on. What is important is that the data is well structured and does not contain missing values, which will be treated in the section about data preprocessing below. It is recommended that the data is in tabular form which will simplify the applicability for a Random Forest and that at least the main data and the weather data described in Section 5.2.1 and 5.2.2 respectively are obtained. Additionally, it is recommended to start the implementation only using this data and then add other features if available.

5.3.3 Data preprocessing

When it comes to data preprocessing, one of the first aspects to consider is how the split between training, validation, and test sets should be. Using the case studies as a base, it seems reasonable to allocate 60% as the training set, 20% should be used for validation, and the last 20% should be designated for testing. This is because the majority of the studies had exactly 60% as training, while the other split varied. In addition, it is recommended to use a random split for the separation since this prevents structural changes in the data due to new bacteria characteristics.

Moreover, depending on the amount of data available, one needs to decide whether to remove the rows with missing values or if it is preferred to use a mathematical model to simulate the missing values. In the assumption that there is a sufficient amount of data available, the first proposal is recommended in the initial step. This is partly because the research engineer highlighted that it was valuable to start simple and then increase the complexity and partly due to the potential error in data values when choosing a representative simulation method. If one has a proper mathematical method, the second option can be considered as well. Whichever alternative is found suitable, it is important to remember that the main objective is to have rows without missing values in order to make the recommended models work properly.

5.3.4 Feature selection

Feature selection has clearly been pointed out to be an important part of machine learning implementation although it needs to be managed carefully since selection methods sometimes can eliminate important predictors. However, if feature selection is applied properly, redundant information will be removed which makes it advantageous. The recommendation for handling this issue is either to use a recursive feature elimination algorithm imposed on a Random Forest such as Stocker et al. (2022) or have an out-of-bag supplement on the Random Forest that selects features during the model fitting in order to mitigate overfitting. Since the recursive algorithm calculates the feature importance after each iteration and then removes the least important feature accordingly, the selection process will proceed satisfactorily and the purpose will presumably be obtained.

5.3.5 Error measures

Regarding the error measure, it was previously mentioned that a combination of different ones is preferred since different intrinsic characteristics are active for separate measures. The majority of the case studies used some kind of standard MSE or RMSE which makes these recommended for this context in an initial step. Moreover, these can be elaborated upon in further development.

5.4 Method discussion

Since this thesis was written in association with a case company, it was natural that the company's specific aims and challenges were considered. This means that the study can be perceived as less generalizable than desirable for a research context in general. As the time frame for the project was relatively limited and implementability was requested by the case company, generalization is a further step and not an aspect this project could be focusing on heavily.

In terms of the data collected, this needs to be seen as reliable since several previous widely different studies were processed. An aspect that highlighted different perspectives where similarities and successful methods were found and applied. In

the initial stage of this project, data sets were desired to be available. However, due to several reasons, this could not be delivered to this study. In such a case, the theoretical model would have been tried which would have been valuable both for the author of this thesis, the case company, and the research quality of this study. Since this was not the case, using previous studies to develop a hypothetical model was an efficient middle ground to give the company a considerably high amount of what they wanted from the project. Additionally, by interviewing the case company and other experts, a nuanced image of the methods and challenges of today could be obtained, something that improved the credibility.

Using semi-structured interviews was an advantageous form for obtaining good answers from the interview objects. Since the format enabled the interpretation of questions and answers that the interviewees thought suited best as an answer, opinions about current systems and challenges could be retrieved which helped develop and validate the theoretical model. In addition, the risk of forced answers was reduced, which is good in terms of research quality and research ethics.

The parts below consist of important aspects regarding the quality and ethics of the study. Aspects that were carefully considered during the research process.

5.4.1 Research quality

To ensure the research quality in an academic study, Halldórsson and Aastrup (2003) suggests trustworthiness as a reliable measure. The measure consists of four dimensions; credibility, transferability, dependability, and confirmability.

Credibility examines the correctness as well as the accuracy of the discoveries during the research (Halldórsson & Aastrup, 2003). This was conducted by a combination of finding patterns from the case studies and the interview answers. In addition, verification of the thesis was performed through an extensive review of the completed thesis by representatives from the case company.

The second dimension to examine is transferability which represents how general the findings are, which means how well they can be applied in different circumstances (Halldórsson & Aastrup, 2003). Since the objectives of the case company mainly were considered, the generality can be seen as limited. However, due to additional expert interviews, some level of transferability was projected to exist in the study.

Dependability is the next dimension to be evaluated. The term regards the data stability over time as well as if the study is conducted in an appropriate independent manner (Halldórsson & Aastrup, 2003). To prove dependability in this study, the data collection procedures are described in detail in Section 3.2.

Finally, the last component of trustworthiness is confirmability and it regards potential biases in the results from the authors (Halldórsson & Aastrup, 2003). In this study, confirmability was less favorable since it is out of qualitative nature which is based on interpretation from the author. However, as a mitigating action, the case company reviewed the interpretations to diminish biased results.

5.4.2 Research ethics

Denscombe (2010) mentions that the researchers need to conduct the tasks with scientific integrity which will not endanger participants interest, as well as phycological or personal harm. Moreover, Bell et al. (2022) mention four different ethical parts that are needed to respect when conducting correct research: avoidance of harm, informed consent, privacy, and preventing deception.

Avoidance of harm is preventing that interview objects and other stakeholders feel stressed or harmed during the research procedure (Bell et al., 2022). In this thesis, this was confirmed by circumventing stressful subjects during the interview process as well as having an ongoing dialogue with the participants during the project.

The second dimension, informed consent, regards that the involved objects are truthfully informed about the purpose so they can decide whether to participate or not (Bell et al., 2022). This was established by letting the participants know and agree with the purpose before the participants involvement began.

Privacy highlights activities that have been conducted in order to protect the privacy of the involved stakeholders during the project (Bell et al., 2022). During this thesis, each interview object chose whether or not they wanted to be anonymized in order to keep their privacy. Additionally, the results and conclusions were described in a general manner, in order to keep a high level of privacy.

The last dimension concerns the prevention of deception in the study (Bell et al., 2022). This was performed by letting the material be constantly reviewed by representatives from the case company as well as the supervising university.

6

Conclusion

The aim of this thesis was to understand if machine learning could be used to predict bacteria content in water and how such a model should be architected. The purpose is of great relevance which was highlighted in several interviews during this project. The current methods for bacteria detection in drinking water have massive challenges where the long lead time is the most prominent. As a result, the water can be contaminated and consumed without anyone knowing until seven days after. To prevent this from happening, a credible forecasting approach is desired and machine learning has shown strong potential in the field. In order to implement such a model properly, one should start simple and then add complexity if the approach turns out to be useful. Also except choosing a promising model, which in this thesis was recommended to be Random Forest, data needs to be available in a structured way. The data that is needed in an initial step is considered to be bacterial data in water as well as weather data. Additionally, if other data features potentially impact the bacteria level, these should be incorporated as well.

Two major challenges with this approach are firstly that the industry might not be prone and ready to be changed since the IT infrastructure and knowledge are limited. Secondly, the different actors in the industry can have distinguished aims with such a technology and the potential improvement would only be optimized if every actor wanted a similar outcome. However, although it may exist better models than the recommended design, an implemented model that would be able to make predictions would facilitate the work massively for the active actors in the industry. Additionally, the demand for the case company's technology would significantly increase. Furthermore, a well-working solution could increase safety and improve proactive water management which would be important for all included parts.

To summarize, even though this thesis did not complete the entire procedure and test whether implementation would be attractive, a review of current studies and interviews was carried out, which gave an overview of the potential. Hence, it can be concluded that machine learning is a promising approach to meet the challenges and that this path should be explored further to improve the current methods successfully.

6.1 Further research

This thesis would be able to expand accordingly in future research in foremost a few different fields. Firstly, if the data would be available, it would be intriguing to implement the recommended model and understand if the template would be useful. Secondly, a progression of the proposed model to make it even more accurate alongside sufficient interpretability would be interesting. Thirdly, it would be appealing to obtain a more nuanced image of the industry by conducting this research with other stakeholders in the industry, such as water treatment plants and wastewater companies to see if the implementability could be even more applicable. Fourthly, an examination of how a model could work together with the data structuring tools that the industry expert highlighted would be favorable. Lastly, further research could include how this model should interact with other parts of the water distribution process in order to provide optimal value for the consumers.

Bibliography

Ahmad, T., & Chen, H. (2020). A review on machine learning forecasting growth trends and their real-time applications in different energy systems. *Sustainable Cities and Society*, *54*, 102010.

Alaliyat, S. (2008). *Video-based fall detection in elderlys houses* (Master's thesis).

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, *9*(5), 272.

Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine learning from theory to algorithms: An overview. *Journal of physics: conference series*, *1142*, 012012.

Aspers, P., & Corte, U. (2019). What is qualitative in qualitative research. *Qualitative sociology*, *42*, 139–160.

Ayodele, T. O. (2010). Types of machine learning algorithms. *New advances in machine learning*, *3*, 19–48.

Baryannis, G., Dani, S., & Antoniou, G. (2019). Predicting supply chain risks using machine learning: The trade-off between performance and interpretability. *Future Generation Computer Systems*, *101*, 993–1004.

Batanlar, Y., & Özuysal, M. (2014). Introduction to machine learning. *miRNomics: MicroRNA biology and computational analysis*, 105–128.

Bell, E., Bryman, A., & Harley, B. (2022). *Business research methods*. Oxford university press.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer.

Brester, C., Ryzhikov, I., Siponen, S., Jayaprakash, B., Ikonen, J., Pitkänen, T., & Kolehmainen, M. (2020). Potential and limitations of a pilot-scale drinking water distribution system for bacterial community predictive modeling. *Science of The Total Environment*, *137249*, 717.

Brownlee, J. (2020). *Data preparation for machine learning: Data cleaning, feature selection, and data transforms in python*. Machine Learning Mastery.

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, *300*, 70–79.

Castleberry, A., & Nolen, A. (2018). Thematic analysis of qualitative research data: Is it as easy as it sounds? *Currents in pharmacy teaching and learning*, *10*(6), 807–815.

Charles, K. J., Howard, G., Prats, E. V., Gruber, J., Alam, S., Alamgir, A. S. M., & Campbell-Lendrum, D. (2022). Infrastructure alone cannot ensure resilience to weather events in drinking water supplies. *Science of The Total Environment*, *151876*, 813.

Chou, J.-S., & Nguyen, T.-K. (2018). Forward forecast of stock price using sliding-window metaheuristic-optimized machine-learning regression. *IEEE Transactions on Industrial Informatics*, *14*(7), 3132–3142.

Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016). Data cleaning: Overview and emerging challenges. *Proceedings of the 2016 international conference on management of data*, 2201–2206.

Cifuentes, J., Marulanda, G., Bello, A., & Reneses, J. (2020). Air temperature forecasting using machine learning techniques: A review. *Energies*, *13*(16), 4215.

Creswell, J. W., & Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications.

da Silva, I., Hernane Spatti, D., Andrade Flauzino, R., Liboni, L., & dos Reis Alves, S. (2017). Artificial neural network architectures and training processes. *Artificial Neural Networks*, *1*.

Denscombe, M. (2010). *The good research guide for small-scale social research projects (6th ed.)* McGraw-Hill Education Open University Press.

Dhal, P., & Azad, C. (2022). A comprehensive survey on feature selection in the various fields of machine learning. *Applied Intelligence*, 1–39.

El Naqa, I., & Murphy, M. J. (2015). *What is machine learning?* Springer.

European Drinking Water. (2017). Nordic drinking water quality. https://www.europeandrinkingwater.eu/fileadmin/edw/documents_links/MaiD_Report_1_final_11.9.2017.pdf

Fratello, M., & Tagliaferri, R. (2018). Decision trees and random forests. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, *374*.

Freitas, A. A. (2019). Automated machine learning for studying the trade-off between predictive accuracy and interpretability. *Machine Learning and Knowledge Extraction: Third IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2019, Canterbury, UK, August 26–29, 2019, Proceedings 3*, 48–66.

Halldórsson, Á., & Aastrup, J. (2003). Quality criteria for qualitative inquiries in logistics. *European journal of operational research*, *144*(2), 321–332.

Højris, B., Christensen, S. C. B., Albrechtsen, H. J., Smith, C., & Dahlqvist, M. (2016). A novel, optical, on-line bacteria sensor for monitoring drinking water quality. *Scientific reports*, *6*(1), 1–10.

Horning, N., et al. (2010). Random forests: An algorithm for image classification and generation of continuous fields data sets. *Proceedings of the International Conference on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences, Osaka, Japan*, *911*, 1–6.

Jain, A., Patel, H., Nagalapatti, L., Gupta, N., Mehta, S., Guttula, S., Mujumdar, S., Afzal, S., Sharma Mittal, R., & Munigala, V. (2020). Overview and importance of data quality for machine learning tasks. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3561–3562.

Jajodia, T., & Garg, P. (2019). Image classification–cat and dog images. *Image*, *6*(23), 570–572.

Jierula, A., Wang, S., OH, T.-M., & Wang, P. (2021). Study on accuracy metrics for evaluating the predictions of damage locations in deep piles using artificial neural networks with acoustic emission data. *Applied Sciences*, *11*(5), 2314.

Jo, J.-M. (2019). Effectiveness of normalization pre-processing of big data to the machine learning performance. *The Journal of the Korea institute of electronic communication sciences*, *14*(3), 547–552.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational researcher*, *33*(7), 14–26.

Jurafsky, D., & Martin, J. (2023). *Speech and language processing*. Stanford University.

Kaggle. (2023). Iris species. https://www.kaggle.com/datasets/uciml/iris

Kamalov, F., Gurrib, I., & Rajab, K. (2021). Financial forecasting with machine learning: Price vs return. *Journal of Computer Science*, *17*(3), 251–264.

Kaur, G., & Oberai, E. N. (2014). A review article on naive bayes classifier with various smoothing techniques. *International Journal of Computer Science and Mobile Computing*, *3*(10), 864–868.

Kavitha, S., Varuna, S., & Ramya, R. (2016). A comparative analysis on linear regression and support vector regression. *2016 online international conference on green engineering and technologies (IC-GET)*, 1–5.

Kolahdouzan, M., & Shahabi, C. (2004). Voronoi-based k nearest neighbor search for spatial network databases. *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, 840–851.

Kramer, O., & Kramer, O. (2013). K-nearest neighbors. *Dimensionality reduction with unsupervised nearest neighbors*, 13–23.

Kumar, S., & Ghosh, A. (2019). Assessment of bacterial viability: A comprehensive review on recent advances and challenges. *Microbiology*, *165*(6), 593–610.

Kyngäs, H. (2020). Qualitative research and content analysis. *The application of content analysis in nursing science research*, 3–11.

LeChevallier, M. W., Schulz, W., & Lee, R. (1991). Bacterial nutrients in drinking water. *Applied and environmental microbiology*, *57*(3), 857–862.

Leedy, P. D., & Ormrod, J. E. (2019). Practical research.

Li, Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart disease identification method using machine learning classification in e-healthcare. *IEEE Access*, *8*, 107562–107582.

Li, Rong, S., Wang, R., & Yu, S. (2021). Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: A review. *Chemical Engineering Journal*, *405*(1), 126673.

Liang, Y., Li, S., Yan, C., Li, M., & Jiang, C. (2021). Explaining the black-box model: A survey of local interpretation methods for deep neural networks. *Neurocomputing*, *419*, 168–182.

Lindholm, A., Wahlstrom, N., Lindsten, F., & Schon, T. (2022). *Machine learn- ing a first course for engineers and scientists*. Cambridge University Press.

Liu, P., Wang, J., Sangaiah, A., Xie, Y., & Yin, X. (2019). Analysis and prediction of water quality using lstm deep neural networks in iot environment. *Sustainability*, *11*(7), 2058.

Maalouf, M. (2011). Logistic regression in data analysis: An overview. *International Journal of Data Analysis Techniques and Strategies*, *3*(3), 281–299.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*(1), 381–386.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PloS one*, *13*(3), e0194889.

Masini, R. P., Medeiros, M. C., & Mendes, E. F. (2023). Machine learning advances for time series forecasting. *Journal of Economic Surveys*, *37*(1), 76–111.

Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, *1*(4), 140–147.

Mohammed, H., Hameed, I. A., & Seidu, R. (2018). Comparative predictive modelling of the occurrence of faecal indicator bacteria in a drinking water source in norway. *Science of the Total Environment*, *628*(1), 1178–1190.

Morgan, G. A., & Harmon, R. J. (2001). Data collection techniques. *Journal-American Academy Of Child And Adolescent Psychiatry*, *40*(8), 973–976.

Muthukrishnan, R., & Rohini, R. (2016). Lasso: A feature selection technique in predictive modeling for machine learning. *2016 IEEE international conference on advances in computer applications*, 18–20.

Nadav-Greenberg, L., & Joslyn, S. L. (2009). Uncertainty forecasts improve decision making among nonexperts. *Journal of Cognitive Engineering and Decision Making*, *3*(3), 209–227.

Nocoli. (2023). Company information. https://chalmersventures.com/startups/nocoli/

Onyema, E. M., Almuzaini, K. K., Onu, F. U., Verma, D., Gregory, U. S., Puttaramaiah, M., & Afriyie, R. K. (2022). Prospects and challenges of using machine learning for academic forecasting. *Computational Intelligence and Neuroscience*.

Opdenakker, R., et al. (2006). Advantages and disadvantages of four interview techniques in qualitative research. *Forum qualitative sozialforschung/forum: Qualitative social research*, *7*(4).

Osisanwo, F., Akinsola, J., Awodele, O., Hinmikaiye, J., Olakanmi, O., Akinjobi, J., et al. (2017). Supervised machine learning algorithms: Classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, *48*(3), 128–138.

Osman, M. S., Abu-Mahfouz, A. M., & Page, P. R. (2018). A survey on data imputation techniques: Water distribution system as a use case. *IEEE Access*, *6*, 63279–63291.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, *2*(11), 100336.

Pereira, L. N., & Cerqueira, V. (2022). Forecasting hotel demand for revenue management using machine learning regression methods. *Current Issues in Tourism*, *25*(17), 2733–2750.

Pickles, J. C., Stone, T. J., & Jacques, T. S. (2020). Methylation-based algorithms for diagnosis: Experience from neuro-oncology. *The Journal of Pathology*, *250*(5), 510–517.

Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In *Machine learning* (pp. 101–121). Elsevier.

Prest, E. I., Hammes, F., Van Loosdrecht, M. C., & Vrouwenvelder, J. S. (2016). Biological stability of drinking water: Controlling factors, methods, and challenges. *Frontiers in microbiology*, *7*(1), 45.

Ray, S. (2019). A quick review of machine learning algorithms. *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*, 35–39.

Rong, M., Gong, D., & Gao, X. (2019). Feature selection and its use in big data: Challenges, methods, and trends. *IEEE Access*, *7*(1), 19709–19725.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Sapsford, R., & Jupp, V. (1996). *Data collection and analysis*. Sage.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, *2*(3), 160.

Schober, P., & Vetter, T. R. (2021). Logistic regression in medical research. *Anesthesia and analgesia*, *132*(2), 365.

scikit-learn. (2023). Machine learning in oython. https://scikit-learn.org/stable/

SMHI. (2023). Prognosuppfoljning. https://www.smhi.se/data/meteorologi/prognosuppfoljning

Smitha, N., & Bharath, R. (2020). Performance comparison of machine learning classifiers for fake news detection. *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, 696–700.

Sokolova, E., Ivarsson, O., Lillieström, N. K., A. andSpeicher, Rydberg, H., & Bondelind, M. (2022). Data-driven models for predicting microbial water quality in the drinking water source using e. coli monitoring and hydrometeorological data. *Science of the Total Environment*, *802*(1), 149798.

Stocker, M. D., Pachepsky, Y. A., & Hill, R. L. (2022). Prediction of e. coli concentrations in agricultural pond waters: Application and comparison of machine learning algorithms. *Frontiers in Artificial Intelligence*, *4*(1), 202.

Stockholm Vatten och Avfall. (2022). Dricksvatenkvalitét. https://www.stockholmvattenochavfall.se/globalassets/pdfer/rapporter/dricksvatten/dricksvattenkvalitet/certificate-kval-dekl-no-lo-eng-2022-04-28.pdf

Subramanian, J., & Simon, R. (2013). Overfitting in prediction models–is it a problem only in high dimensions? *Contemporary clinical trials*, *36*(2), 636–641.

Suthaharan, S., & Suthaharan, S. (2016). Support vector machine. *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, 207–235.

Svenskt Vatten. (2022). Mikroorganismer i vatten. https://www.svensktvatten.se/vattentjanster/dricksvatten/riskanalys-och-provtagning/mikroorganismer-i-vatten/

UNESCO World Water Assessment Programme. (2021). The united nations world water development report 2021: Valuing water. https://unesdoc.unesco.org/ark:/48223/pf0000375724

UNICEF. (2022). Water scarcity. https://www.unicef.org/wash/water-scarcity

Vaismoradi, M., Turunen, H., & Bondas, T. (2013). Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. *Nursing & health sciences*, *15*(3), 398–405.

Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine learning*, *109*(2), 373–440.

Vandeput, N. (2021). Data science for supply chain forecasting. *De Gruter*.

Webb, B., & Nobilis, F. (1997). Long-term perspective on the nature of the air–water temperature relationship: A case study. *Hydrological Processes*, *11*(2), 137–147.

Weller, D. L., Love, T. M., & Wiedmann, M. (2021). Interpretability versus accuracy: A comparison of machine learning models built using different algorithms, performance measures, and features to predict e. coli levels in agricultural water. *Frontiers in Artificial Intelligence*, *4*(1), 628441.

Whitman, R. L., & Nevers, M. B. (2003). Foreshore sand as a source of escherichia coli in nearshore water of a lake michigan beach. *Applied and environmental microbiology*, *69*(9), 5555–5562.

Williams, C., et al. (2007). Research methods. *Journal of Business & Economics Research (JBER)*, *5*(3).

Worlds Health Organisation. (2022). Drinking-water. https://www.who.int/news-room/fact-sheets/detail/drinking-water

Zarei, M. (2020). Wastewater resources management for energy recovery from circular economy perspective. *Water-Energy Nexus*, *3*(1), 170–185.

Zhang, Gao, X., Smith, K., Inial, G., Liu, S., Conil, L. B., & Pan, B. (2019). Integrating water quality and operation into prediction of water production in drinking water treatment plants by genetic algorithm enhanced artificial neural network. *Water research*, *164*(1), 114888.

Zhang & Hong, W.-C. (2021). Application of variational mode decomposition and chaotic grey wolf optimizer with support vector regression for forecasting electric loads. *Knowledge-Based Systems*, *228*, 107297.

A

Interview Questions

A.1 Questions asked to the case company

- Can you please introduce yourself?
- Please elaborate a bit about the drinking water industry.
  - Current actors?
  - Methods?
  - Familiar projects with machine learning included?
- What are the significant challenges within the bacteria level detection industry?
- What do you want to achieve by using machine learning in the bacteria detection process?
- How many days in advance do you need reliable information about bacteria level to be able to implement it into the devices?
- Will this be a fully automated process or are you expecting manual involvement?
- What would you say is preferred by the machine learning process, high complexity (potentially better accuracy) or interpretability (potentially worse accuracy) in the results?
- How detailed predictions are required by the model? An interval in which the bacteria level lies within or a more detailed value?
- In this stage of the research process, we have collectively decided to predominantly look at weather factors to understand forecasting possibilities. Can you think of other aspects that you also would find interesting to examine?
- Currently, there exists research saying machine learning is promising in the field, but so far, a lot of parameters are included in the predicting models. Do you know if parameters such as level of copper or electrical conductivity will be available?

A.2 Questions asked to the industry expert

- Can you please introduce yourself?
- Please elaborate a bit about the drinking water industry.
  - Current actors?
  - Methods?
  - Familiar projects with machine learning included?
- What are the major challenges within the industry in general?
- What are the significant challenges within the bacteria level detection industry?
- Do you think machine learning could be a tool to mitigate the current challenges?
- Do you see other ways of obtaining improved methodologies than machine learning?
- What would you say is preferred by the machine learning process, high complexity (potentially better accuracy) or interpretability (potentially worse accuracy) in the results?
- In this stage of the research process, we have collectively decided to predominantly look at weather factors to understand forecasting possibilities. Can you think of other aspects that you also would find interesting to examine?

A.3 Questions asked to the research engineer

- Can you please introduce yourself?
- How come you and your fellow researchers were interested in the study you conducted that is closely related to the research field of this report?
- From my perspective, I find the study really interesting. Did you feel that the study was successful?
- You tested quite a few models, if you would choose only one to implement in real-life practice, which would it be?
- Did you consider the trade-off between high complexity (potentially better accuracy) or interpretability (potentially worse accuracy)?
- Did you ever discuss other data features (pH, temperature etc.) than those mentioned in your study?
- Do you have other recommendations for me that you obtained during your project?

B

Possible implementation

This appendix contains a possible implementation of the model in this project. The used programming language is Python and consequently notations from the standard libraries. A common example is the lines preceded by # indicating that the lines consist of comments that do not affect the code. Other lines impact the outcome of the implementation. To simplify the implementation, scikit-learn's standard libraries are often used scikit-learn (2023). Additionally, in this implementation, three different models are included. The first one is a standard version of a Random Forest, the second is a Random Forest with an out-of-bag supplement and the third is a Random Forest including recursive feature elimination. The reason for several models is that these need to be evaluated individually with the specific data set. Consequently, all are worth implementing in order to obtain good results.

B.1 Loading the data set

```
# Import Pandas for data manipulation
import pandas as pd
# Read in the data set data
data = pd.read_csv('dataset.csv')
# Creates a list with the headings of the input features
features = data[list(data[data.columns[:-1]])]
# Displays the five first rows from the dataset
data.head()
```

B.2 Splitting the data set properly

```
# Import train_test_split from SciKit Learn to
# split the data
from sklearn.model_selection import train_test_split
# Seperates the last column containing the values that
# the model will predict
x = data.iloc[:, :-1].values
y = data.iloc[:, -1].values
# Split the dataset into training, test,
# .-and validation sets test_size
```

```
x_train, x_test, y_train, y_test =
train_test_split(x, y, test_size=0.2)
x_train, x_val, y_train, y_val =
train_test_split(x_train, y_train, test_size=0.5)
```

B.3 Standard Random Forest implementation

```
# Import RandomForestRegressor from SciKit Learn
# as the model
from sklearn.ensemble import RandomForestRegressor
# Setting up the standard version of random forest
random_forest = RandomForestRegressor()
# Fits the standard version of random forest to the data
random_forest.fit(x_train,y_train)
# The model makes the predictions
y_pred = random_forest.predict(x_test)
```

B.4 Out-of-bag Random Forest implementation

```
# Setting up the random forest including the
# out-of-bag supplement
random_forest_oob = RandomForestRegressor(
oob_score=True, max_features="sqrt")
# Fits the out-of-bag version of random forest to the data
random_forest_oob.fit(x_train,y_train)
# The model makes the predictions
y_pred_oob = random_forest_oob.predict(x_test)
```

B.5 Recursive feature elimination Random Forest implementation

```
# Import Recursive feature eliminating RandomForestRegressor
# from SciKit Learn as the model
from sklearn.feature_selection import RFECV
# Import matplotlib.pyplot for visualizing the
# feature importance
import matplotlib.pyplot as plt

# Visualising the feature importance
f_i = list(zip(features,random_forest.feature_importances_))
f_i.sort(key = lambda x : x[1])
plt.barh([x[0] for x in f_i],[x[1] for x in f_i])
plt.show()
```

```
# Setting up the random forest including recursive
# feature elimination
random_forest_rfe = RFECV(random_forest,cv=7)
# Fits the random forest including recursive
# feature elimination version to the data
random_forest_rfe.fit(x_train,y_train)
# The model makes the predictions
y_pred_rfe = random_forest_rfe.predict(x_test)

# Obtains the columns that contain the most
# important features
selected_features = random_forest_rfe.get_support()
# Adds the columns to a list
rfe_feature = features.loc[:, selected_features].columns.
tolist()
# Prints the result
print(str(len(rfe_feature)), 'selected_features')
print('RFE features')
print(rfe_feature)
```

B.6 Evaluating the models

```
# Import mean_squared_error from SciKit Learn as
# error measures
from sklearn.metrics import mean_squared_error
# Calculates and prints the RMSE for the actual and
# predicted values
print('Root Mean Squared Error:', mean_squared_error
(y_test, y_pred, squared=False))
# Calculates and prints the MSE for the actual and
# predicted values
print('Mean Squared Error:', mean_squared_error
(y_test, y_pred))

# Calculates and prints the RMSE for the actual and
# predicted values
print('Root Mean Squared Error:', mean_squared_error
(y_test, y_pred_oob, squared=False))
# Calculates and prints the MSE for the actual and
# predicted values
print('Mean Squared Error:', mean_squared_error
(y_test, y_pred_oob))

# Calculates and prints the RMSE for the actual and
# predicted values
print('Root Mean Squared Error:', mean_squared_error
```

```
(y_test, y_pred_rfe, squared=False))
# Calculates and prints the MSE for the actual and
# predicted values
print('Mean Squared Error:', mean_squared_error
(y_test, y_pred_rfe))
```