# Making Machine Learning Accessible for SMEs: Framework Requirements and Clustering Prototype

**Bachelor thesis**

Literature Research | Software Prototype

Bachelor studies

**"Digital Business & Software Engineering"**

MCI | DIE UNTERNEHMERISCHE HOCHSCHULE®

Supervisors:

**Dr. Pascal Schöttle**

**Deepak Khazanchi, PhD**


Author:

**Marijan Divkovic, BA**

**2018**

**1810836044**


29.05.2021

# Abstract

Machine learning is one of the most promising technologies in tech nowadays. It offers companies various advantages resulting in increased revenues, reduced costs, or improved customer satisfaction. However, research focuses mainly on large enterprises. Moreover, research shows that especially small- and medium-sized enterprises (SMEs) struggle to implement this promising technology. The reason can be partly explained by the specific challenges that such companies deal with. For example, due to their size, they have only limited resources and struggle to attract the talent that is required for the implementation of machine learning. Therefore, this paper focuses on developing the requirements of a framework that supports SMEs in the implementation of machine learning. To do so, a literature review was conducted. As part of the framework, the following eight distinct requirements are described: graphical user interface, data pre-processing, data sources, models, technology acceptance model, fairness, modularity, and local running environment. Each of the requirements is clearly described in the paper. Furthermore, five use cases are detailed that explain potential applications that could be used. Apart from developing the framework, this paper also implements one of the use cases as a prototype. Thus, as part of the second research question, the use case “product recommendation” is implemented, and assumptions and limitations are highlighted. The prototype is implemented as a deep learning-based recommender system following the idea of neural collaborative filtering using an already pre-processed data set.

# Kurzzusammenfassung

Maschinelles Lernen ist heutzutage eine der vielversprechendsten Technologien in der Tech-Branche. Es bietet Unternehmen verschiedene Vorteile, die sich in höheren Umsätzen, geringeren Kosten oder höherer Kundenzufriedenheit niederschlagen. Die Forschung konzentriert sich jedoch hauptsächlich auf große Unternehmen. Darüber hinaus zeigen Untersuchungen, dass vor allem Klein- und Mittelunternehmen Schwierigkeiten haben diese vielversprechende Technologie zu implementieren. Der Grund dafür liegt zum Teil in den spezifischen Herausforderungen, mit denen diese Unternehmen zu kämpfen haben. Beispielsweise verfügen sie aufgrund ihrer Größe nur über begrenzte Ressourcen. Daher haben sie Schwierigkeiten die erforderlichen Mitarbeiter zu gewinnen, die für die Implementierung von Projekten im Bereich des maschinellen Lernens wichtig sind. Daher konzentriert sich diese Bachelor Arbeit auf die Entwicklung eines Frameworks, das KMUs bei der Implementierung von maschinellem Lernen unterstützt. Zu diesem Zweck wurde eine Literaturrecherche durchgeführt. Als Teil des Frameworks werden die folgenden acht unterschiedlichen Anforderungen beschrieben: Grafische Benutzeroberfläche, Datenvorverarbeitung, Datenquellen, Modelle, Technologieakzeptanzmodell, Fairness, Modularität, lokale Betriebsumgebung. Jede der Anforderungen wird in der Bachelor Arbeit eindeutig beschrieben. Darüber hinaus werden fünf Anwendungsfälle im Rahmen des Frameworks erklärt. Neben der Entwicklung des Frameworks wird in dieser Bachelor Arbeit auch einer der Anwendungsfälle als Prototyp implementiert. So wird im Rahmen der zweiten Forschungsfrage der Anwendungsfall "Produktempfehlung" implementiert sowie Annahmen und Einschränkungen aufgezeigt. Der Prototyp ist als Deep Learning basiertes Empfehlungssystem implementiert. Die Implementierung folgt der Idee des neuronalen kollaborativen Filterns. Dabei kommt es zur Anwendung eines bereits vorverarbeiteten Datensatzes.

# Table of contents

# List of figures

# List of abbreviations

| | |
|---|---|
| AI | Artificial intelligence |
| AutoML | Automated machine learning |
| CRISP DM | Cross-industry standard process for data mining |
| DL | Deep learning |
| GUI | Graphical user interface |
| IT | Information technology |
| ML | Machine learning |
| SME | Small and medium-sized enterprise |
| TAM | Technology acceptance model |

# 1 Introduction

In an ever-faster moving and more globalized world where technology is continuously growing importance, one primary economic driver across many countries are small and medium-sized enterprises (SMEs) [1, p. 84], [2, p. 438]. Due to their nature of being small, SMEs face unique challenges. The most predominant one is their limited availability of resources. This challenge impedes SMEs from investing considerable amounts of money into R&D, innovation, and new technologies [3, p. 5]. Concurrently, SMEs are in greater need of innovation to remain competitive than large enterprises [2, p. 438].

One promising technology that gained traction in recent years and could offer promising opportunities to SMEs is artificial intelligence (AI), and more specifically, the field of machine learning. According to Tom Mitchell, machine learning can be described as "*any computer program that improves its performance at some task through experience*" [4, p. 2]. The reason for the rise of machine learning can be explained by the availability of a massive volume of data, cheap cloud storage, enormous processing power, and infrastructure that allows the smooth transfer of large data amounts [5, pp. 13–18]. Besides, machine learning also emerged because enormous data volumes became too large for humans to process manually. Therefore, even rule-based systems have reached their limits. Rule-based systems are characterized by a fixed set of rules that emerge from human experience [6, p. 921]. For example, in comparison to rule-based systems, machine learning models achieve better results for the use of chest image analysis than rule-based systems, as outlined by Ginneken [7].

The significant advantages of machine learning are its ability to adapt to altering environments based on the data that is supplied to machine learning models, as well as its ability to identify unknown patterns while analyzing large amounts of data [4, p. 2]. As a result, many new opportunities arise from the field of machine learning. According to a McKinsey Global Institute report, machine learning offers businesses across various industries many opportunities to enhance their performance [8]. Overall, machine learning can help SMEs boost revenues, reduce costs, and support them to analyze customer data or make decisions [9, p. 42].

## 1.1. Problem statement and research question

To benefit from the previously mentioned advantages, companies first must pursue a machine learning implementation. However, such deployments bring various barriers and challenges. First, some firms are unsure about the topic or do not know how to kick off respective initiatives [10, p. 19]. Moreover, companies usually do not have a distinct AI strategy [11]. In addition, many businesses miss either acceptance from top-management or organizational alignment [12, p. 18]. Finally, one of the most significant challenges is the lack of skilled employees [11], [13, p. 200].

Focusing more specifically on SMEs, Bauer et al. [14] highlight that SMEs especially experience difficulties with the implementation of machine learning and, as a result, have a lower adoption rate. Additionally, they suggest that frameworks designed to facilitate machine learning implementations could assist SMEs. Such frameworks could help as no employees with advanced technical knowledge would be required to utilize machine learning.

Therefore, to allow these kinds of companies to leverage this technology's vast benefits and reduce the barriers posed on them, this paper aims to describe the requirements of a framework that should support SMEs with implementing and evaluating machine learning. This framework will consist of several modules. Each module reflects one part of a more extensive software. In the end, the goal is to create software that can be easily used and applied by SMEs to experiment with machine learning. However, due to this paper's nature of being a bachelor's thesis, certain limitations must be considered. Hence, this paper will only detail the requirements for such a framework and define the respective modules that constitute the more extensive software. Thus, the following eight different requirements will be discussed:

- Graphical user interface
- Data pre-processing
- Data sources
- Data models
- Technology acceptance model
- Fairness
- Modularity
- Local running environment

Firstly, the research will focus on data pre-processing and data sources. Secondly, this paper will evaluate various machine learning models and evaluate which models to apply for different business cases. Thirdly, the role of fairness as well as the technology acceptance model. Finally, the paper examines requirements related to the graphical user interface (GUI), modularity, and local running environment.

Apart from identifying the framework requirements, this paper shall also focus on implementing a clustering prototype. This prototype is part of the clustering module that could be used, for example, to cluster customer data. This paper will not involve efforts to automize the data collection process. Hence, the prototype will use a pre-processed data set to produce meaningful results. Automatic data pre-processing shall either be implemented as part of another paper or shall be conducted manually. Moreover, the other modules will not be implemented as part of this research. They shall be implemented as part of future research regarding this framework.

To summarize, the main aim of this paper is twofold. On the one hand, this paper aims to describe the requirements of a machine learning framework that supports SMEs in evaluating how and whether machine learning is beneficial for them while serving as a tool that can be used to deploy the first machine learning models. On the other hand, it focuses on implementing a prototype representing the larger framework's clustering module.

Based on the mentioned research objectives, this paper should tackle the following two research questions:

**What are the requirements for an ML framework such that it can support SMEs?**

**How could a prototype of such a framework with the focus on clustering look like?**

## 1.2. Methodology

The initial idea for the research questions was suggested by the supervisor Pascal Schöttle. He has a strong background in the fields of machine learning and cybersecurity. Afterward, the idea was further developed jointly.

To answer the first research question, a literature review will be conducted. As reasoned by Jesson et al. [15, p. 140], the literature review will help to gain an in-depth understanding of the current state of research regarding the respective research topic. On the one hand, the literature review will focus on gaining a thorough understanding of machine learning and SMEs' topics in general. On the other hand, it will emphasize the overall requirements that must be considered for a machine learning framework to support SMEs. The first research question is answered if the requirements towards such a framework are identified and clearly described. To ensure proper academic standards while conducting the research, various quality criteria will be followed. Therefore, it will be assured that the most recent research papers will be used [16, p. 130]. The reliability of corresponding sources will be guaranteed by focusing on the journal's name, the respective authors and other articles they have written, and whether the individual papers were peer-reviewed and have a valid DOI number [17, p. 77].

To answer the second research question, the clustering prototype will be developed. In this regard, the respective implementation approach and assumptions will be outlined in detail. For example, as mentioned before, it will be assumed that the data pre-processing module is already implemented. Therefore, an existing and already pre-processed data set from Kaggle[1] will be used. In general, the entire prototype is programmed and tested in a controlled laboratory environment. The clustering prototype will be implemented using Python. Collaborative filtering will be applied to cluster the data. Further assumptions and details will be described in the relevant chapters. After the implementation is completed, the clustering prototype will be clearly explained. Lastly, the overall learnings from the development of the prototype will be highlighted and discussed. This research question is answered if the programming assumptions are clearly outlined, and the working prototype is programmatically implemented in Python.

[1] https://www.kaggle.com/

## 1.3. Structure

In general, the paper consists of six chapters. The introduction explains the overall research goals, research questions, and the methodology that this paper will follow. Chapters two and three provide theoretical background information on SMEs and machine learning. Chapter four answers the following research question: "What are the requirements for a machine learning framework such that it can support SMEs?". As a result, it discusses different requirements that must be considered for the framework. Chapter five describes the practical implementation of the customer clustering prototype to answer the following research question: "How could a prototype of such a framework with the focus of clustering look like?". Hence, this chapter ensures the comprehensibility of the clustering prototype's programming and highlights assumptions that have been made during the implementation. The last chapter discusses the overall findings regarding the research questions that have been defined.

# 2 Small and medium-sized enterprises

This chapter provides further theoretical information about SMEs. Therefore, it discusses the broad definition of SMEs, challenges related to the implementation of machine learning, as well as other characteristics of such companies.

To begin with, SMEs are considered one of the main economic drivers across many countries [1, p. 84], [2, p. 438]. In general, several distinct definitions exist for the term SME, and there is no universal definition applied across the world. For example, the European Union (EU) characterizes SMEs as entities with less than 250 employees working for the respective company while not surpassing an annual revenue of 50 million euro and/or an annual balance sheet total of 43 million euro [18, p. 10]. Simultaneously, in the United States (US), different definitions exist even between government institutions. However, the US International Trade Commission [19] suggests using the definition that is proposed by the SBA Advocacy, which states that SMEs are organizations with less than 500 employees. This definition is widely used and suggested as it is the most practical one to apply compared to the definitions described by other US government institutions. However, this paper will follow the definition that the EU proposes.

In general, SMEs deal with various challenges. Firstly, they have only limited resources, limiting their investments into new innovation [3, p. 5]. However, according to Cefis and Marsili [20, p. 626], there is clear evidence that innovation has a considerable impact on small firms' long-term survival. However, Werner et al. [21] also outline that in family-run SMEs, succeeding generations seem to be less innovative than the previous ones due to an increased level of risk averseness. At least for family-run businesses, this should be considered for the implementation of new innovative technologies such as machine learning.

Focusing on the implementation of IT (Information Technology) initiatives, according to Nguyen [22, p. 163], SMEs have a slow adoption rate and a high failure rate. Nguyen [22, p. 163] identified three main reasons for that. First, employees do not have a clear understanding of how and why IT is adopted. Secondly, misunderstandings exist regarding the IT adoption process. Lastly, IT resources are underdeveloped due to a missing alignment between business and IT strategy, limited financial resources, and overall IT expertise.

According to Devos et al., many SMEs do not even have their own dedicated IT department [23, p. 207]. Furthermore, Thong [24, p. 153] highlights that SMEs strongly rely on external partners regarding IT implementations due to time, financial, and expertise constraints. However, getting external support to conduct IT initiatives leads to certain risks and problems, resulting even in failing IT implementations [23, p. 208]. Moreover, Devos et al. [23] highlight that CEOs of SMEs often have little knowledge, limited time, or commitment despite them being significant stakeholders for technical implementations. This should be specially considered by external implementation partners who work together with SMEs on IT implementations.

Wang et al. [25] also emphasize that many SMEs do not engage in strategic planning due to missing aspirations by owners. Other reasons are inadequate time for strategic planning and a lack of experience [26] or internal implementation barriers [27]. Due to the lack of IT departments and strategic planning, it can be assumed that many SMEs do not even have a proper IT strategy. As a result, these findings might explain why many SMEs do not have a clear AI strategy and why they face a lack of internal know-how to conduct a machine learning implementation.

Focusing on machine learning, Bauer et al. [14] identified that both large companies and SMEs face various challenges regarding machine learning implementations. One of the biggest challenges is a lack of employees with adequate machine learning knowledge. Furthermore, another issue was the limited budget that usually competes with other initiatives that are not focused on machine learning. However, some companies also find it challenging to identify suitable business cases. Some companies are also worried regarding data privacy and the infringement of data privacy regulations such as GDPR. Finally, there is also missing acceptance from decision-makers and even among users regarding machine learning. However, regardless of these challenges, Bauer et al. also clearly explain that SMEs are considerably less likely to utilize machine learning applications than larger companies. SMEs also cope more with challenges concerned with entry barriers such as poor data quality or a lack of knowledge on potential use cases. However, regardless of their size, start-ups and small companies that sell machine learning applications as distinguished products seem to be outliers. Finally, to support SMEs, Bauer et al. propose to support SMEs with machine learning by offering them a machine learning framework that reduces the need for specialized resources. In addition, they also propose that partnerships with external companies can help SMEs overcome the entry barriers.

# 3 Machine learning

This chapter focuses on providing further theoretical background. Thus, it describes the current literature on machine learning. As a result, it provides information on various definitions, areas of application, types of machine learning, challenges regarding machine learning, as well as an overview of automated machine learning (AutoML).

Before focusing on machine learning, it is relevant to introduce the term AI. John McCarthy initially outlined the term AI as “the science and engineering of making intelligent machines” [28, p. 2]. A more recent definition is provided by Russell & Norvig who describe AI “as the study and construction of rational agents” [29, p. 7]. In this context, an agent perceives and takes actions to reach specific objectives based on certain beliefs. Russell & Norvig also describe other categorizations of the term AI [29, p. 5]. However, this paper will focus on the category that concentrates on rational agents. This is because this category is beneficial to the application of machine learning within AI [30, p. 4].

Besides, an essential distinction regarding AI is the differentiation between weak and strong AI. Weak AI can only be seen as a tool, and hence, it only pretends to think. Therefore, it can only perform specific tasks. At the same time, strong AI can be seen as a mind that thinks. Thus, it can conduct a variety of different or even new tasks [31].

Focusing on machine learning, one of the most popular definitions is framed by Tom Mitchell. He describes machine learning as "any computer program that improves its performance at some task through experience" [4, p. 2].

He further details machine learning in the following manner:

> "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." [4, p. 2].

The last definition can be applied to various problems. For example, this definition could be applied to the game checkers. The task T would be to play checkers. The performance P would be the percentage of the games won against another player. Moreover, the experience E could be described as the games played to practice the game. Thus, by practicing the game checkers by playing matches against oneself, a computer would gain more experience playing the game. As a result, the computer would win more games.

However, the terms machine learning, deep learning, and AI are often confused with each other. Thus, it is relevant to outline that each term is distinct, as shown in figure 1. As a result, machine learning is just a subset of AI and, as such, only describes methods used to learn specific patterns from data [30]. Moreover, deep learning is a subset of machine learning, and it describes models that use several layers to learn from data to make respective predictions [32].

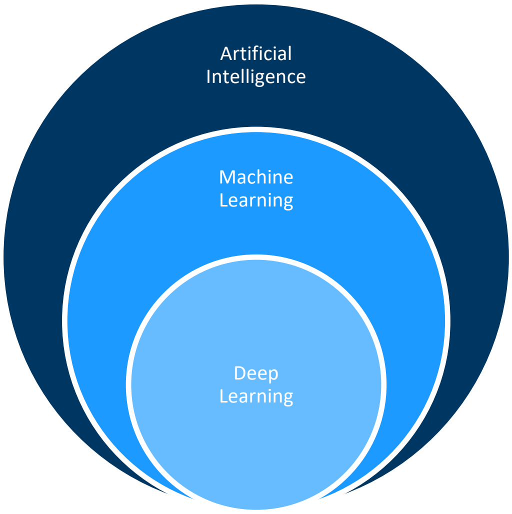


Fig. 1: Differentiation between AI, ML, and DL. Adapted from [33, p. 37624]

In general, many different applications are using machine learning, including various areas such as self-driving cars [34], robotics [35], image classification[36], or speech recognition [37]. However, machine learning should not be the default method. Machine learning makes especially sense in the following cases: analyzing substantial data amounts and understanding complicated problems, fast-changing environments in which adaptiveness towards new data is required, complicated challenges that cannot be solved using traditional approaches, or problems that require extensive manual work or roles if solved using traditional approaches [38, p. 7].

Overall, machine learning systems can be classified in different ways. Firstly, they can be grouped based on the amount and type of training under human supervision. Secondly, it is possible to classify them based on their ability to learn gradually from provided data. Lastly, they can be categorized based on how they conduct generalization. This paper focuses on the first categorization. Based on this categorization, it can be distinguished between supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning [38, p. 8].

Supervised learning algorithms require a high level of supervision as labeled training data sets are required that contain the desired solution. Such algorithms are usually used for tasks involving classification or regression. Unsupervised learning algorithms learn without any supervision as they try to learn with unlabeled training data sets. They are usually used for clustering, anomaly detection, visualization reduction, or association rule learning. Semi-supervised learning combines aspects of two supervised and unsupervised learning. They mainly use unlabeled data but require a small amount of labeled data to make better predictions. Finally, reinforcement learning is different from the other types of machine learning models. It involves an agent that is given a specific goal. However, the agent can freely decide how to achieve the goal. A reward/penalty system is used to improve the learning curve. Iteratively, the agent will become better [38, pp. 8–15].

Usually, while engaging with machine learning, various challenges must be considered. First, appropriate availability of data needs to be ensured to train the models. However, apart from ensuring data quantity, it is also significant to ascertain data quality. Thus, data must be representative enough to generalize well to new data. The data must also be free of errors or outliers. In addition, data must contain the right features to make the predictions more accurate, which is part of the feature engineering process. Apart from ensuring data quality and quantity, overfitting and underfitting must be avoided. Overfitting occurs if machine learning models only perform well on training data but not on new data. Underfitting happens if machine learning models do not even perform well on the training data. In both cases, different measures can be taken that usually include changing models, using different hyperparameters, improving data quantity or quality [38, pp. 24–30].

To overcome all these challenges, companies usually require data science and machine learning experts. Moreover, even if a company employs such experts, it still requires an extensive amount of time until such experts can prepare the data appropriately and select the most effective model. As a result, to make the application of machine learning more affordable and easier applicable, in recent years, the concept of AutoML emerged [39, p. 9].

In general, there are many similar definitions of AutoML. Zöller and Huber [40] describe it as a field of machine learning research that empowers people to deploy machine learning applications without advanced machine learning or statistics knowledge. Thus, AutoML makes experts of such fields redundant in the overall process. Yao et al. [41] also highlight in their definition that AutoML helps to replace human beings while automating the overall process of creating and deploying machine learning. Also, He et al. [39] underline that AutoML automates the creation of machine learning pipelines.

Focusing on all these definitions, it becomes clear that AutoML offers many advantages. Some advantages of AutoML are faster deployments of machine learning applications, more efficient performance validation of deployed machine learning applications, reduced need for experts, and an increased level of accessibility of machine learning [41, p. 1]. However, AutoML is still in its infancy. As a result, many problems still need to be addressed, such as the mathematical interpretability, the reproducibility of AutoML results, robustness towards adversarial data, the lack of a library that focuses on the complete AutoML pipeline, the missing abilities to use former data to tackle new tasks or to remember knowledge from old data [39, pp. 23–26].

# 4 Framework requirements

This chapter deals with answering the first research question. As such, it provides information on existing literature and describes each of the individual requirements. In the end, it summarizes the findings and discusses each of the requirements.

To answer the first research question and identify relevant framework requirements, it is significant to outline that there are already some ML frameworks being proposed.

First, Gupta and Pathak [42] describe an ML framework that predicts customers' purchase decisions based on dynamic pricing. Their framework proposes several stages. It starts with the collection of data from databases and several data points. Afterward, the data is pre-processed to evaluate the data's relevancy and prepare it for data analysis. Then, relevant attributes are selected to support customer segmentation. By utilizing these attributes, customer segmentation is conducted. Subsequently, a dynamic price range is defined for each customer group. In the end, the likelihood of a purchase is identified based on the customer groups and their price range.

Moreover, Jain et al. [9] describe a performance framework helping SMEs predict their strategic planning impact to stay competitive. It consists of four modules, each of which covers specific functionality and uses different machine learning techniques. The first module captures customers' facial expressions and voice responses to evaluate their current satisfaction and mood. The second module makes personalized product recommendations by taking customer purchase history, preferences, and interests into consideration. The third module makes individualized customer offers based on customers' purchase history, billing information, and account information. The last module predicts demand forecasts using sales data, seasonality data, product data, and other relevant information.

Furthermore, Huang and Rust [43] developed a strategic AI framework for marketing planning. It consists of three circular stages – marketing research to understand the market and competitors, marketing strategy to develop a strategy, and marketing action to execute the respective strategy. The feedback of the last stage will loop again into future marketing research and close the circle. The authors propose that each of the steps could be supported by AI. During each of the three stages, they propose to use different forms of AI to automate recurring tasks, the processing of data, and the analysis of interactions and emotions.

In addition, Mahmoud et al. explain in their work a model that utilizes AI to improve the hiring and performance appraisals process of companies [44]. Their idea is to use various data points to predict the performance of new hires. Thus, personal information of current and previous employees, appraisal information, as well as candidate information are used. To produce respective results, they use a decision tree that is based on one class. The results of the decision tree are the potential performance of the candidate in defined areas.

Finally, Pittino et al. discuss the use of automatic anomaly detection for production manufacturing machines using supervised and unsupervised machine learning [45]. Thus, they made use of regression model control charts and anomaly classification. They also mentioned that due to the scarcity of data, they were not able to use deep learning techniques.

Apart from such frameworks, it is also relevant to look at the steps involved in a machine learning initiative. Focusing on the literature used in the following pages, one can distinguish between a strategic part and a technical part of a machine learning implementation.

First, starting with the strategic part of a machine learning implementation, it is relevant to understand its business strategy and strategic objectives. Thus, it will become clear how machine learning can support the business and generate value towards achieving the set goals [5, pp. 92–93].

The next step is to gain an understanding of the ambitions of the company and different departments. Not every department requires to implement processes that are backed by sophisticated machine learning algorithms. Some departments might even show resistance. Therefore, it is also necessary to keep change management in mind and start the journey first with those departments and individuals who want to engage with machine learning [5, pp. 94–96], [10, p. 93].

Subsequently, companies need to identify and evaluate opportunities on possible use cases for machine learning together with the identified supporters. This can be done by utilizing, for example, the AI canvas by Dewalt [10, p. 77] or the AI maturity in combination with the AI heat map by [5, pp. 97–105]. The result of these will constitute the AI strategy.

After creating an AI strategy aligned with the business strategy and identifying a motivated team, companies can start the technical part of the implementation.

First, firms need to decide whether they want to focus on a bespoke solution or an off-the-shelf solution. The first provides more flexibility but also requires more resources. The latter can be easier implemented but provides less flexibility. Which one to choose depends strongly on the companies' resources and the use case [5, pp. 118–121], [10, pp. 91–93].

When it comes to building a machine learning system, according to Raschka [46, p. 11], the following four distinct steps can be distinguished: pre-processing, learning, evaluation, and prediction.

The first and most crucial step is the pre-processing of data. Its goal is to transform the raw data into a form and shape required for the respective learning algorithm. Part of pre-processing is feature extraction, feature scaling, feature selection, dimensionality reduction, and sampling. The last is concerned with splitting the data set into a training and test set to ensure that the model does not only perform well on the training data set. The next step is the training. It involves choosing the suitable model, choosing performance metrics to measure the performance, applying cross-validation techniques to estimate the model's generalization performance, and optimizing hyperparameters to fine-tune the final performance. After an appropriate model is chosen and trained with the training dataset, the evaluation phase begins. It is concerned with evaluating how well the model performs with the test dataset. If the results are correct, the model could be used to predict using new data as part of the last phase [46, pp. 12–13]. Apart from the just mentioned steps that are described by Raschka [46, p. 11], a prevalent model to describe the data science life cycle is the Cross Industry Standard Process for Data Mining (CRISP DM) [47]. It is used in data science or machine learning projects for organization, planning, and implementation. It consists of 6 steps. Business understanding to identify the business requirements. Data understanding to understand the available and required data. Data preparation to identify how data needs to be adapted. Modeling describes the modeling techniques that should be used. Evaluation to identify the best model. Finally, deployment to ensure that stakeholders can access the results of the model.

In general, there are different potential use cases for AI within companies. In this regard, Dewalt [10, p. 48] frames the word AI product pattern to describe how various AI technology applications solve similar repetitive business cases. Hence, the most promising AI product patterns are computer vision, natural-language processing NLP, next-in-sequence predictions, and collaborative filters, according to Dewalt [10].

Computer vision applications use large amounts of images and videos to make predictions. They can be used to classify images, identify objects, or for image restoration. Business processes that involve the processing of photos or videos are a good starting point for utilizing computer vision. NLP applications make predictions by using language data. Potential NLP solutions are machine translation, speech recognition, speech generation, entity recognition/extraction, text generation, or sentiment analysis. NLP can be primarily utilized for the process that involves the reading and classifying of documents. Next-in-sequence predictions predict the subsequent results based on the previous ones. As a result, they solve many different business problems such as predicting futures sales, detecting anomalies, predicting user behavior, or predicting KPIs. Collaborative filtering predicts user behavior based on the behavior of other similar users. It is used mainly for recommendation systems. Hence, it could support business processes where users are given many choices, such as e-commerce stores or content providers such as Amazon Prime or Netflix [10, pp. 49–65].

The subsequent sub-chapters will describe eight requirements and respective use cases that need to be considered for a machine learning framework that supports SMEs with the implementation of machine learning, as shown in figure 2.


Fig. 2: Overview of requirements and use cases.

## 4.1. Data pre-processing and data sources

To begin with, the most crucial aspect of every machine learning initiative is the selection of appropriate data sources and the respective pre-processing of the data. These two aspects also require a lot of in-depth knowledge and expertise. Thus, companies usually require experts who know how to find adequate data and how to process the data adequately to be used for machine learning algorithms. Therefore, these aspects were chosen to be essential requirements and will be further described in this chapter in more detail.

First, data pre-processing can be described as a process that transforms raw data from the real world into data that can be used for data analysis, according to Famili et al. [48, p. 5]. The process involves all activities that need to be conducted before the data can be analyzed. The process needs to be conducted to solve issues with data that restrict data analysis to understand the nature of the data better to facilitate the data analysis and gain more significant insights from the data. Garcí et al. [49, pp. 10–16] distinguish between data preparation and data reduction.

Data preparation contains all techniques used to prepare the data so that it can be used as an input for an algorithm. Thus, it follows the definition as mentioned above of data pre-processing. Usually, it must always be conducted. If it is not conducted, the result might be wrong results, or the algorithm might not work. In general, data preparation is required to tackle many different distinct issues using the following techniques: data cleaning, data transformation, data integration, data normalization, missing data imputation, and noise identification. Data cleaning describes all methods that are used to correct corrupt data, remove inaccurate data from the data set, and reduce details from the data that are not required. Various data cleaning activities intersect with other data preparation techniques. Data transformation is concerned with transforming data from one format to another to ensure that it can be used appropriately for specific algorithms and improve the efficiency of these algorithms. Data integration refers to the process of combining data that originates from different data stores. This process is relevant to prevent redundancies and inconsistencies within the final data set. Data normalization ensures that similar measurement units, scales, or ranges are used across all attributes. This helps to ascertain equal weights across the attributes. Missing data imputation replaces missing values with intuitive data as algorithms produce mostly better results with a well-estimated value than with a missing value. Noise identification has the primary goal of identifying errors or variances in the data set. This technique also intersects with data cleaning and data transformation [49, pp. 10–16].

Data reduction consists of all techniques used to reduce the original data set to a version containing a reduced amount of data. Thus, data reduction does not change anything about the structure of the original data. It only reduces the size of the data set in a particular manner. In comparison to data preparation, data reduction might be an optional step. However, most of the time, there are various use cases when it makes sense to use data preparation. In general, the following data reduction techniques can be distinguished: feature selection, instance selection, discretization, feature extraction, and instance generation. Feature selection is focused on getting rid of all features that are not relevant and meaningful to make good predictions. Thus, the goal is to choose a set of features that reflect the original distribution as similarly as possible while having a reduced set of features. Instance selection is a method that uses a subset of the overall data to achieve the same result as with the complete data set. This method is often used for internal validation and is utilized to avoid over-fitting. Discretization is a technique that helps to convert quantitative data into qualitative data. It is a technique that can be part of both data reduction and data preparation. Finally, feature extraction and instance generation alter values of data subsets or attributes [49, pp. 10–16].

Discussing these techniques, it is also relevant to mention the problems that go hand-in-hand with data. According to Famili et al. [44], data problems can be clustered into the following three categories: too much data, too little data, or fractured data. As part of the first category, data can be corrupt or noisy. Both might originate from various problems during the data collection process. Another issue within the first category is irrelevant data. A bigger problem than the first category is usually the second category. The second category involves missing attributes, missing attribute values, or the overall data set might just be too small. The last category contains incompatible data, multiple sources of data, or data from multiple levels of granularity. Each of the problems within the distinct categories needs to be appropriately distinguished during the data pre-processing phase to make sure that the algorithms can work properly.

Another critical aspect of the technical implementation is the determination of data sources and data availability, as machine learning models require a considerable amount of data to make good predictions [5, pp. 123–124], [38, p. 24]. The right amount of data depends strongly on the representativeness and quality of the data and the number of relevant features and the chosen model [38, pp. 24–28]. As a result, it is also relevant to make sure that the data is appropriately adjusted to the needs of the specific algorithm. Different ML algorithms have different requirements for how the data that servers as an input needs to be pre-processed. In general, data can either be gathered internally or externally, and potential data sources might be databases, data warehouses, or the internet [50, p. 8]. However, Yan et al. [51] argue that external data might be untrustworthy, and as a result, they also propose a framework on how to ensure high data quality while dealing with external data.

## 4.2. Use cases and models

Like in the case of data pre-processing, a profound understanding of machine learning is required to choose a suitable machine learning model. To make the selection of a machine learning model easier and to ensure that a suitable machine learning model is selected for the right use case, this sub-chapter describes some possible use-cases and respective models.

To make the selection of models easier, cheat sheets have been created. One widely used cheat sheet is the one that is provided by Microsoft, as illustrated in figure 3. As part of this research, it will be used to describe potential machine learning models for the use cases discussed within this chapter.

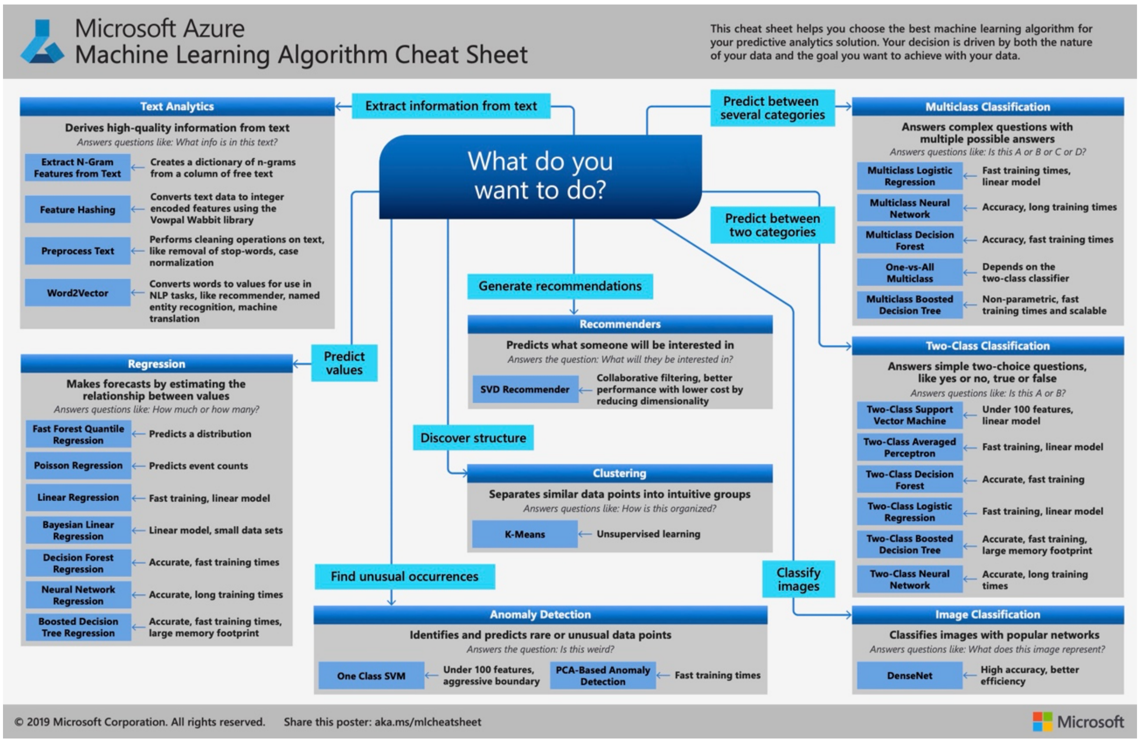


Fig. 3. Machine learning algorithm cheat sheet. [45]

To begin with, several machine learning use cases within SMEs can be derived from customer data. This would be a good start for every SME because usually, each company deals with customers. Furthermore, even if respective customer data is not available right from the beginning, it could be gathered within a reasonable period. Overall, one potential use case would be to recommend products to customers, as also suggested by Jain et al. [9]. Thus, a recommender model could be used to make the respective recommendations.

Moreover, clustering and, more specifically, k-means could be used to group customer data. This could help to identify new customer groups and could help to develop a more targeted marketing strategy. An example of this can be found in the papers of Ezenkwu et al. [49], who use k-means to segment customers while describing its benefit for targeted customer services.

Customer data could also be used with a two-class classification to identify whether a user will or will not buy a specific product. An example of this use case can be seen in a research paper by Gupta and Pathak [40], which discusses the prediction of online purchases based on dynamic pricing.

Another promising area where machine learning can be applied within SMEs is the hiring process. Every company is involved in hiring employees, and usually, data is already available or can be gathered. Machine learning could be used to predict whether candidates are a good fit for a particular job. In this regard, as mentioned before, Mahmoud et al. [44] suggest a model that predicts the expected performance of candidates. However, they also mention that their model requires plenty of data. Thus, especially for SMEs, it would be essential to introduce a machine learning model that does not require massive data sets as it might be difficult for small companies to gather extensive data sets. A combination of text analytics to gain data from CVs and regression to evaluate the performance could be used to build the model. Mahmoud et al. [44] suggest in their paper a decision tree. However, especially for this use case bias must be considered carefully. Moreover, this tool can only serve as a supporting tool within the hiring process and cannot make the final decision.

Finally, SMEs producing goods by themselves could also profit from machine learning by using anomaly detection. It could be used to identify anomalies within production machines and help to repair machines in advance. For this purpose, the respective anomaly detection machine learning models of the Microsoft framework could be applied. As mentioned before, Pittino et al. [45] describe in their paper different models that can be applied to implement anomaly detection on in-production manufacturing machines.

## 4.3. Fairness

Due to continuous advances in machine learning over the past decades, the technology is being used continuously for more sophisticated use cases. As a result, machine learning algorithms are utilized to make crucial decisions, such as whether someone should be granted a loan or whether some content should be removed from social media platforms or not. This development has resulted in concerns on whether these algorithms might be biased, perceived as unfair, or even institute and foster discriminatory behavior. These concerns have also been proved by research within the past years. As a result, a whole research field has developed on the fairness of machine learning algorithms [54], [55]. Thus, this is an essential requirement that needs to be covered within this framework.

However, what is fairness? Chouldechova and Roth [54] identified that two main groups towards the definition of fairness exist within the literature. On the one hand, one group of definitions is concerned with the statistical definition of fairness which focuses on fairness between demographic groups. On the other hand, definitions of fairness can be clustered according to the individual definitions of fairness. It puts its emphasis on comparing fairness between individuals instead of larger demographic groups. Mehrabi et al. [55] also point out that there are many definitions of fairness and that it is still one of the challenges within the overall research that needs to be tackled.

In general, there might be different sources for bias. Mehrabi et al. [55] define two main categories. First, they identified bias arising from data. For example, if training data already discriminates against people of color, it can be expected that the machine learning algorithm that might be used for a loan granting the application, will also engage in such discriminatory behavior. Chouldechova and Roth [45] also emphasize that bias might be encoded in the training data that serves as an input for machine learning algorithms. The other main category identified by Mehrabi et al. [55] is a bias that originates from how algorithms work and impede them from making unbiased decisions even with unbiased data.

According to researchers, to tackle bias, three types of methods can be distinguished [55, pp. 13–14] [56, pp. 4–5]. First, pre-processing techniques can be used to tackle discrimination before the algorithm uses the training data. Secondly, in-processing techniques can modify and change machine learning algorithms to make sure that discrimination during the training process within the model is removed. Thus, fairness metrics are often integrated to optimize the models appropriately. Thirdly, post-processing techniques can be used after the model training has been conducted to transform the model's output and improve the prediction fairness without knowing the actual machine learning algorithm.

## 4.4. Technology acceptance model

To implement new technology such as machine learning within a company and take advantage of it, it is relevant to gain acceptance across all company levels [57]. In this regard, Bauer et al. [14] identified that one challenge within SMEs regarding the implementation of machine learning is the lacking acceptance of decision-makers and users towards machine learning. One theory that can be utilized is the TAM to identify the reasons for this development and to ensure performance gains through machine learning within SMEs. It is one of the most prominent theories within the research field of information systems. It is used to describe the acceptance of individuals towards an information system [58]. The theory was developed by Davis [57]. He developed it based on the theory of reasoned action by Fishbein & Ajzen [59]. The TAM assumes that a user's intention to use a system is mainly determined by the perceived usefulness and the perceived ease of use, as shown in figure 4.

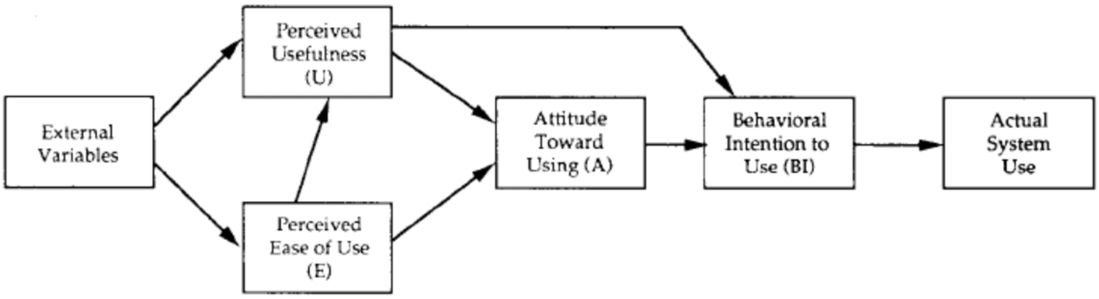


Fig. 4: Technology acceptance model (TAM). [60, p. 985]

The perceived usefulness describes the level an individual believes that using technology is beneficial for their work. The perceived ease of use describes to which level a user believes that using technology is effortless. Thus, the more technology helps to boost job performance, and the easier it is to use the technology, the more accepted it is according to the model [57]. The model has already been used by researchers such as Alhashmi et al. [58]. They found out that both factors positively affect the implementation of AI projects within the healthcare sector. Thus, focusing on the findings of Bauer et al. [14], the reason for the missing acceptance of machine learning within SMEs could be explained due to a negative perception towards usefulness or ease of use. Thus, to take such negative perception into consideration of ML implementations in SMEs, it is relevant to identify the relevant factors that might influence perceived usefulness and ease of use for each of the use cases that were outlined previously as part of this framework.

In general, there are various factors that influence perceived usefulness and ease of use. Alhashmi et al. [61] identified that managerial, organizational, operational, and IT factors impact both. Managerial factors refer to decision-makers who need to build trust towards the use of technology. Trust plays a significant role in SMEs [23]. Organizational factors refer to variables such as training programs and local expertise among employees that enable and help other employees to use new technology. Finally, operational factors refer to variables such as the perceived enjoyment. Further factors that could be analyzed to better understand the perceived usefulness within SMEs could be factors such as the impact of machine learning applications on job performance, productivity, quality of results, or usefulness [57]. Moreover, perceived ease of use could be measured by analyzing how difficult it is to integrate machine learning applications into the existing workflow. Moreover, it could be analyzed whether such applications are understandable and whether the behavior of the applications makes sense to employees. Besides, frustration and confusion while using such applications could also be measured [57].

## 4.5. Miscellaneous (GUI, modularity, local running environment)

This framework leans toward the idea of AutoML. As such, the GUI needs to be designed to be easy to use by individuals who have no in-depth experience with machine learning or statistics. As a result, after opening the framework, it should show an overview of the use cases that can are covered, as shown in figure 5.

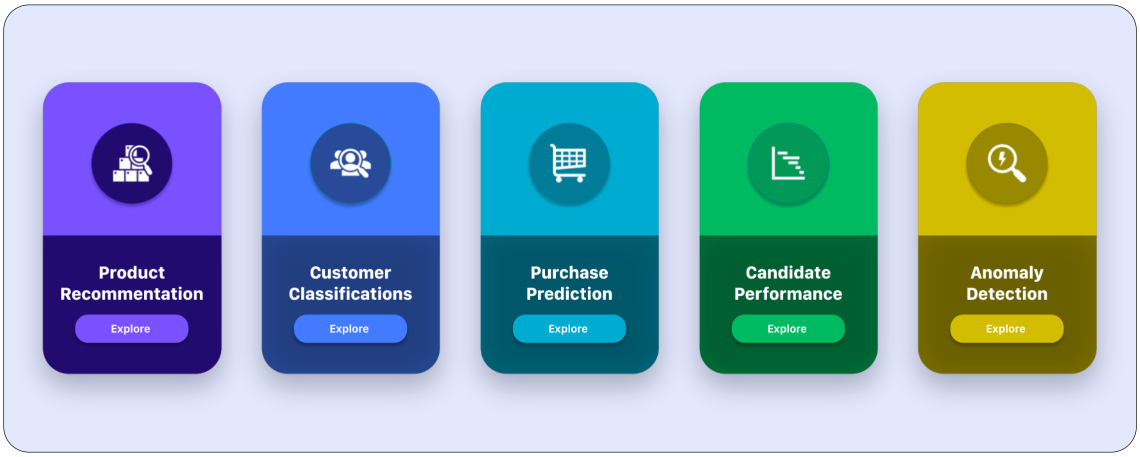


Fig. 5: Potential GUI of framework.

Moreover, the user experience behind each of the use cases should be built around the idea of a machine learning pipeline, as illustrated in figure 6. The AutoML product of Google[2] inspires the design for the prototype.

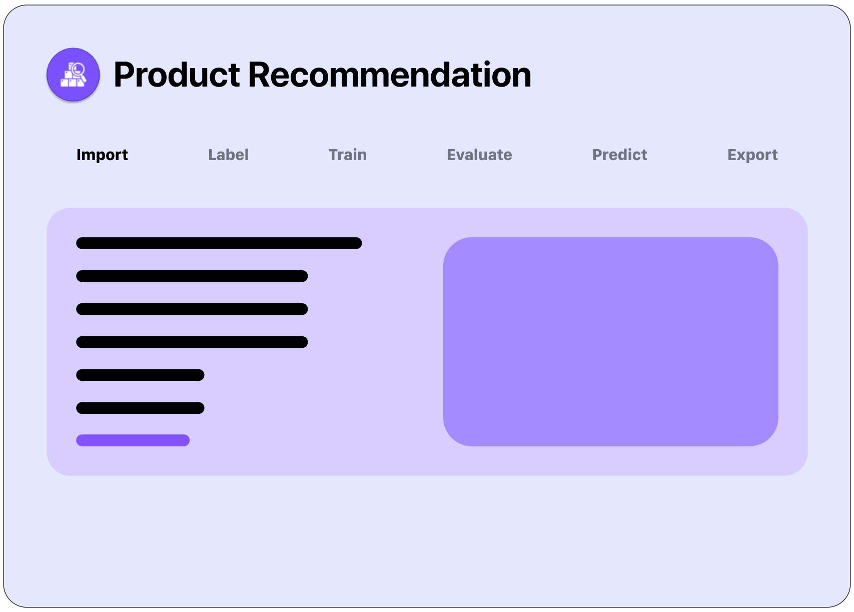


Fig. 6: Prototype of use case page.

[2] https://cloud.google.com/automl

As the framework covers different use cases, modularity is a relevant requirement and plays an influential role. By focusing on modular programming, changes can be easily applied to the respective modules, while new modules can be added without affecting the existing ones. Such an approach also ensures that change is only limited to the modules that are being modified or refactored. Thus, more flexibility can be provided [62, pp. 10–15].

Finally, it is relevant to ensure that the framework runs locally on the computers of the respective companies as an on-premises solution. This requirement originates from the fact that SMEs have considerable concerns regarding data security, which creates a barrier towards the usage of data-driven solutions, according to Coleman et al. [63]. Thus, running such software locally also ensures that the respective companies control the software and the data.

### 4.6. Summary

Overall, the main aim of this chapter was to identify the requirements of a machine learning framework that supports SMEs in their efforts to evaluate and implement machine learning. As a result, five use cases and eight essential requirements have been identified and described.

To begin with, customer data could be used to provide better product recommendations to customers and increase the number of sales. Moreover, it could be utilized to segment customers better, which would help to target customers better. Sales data could also predict whether customers will buy a product, which would help reduce customer churn. Furthermore, machine learning could also be used within the recruiting process. With the help of machine learning, it would be possible to predict better whether potential candidates might be a good fit for the respective role. The last identified use case is concerned with anomaly detection within production companies. The machine data could be analyzed continuously to identify anomalies in advance.

For each of these modules, potential data sources must be clearly described. Moreover, the data pre-processing should be automized as much as possible. Where automation is not possible, at least clear descriptions should be maintained on how the data needs to be appropriately prepared. Furthermore, each module must implement appropriate machine learning algorithms that fit the respective use case. To provide users with some flexibility, users should be given the optional possibility to define some hyperparameters themselves. To ensure flexibility, the overall framework should be built based on modularity to ensure that existing modules can be easily adapted and new modules can always be added. Moreover, fairness needs to be considered within each of the modules to avoid biased results of the machine learning models. Besides, the TAM should also be considered while developing each of the modules to ensure a high acceptance level of employees who use the framework. Furthermore, the overall framework and all modules should be developed as an on-premises solution to consider the data security concerns of SMEs. Finally, the first draft of a potential GUI has been described. However, for each of the modules, a specific GUI must be developed and tested.

To summarize, AI and, in particular, machine learning is one of the most promising technologies for companies these days to improve their business processes. However, the implementation of machine learning currently focuses mainly on large enterprises.

Moreover, implementing machine learning brings apart from opportunities also many challenges – especially for SMEs. Therefore, the framework should be built upon the idea of AutoML and automize as much as possible to make its application easy for people who do not have in-depth knowledge about machine learning or statistics. Such a framework would help to democratize machine learning for SMEs.

# 5 Customer clustering prototype

Every company deals with customers, and hence, customer data can be confidently accumulated over time for machine learning purposes, making individual use cases very attractive for SMEs. As a result, the following chapter will cover the second research question on how customer data can be clustered to generate product recommendations.

Most businesses seek to maximize their revenue while realizing a competitive advantage. To achieve this, many companies make use of market segmentation. Market segmentation describes the fact that businesses must understand who their customers are by identifying specific customer segments related to specific product demand. Businesses can target specific customer clusters after understanding profitable customers and create a specific marketing mix for each of the identified clusters. By applying a specific marketing mix to each segment, a more targeted message can be sent to the right people instead of sending a very generic message to all customers. Thus, this helps to maximize profits while creating a competitive advantage [64, p. 79].

One of the primary use cases that were outlined in the last chapter was product recommendation. It helps SMEs to recommend products to customers by clustering customers into groups. Focusing on machine learning, this use case can be implemented with recommender systems. In the next sub-chapters, a theoretical overview of recommender systems, the implementation as well as the limitations of the prototype will be discussed.

## 5.1. Recommender systems

According to Falk [65, p. 13], recommender systems can be defined as a system that utilizes user information, content, and interactions between users and items to compute and provide appropriate content to users. Isinkaye et al. [66] describe them as information filtering systems that help predict whether an item should be recommended to users based on their user profile. As a result, recommenders help deal with the problem of information overload that users must deal with while being offered many items.

Recommender systems provide various advantages for users and companies. They help users to find appropriate items and improve the overall decision-making process. As a result, they make the sales process more effective and help to boost sales revenues [66].

In general, the recommendation process of recommenders can be divided into three main phases, as shown in figure 7.

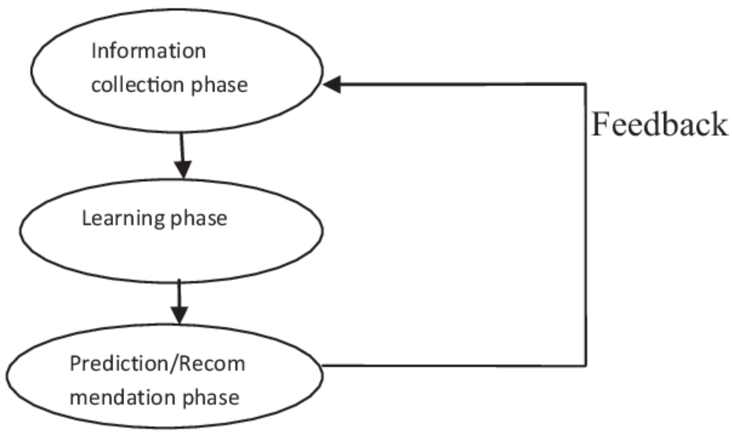


Fig. 7: Recommendation phases. [66, p. 263]

During the information collection phase, the relevant information is gathered to create a user profile, including user characteristics, user behavior, or content of the resources that the user accessed. Afterward, during the learning phase, the user profiles are used to extract relevant features which are required for the respective recommender algorithm to make predictions in the last phase. In the prediction phase, the final recommendations are made. The user behavior that results from the prediction is then again considered as feedback for future predictions. Feedback can be divided into implicit and explicit feedback. The former describes ratings that the user gives to a specific item. The more ratings are provided by the user; the higher is the accuracy of the model. However, the problem of explicit feedback is that users are not always ready to rate items as it requires more work. The latter observes the user behavior. Thus, the purchase history, clicks on the website, time spent on certain parts of the website, viewed items, and other aspects are viewed. By combining the feedback of all this information, recommendations are made. The great advantage of implicit feedback is that users are not required to make an extra effort. Therefore, implicit feedback is also less accurate [66].

To implement a recommender system, it is significant to choose a suitable recommendation filtering technique as it affects the usefulness of recommendations. Figure 8 provides an overview of the essential filtering techniques.

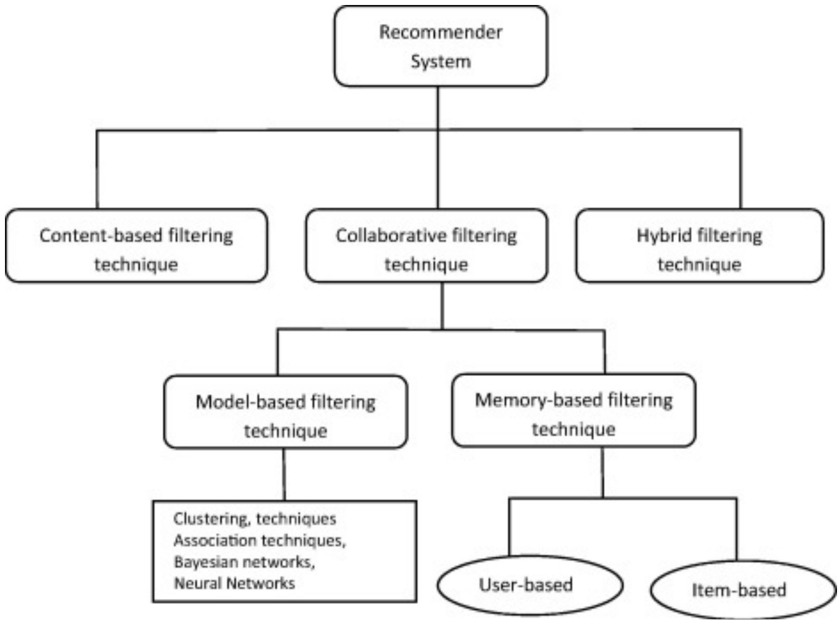


Fig. 8: Recommendation techniques. [66, p. 264]

Content-based filtering makes recommendations based on similarities in features. Features are selected based on items that a user has evaluated. Then, items are recommended that are most like the ones that have been rated positively by a user. To find the similarity between features, it uses either vector space models or probabilistic models. The main advantage of this method is that it overcomes the cold-start problem and can recommend even new items that have not been rated by users. However, the effectiveness of this method strongly depends on the availability of in-depth descriptions of item features. In addition, another issue is over-specialization which means that users mostly get only the items that are like the ones they liked in the past [66].

Collaborative filtering recommends items that a user might like based on the reaction of similar users. Thus, it creates a user-item matrix that is used to calculate similarities between user profiles to recommend relevant items. Collaborative filtering can be divided into memory-based and model-based techniques. The former techniques make predictions by identifying first users who are like the one user for which a prediction should be made. Afterward, they use the ratings of similar users to calculate a prediction for the user in question. Memory-based techniques can also be further distinguished into user-based techniques that make predictions using the similarity between users and item-based techniques that use the similarity between items. Model-based techniques use machine learning or data mining techniques to train a model based on ratings to get better recommendations. The great advantage of collaborative filtering is that it can be used whenever not much content is available related to items and if the content analysis might be complicated. Moreover, it can recommend items that are not associated with the profile of a user. However, one of the main issues is the cold-start problem which describes a situation during which recommenders cannot make recommendations as no information are available on a user or an item as they are newly added to the database [66].

Hybrid filtering uses combines content-based and collaborative filtering to make better recommendations. By taking advantage of two algorithms, one can profit from the advantages of both. For example, the combination of both ensures higher accuracy and effectiveness of the recommendations that are generated by the overall recommender. In addition, such a hybrid approach can help to overcome the disadvantages and limitations of the single algorithms. To implement such a hybrid recommender, both algorithms are implemented separately, and results are combined later. Thus, either content-based filtering is used in collaborative filtering or the other way around. However, to use such a hybrid filtering approach, it is significant to ensure that appropriate data for both algorithms are available. Thus, collaborative filtering requires data that contains users, items, and ratings that are provided by the users of the respective items. For content-based filtering, descriptive data of items needs to be ensured [66].

## 5.2. Implementation

To begin with, extensive research about recommender systems has been conducted and summarized in the last sub-chapter. The findings of the research were utilized to identify an appropriate implementation of a potential prototype. As a result, collaborative filtering and more specifically a model-based filtering approach were chosen as a recommendation technique. Collaborative filtering was selected due to the advantages that were outlined in the previous sub-chapter. Its ability to make meaningful predictions even if not a lot of content is available makes it very useful for SMEs despite its cold-start problem.

Furthermore, the prototype was created based on the idea around neural collaborative filtering described by He et al. [67]. Instead of using the inner product on the latent features of users and items, they suggest applying a neural architecture that can learn the user-item interaction function from the respective data. For this neural architecture, multiple layers are applied where the output of each layer is used as an input for the next one. The sparse user and item data are used at the beginning as inputs for the creation of the embedding layers. Due to the high number of categories involved, embedding layers are used instead of one-hot encoding [38, p. 422]. The advantage of using embedding layers is that fewer weights have to be trained due to the reduced input size leading to a computational advantage [38, p. 422]. Within the embedding layer, the spare user and item data are transformed into dense vectors. Afterward, the embedding layers are then used as an input for a multi-layer neural architecture. Each layer helps the model to learn the user-item interaction structure. Finally, the output layer results in a predicted score that represents the loss of the model. By continuously training the model through several epochs, the goal is to minimize the loss. The explained architecture can also be observed in figure 9 and follows the explanation of the respective paper. It also represents the architecture of the prototype.

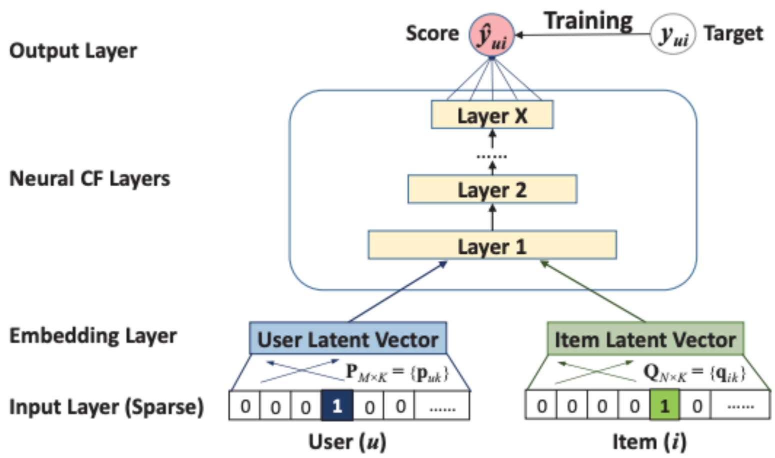


Fig. 9: Neural collaborative filtering framework. [67, p. 175]

To implement the prototype, Python was used as a programming language as it offers many valuable libraries and examples for machine learning. The neural network was built using the library Keras [3]. It is a popular library for deep learning. Other libraries that were used involve standard libraries such as numpy, pandas, matplotlib, or seaborn. Other libraries and exact imports can be viewed in more detail in the code. A summary of the applied model with the respective layers, parameters, and shapes can be observed below in figure 10.

[3] http://keras.io/

```
Model: "model_4"
____________________________________________________________________________________________________
Layer (type)                    Output Shape         Param #     Connected to
====================================================================================================
input_7 (InputLayer)            (None, 1)            0
____________________________________________________________________________________________________
input_8 (InputLayer)            (None, 1)            0
____________________________________________________________________________________________________
embedding_7 (Embedding)         (None, 1, 50)        6618400     input_7[0][0]
____________________________________________________________________________________________________
embedding_8 (Embedding)         (None, 1, 50)        486700      input_8[0][0]
____________________________________________________________________________________________________
reshape_7 (Reshape)             (None, 50)           0           embedding_7[0][0]
____________________________________________________________________________________________________
reshape_8 (Reshape)             (None, 50)           0           embedding_8[0][0]
____________________________________________________________________________________________________
concatenate_4 (Concatenate)     (None, 100)          0           reshape_7[0][0]
                                                                 reshape_8[0][0]
____________________________________________________________________________________________________
dropout_7 (Dropout)             (None, 100)          0           concatenate_4[0][0]
____________________________________________________________________________________________________
dense_7 (Dense)                 (None, 10)           1010        dropout_7[0][0]
____________________________________________________________________________________________________
activation_7 (Activation)       (None, 10)           0           dense_7[0][0]
____________________________________________________________________________________________________
dropout_8 (Dropout)             (None, 10)           0           activation_7[0][0]
____________________________________________________________________________________________________
dense_8 (Dense)                 (None, 1)            11          dropout_8[0][0]
____________________________________________________________________________________________________
activation_8 (Activation)       (None, 1)            0           dense_8[0][0]
____________________________________________________________________________________________________
lambda_4 (Lambda)               (None, 1)            0           activation_8[0][0]
====================================================================================================
Total params: 7,106,121
Trainable params: 7,106,121
Non-trainable params: 0
```

Fig. 10: Visualization of neural network.

The “Amazon Product Reviews” [4] dataset from Kaggle [5] was chosen as a data source. The data that was provided in this data set was used to train the model. The dataset contains 7,824,482 rows and includes the following four columns: userId, productId, rating, and timestamp. The timestamp column was not used in the implementation. Due to the limited available computing power, only 150,000 rows were used. It was assumed that a respective data pre-processing module was already implemented. Thus, no missing values were identified in the data set. However, the ratings are strongly skewed, with more than half of the ratings being rated with 5.0, as shown in figure 11. The mean is 4.01.

[4] https://www.kaggle.com/saurav9786/amazon-product-reviews
[5] https://www.kaggle.com/

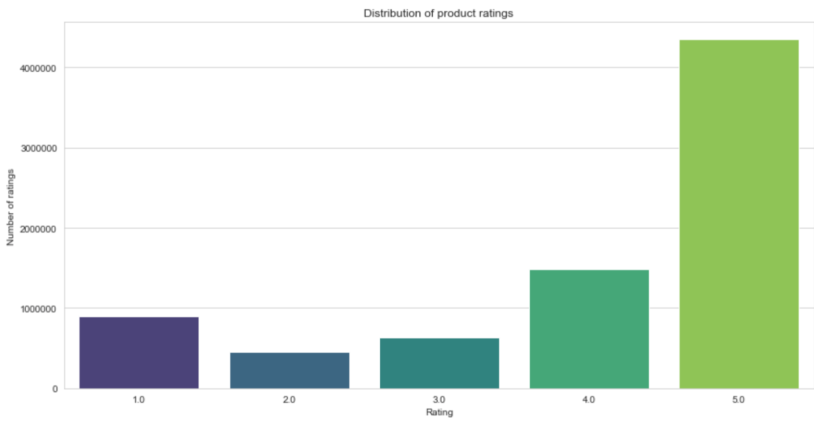


Fig. 11: Distribution of product ratings.

Moreover, there are also clear outliers, as shown in figure 12, regarding the number of products that each user rated and the number of ratings that each product received. On average, each user rated 1.86 ratings, and each product was rated 16.44 times. However, the two plots below clearly show that there are also strong outliers. Further analysis shows that the top 4.25 % of the users make more than five ratings per user. In addition, the top 6.74 % of the products receive more than 20 ratings per product.

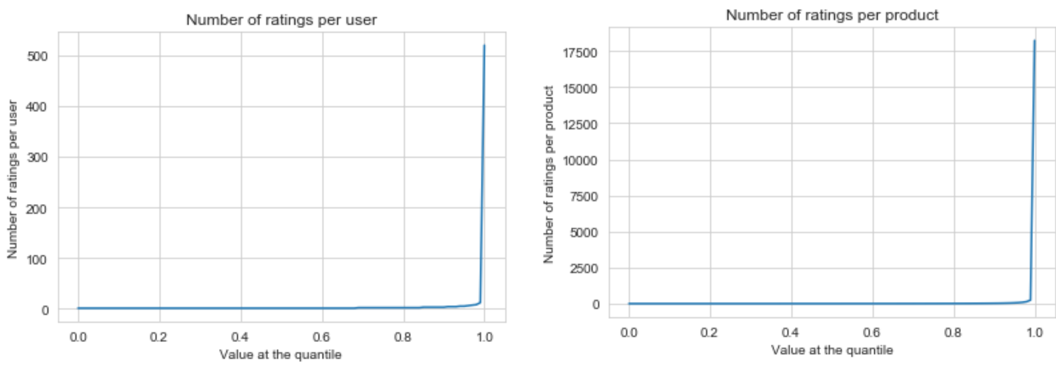


Fig. 12: Number of ratings per user/product

To measure the success of the model, the mean squared error was applied. Figure 13 shows the results that were achieved after five epochs. Overall, 135,000 data points were used to train the model and 15,000 data points to validate the model. Various learning rates have been tested. However, a learning rate of 0.00003 seemed to perform good results avoiding overfitting and underfitting. However, a popularity bias can be observed in the model as often similar

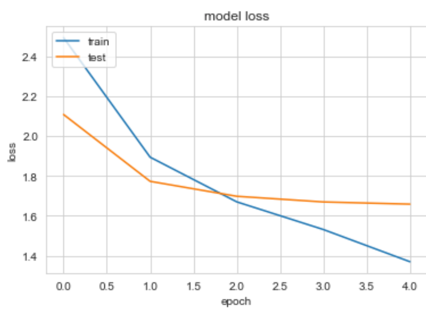


Fig. 13: Model loss

products are being recommended. Popularity bias describes the problem that a few popular items are recommended more often than most other items [68]. Overall, the entire implementation can be observed in the following GitHub repository[6]. In addition, the code has also been uploaded to the internal learning platform. A readme file has also been created and uploaded that includes information on the steps that need to be taken to run the prototype. Within the implementation, the entire code is clearly described. In addition, references to other code implementations and websites used to overcome potential issues and maximize the performance of the model have been made.

[6] https://github.com/marijand/recommender_implementation

## 5.3. Discussion and limitations

In the last sub-chapter, the assumptions of the prototype implementation were clearly described. However, some limitations must be mentioned.

To begin with, the implementation of the prototype focused only on the implementation of the recommender. It did not consider the pre-processing module. However, pre-processing is a crucial aspect of the overall machine learning process. For future research and the implementation of the respective module, it is relevant to implement an automized pre-processing module. This is especially significant for SMEs who lack the capabilities to conduct the pre-processing themselves.

Moreover, another big issue that was not tackled during the implementation was the cold-start problem. The used Kaggle data set only included data that can be used for collaborative filtering. For the further development of the module, it would be essential to implement a hybrid solution that utilizes content-based filtering to deal with the cold-start problem. Thus, it is also essential to make sure that in-depth descriptions of item features are also gathered by companies.

Furthermore, another limitation was computational power. The data set included 7,824,482 data points. However, due to the limited computational power available for testing, only 150,000 random data points were used. While testing the prototype for testing purposes at the beginning, even fewer data points were used. By increasing the number of data points, the accuracy of the model improved. Thus, ensuring enough computational power is an important aspect that needs to be considered by companies who want to engage in machine learning.

In addition, there is also a bias that needs to be considered. For example, ratings of products might be faked, which would result in lousy accuracy. Thus, wrong products might be recommended to customers. In addition, algorithm bias also must be considered. In this regard, it is complicated to understand why a neural network comes up with a specific recommendation. Therefore, some suggestions might be counter-intuitive for employees of SMEs. As a result, the explainability of the models must be considered as well. In addition, popularity bias also has not been considered appropriately, which can be observed in the recommendations that are being provided by the model.

Moreover, the second research question only dealt with the implementation of a potential prototype. As part of this research, the respective use case was not evaluated according to all the requirements that were outlined in the framework previously. Thus, further research should be conducted to cover all requirements that were outlined as part of the whole framework. In this regard, the concept of AutoML should also be incorporated to ensure that someone within an SME who does not have a strong programming background would be able to run the prototype without any problems.

Overall, the main aim of this chapter was to answer the second research question, which is concerned with identifying a potential clustering prototype as part of the previously described machine learning framework. As a result, the first sub-chapter described various aspects behind recommenders. This was used as a foundation for identifying the proper recommendation technique implemented as part of the prototype. The current sub-chapter described the assumptions behind the prototype implementation that was uploaded on GitHub. As all research objectives that were defined for the second research question have been fulfilled, the second research question is successfully completed.

# 6 Conclusion

SMEs are the backbone of the economy. Especially for them, innovation plays a crucial role in staying competitive. However, when it comes to machine learning, SMEs particularly struggle with respective implementations. The main challenges are a lack of skilled employees, limited budget, data privacy concerns, or unclear business cases to utilize machine learning. Therefore, this paper aimed to identify the main requirements for a framework that should support SMEs with the implementation of machine learning. Furthermore, it also describes the implementation of a clustering prototype as part of the overall framework.

The first research question was answered by identifying and detailing the requirements of such a framework. In addition, five different use cases were described. For each of the use cases, the following requirements need to be considered. Potential data sources need to be clearly outlined, and a high level of automation of data pre-processing needs to be ensured. Especially, the latter is highly relevant as it would help to reduce the demand for highly skilled data scientists. Another important requirement is the automized selection of the machine learning model and parameters based on the use case. Some flexibility could be provided with optional hyperparameters. In addition, fairness needs to be considered to reduce potential bias. Also, to ensure high acceptance among employees, the TAM should be considered. To deal with data security concerns of SMEs, the implementation should be focused on an on-premises solution. To ensure flexibility while developing the tool, everything should be modular. Thus, new use cases could be easily added, or existing ones could be adapted without affecting others. Finally, a GUI needs to be designed for each of the use cases and the overall tool. A potential example of such a GUI was outlined in the respective chapter. More details on each of these requirements were detailed in the respective chapter.

As part of the first research question, a product recommendation use case was outlined. It recommends products by clustering customers. This use case was the basis for the second research question. To implement this use case, a deep learning-based recommender was implemented in Python with the help of the Keras library. Therefore, various recommendation techniques, as well as their advantages and disadvantages, were outlined. Various limitations that relate to this prototype were also discussed. As collaborative filtering and, more specifically, model-based filtering was used, one big issue is the cold-start problem that could not be tackled due to missing descriptive data to build a hybrid model. In addition, the prototype also assumed that a data pre-processing module was already implemented. As such, a pre-processed data set from Kaggle was used. Other limitations, including popularity bias, lack of explainability of the model, or limited computational power were also described.

To summarize, AI and, in particular, machine learning is at the moment one of the most promising technologies for companies to improve their business processes. However, the implementation of machine learning currently focuses mainly on large enterprises. Moreover, implementing machine learning brings, apart from opportunities also many challenges. Therefore, the discussed framework that follows the idea of AutoML should help SMEs and, as a result, enhance the adoption rate of machine learning among SMEs. However, this paper is limited in its scope as it is only a bachelor's thesis. Thus, only a potential prototype of one of the use cases was discussed. Future research should be conducted to evaluate the usefulness of the framework. Thus, interviews with SMEs should be conducted to gain feedback on the proposed framework, the respective use cases, and requirements. This would help to understand how well it tackles the problems of such companies and where improvements might be required. In addition, a detailed GUI needs to be developed. Such a GUI could guide the overall implementation of the respective use cases, which also need to be developed as part of further research.

All in all, there is still a lot of research required. However, further research could help SMEs to be more competitive in an ever more digitalized world in which all businesses have to learn to operate. Furthermore, it would further democratize machine learning.

# Literature


[1] P. D. Student and M. Robu, “the Dynamic and Importance of Smes in Economy,” *USV Ann. Econ. Public Adm.*, vol. 13, no. 1, pp. 84–89, 2013.

[2] S. Radas and L. Božić, “The antecedents of SME innovativeness in an emerging transition economy,” *Technovation*, vol. 29, no. 6–7, pp. 438–450, 2009, doi: 10.1016/j.technovation.2008.12.002.

[3] N. Yoshino, “Major Challenges Facing Small and Medium-Sized Enterprises in Asia and Solutions for Mitigating Them,” *SSRN Electron. J.*, no. November 2018, 2016, doi: 10.2139/ssrn.2766242.

[4] T. M. Mitchell, *Machine Learning*, 1st ed. McGraw-Hill Science/Engineering/Math, 1997.

[5] A. Burgess, *The Executive Guide to Artificial Intelligence*. Cham: Springer International Publishing, 2018.

[6] F. Hayes-Roth, “Rule-based systems,” *Commun. ACM*, vol. 28, no. 9, pp. 921–932, 1985, doi: 10.1145/4284.4286.

[7] B. van Ginneken, “Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning,” *Radiol. Phys. Technol.*, vol. 10, no. 1, pp. 23–32, 2017, doi: 10.1007/s12194-017-0394-5.

[8] J. Manyika and B. Jacques, “The promise and challenge of the age of artificial intelligence,” *McKinsey Global Institute*, 2018. [Online]. Available: https://www.mckinsey.com/featured-insights/artificial-intelligence/the-promise-and-challenge-of-the-age-of-artificial-intelligence. [Accessed: 01-Aug-2020].

[9] N. Jain, A. Tomar, and P. K. Jana, “Novel Framework for Performance Prediction of Small and Medium Scale Enterprises: A Machine Learning Approach,” *2018 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2018*, no. December 2019, pp. 42–47, 2018, doi: 10.1109/ICACCI.2018.8554747.

[10] K. Dewalt, *Become an AI Company in 90 Days: The No-Bullshit Guide for Understanding AI, Identifying Opportunities, and Launching Your First Product*. 2018.

[11] M. Chui and S. Malhotra, “AI adoption advances, but foundational barriers remain,” *McKinsey & Company*, 2018. [Online]. Available: https://www.mckinsey.com/featured-insights/artificial-intelligence/ai-adoption-advances-but-foundational-barriers-remain. [Accessed: 01-Aug-2020].

[12] Algorithmia, “2020 State of Enterprise Machine Learning,” 2020.

[13] OECD, *OECD SME and Entrepreneurship Outlook 2019*. 2019.

[14] M. Bauer, C. Van Dinther, and D. Kiefer, “Machine Learning in SME : An Empirical Study on Enablers and Success Factors,” *Am. Conf. Inf. Syst.*, no. October, pp. 0–10, 2020.

[15] J. Jesson and F. Lacey, “How to do (or not to do) a critical literature review,” *Pharm. Educ.*, vol. 6, no. 2, pp. 139–148, 2006, doi: 10.1080/15602210600616218.

[16] D. R. Cooper and P. S. Schindler, *Business Research Methods 12th Edition*. 2014.

[17] W. Booth, G. Colomb, and J. Williams, *The craft of research*, 2nd ed. Chicago: The

University of Chicago Press, 2003.

[18] E. and Sme. Directorate-General for Internal Market, Industry, “User guide to the SME definition,” 2017.

[19] United States International Trade Commission, “Small and Medium-Sized Enterprises: Overview of Participation in US Exports,” Washington, DC, 2010.

[20] E. Cefis and O. Marsili, “Survivor: The role of innovation in firms’ survival,” *Res. Policy*, vol. 35, no. 5, pp. 626–641, 2006, doi: 10.1016/j.respol.2006.02.006.

[21] A. Werner, C. Schröder, and S. Chlosta, “Driving factors of innovation in family and non-family SMEs,” *Small Bus. Econ.*, vol. 50, no. 1, pp. 201–218, 2018, doi: 10.1007/s11187-017-9884-4.

[22] T. U. H. Nguyen, “Information technology adoption in SMEs: An integrated framework,” *Int. J. Entrep. Behav. Res.*, vol. 15, no. 2, pp. 162–186, 2009, doi: 10.1108/13552550910944566.

[23] J. Devos, H. Van Landeghem, and D. Deschoolmeester, “Rethinking IT governance for SMEs,” *Ind. Manag. Data Syst.*, vol. 112, no. 2, pp. 206–223, Mar. 2012, doi: 10.1108/02635571211204263.

[24] J. Y. . Thong, “Resource constraints and information systems implementation in Singaporean small businesses,” *Omega*, vol. 29, no. 2, pp. 143–156, Apr. 2001, doi: 10.1016/S0305-0483(00)00035-9.

[25] C. Wang, E. Walker, and J. Redmond, “Explaining the Lack of Strategic Planning in SMEs: The Importance of Owner Motivation,” *Int. J. Organ. Behav.*, vol. 12, no. 121, pp. 1–16, 2007.

[26] R. B. Robinson and J. A. Pearce, “Research Thrusts in Small Firm Strategic Planning,” *Acad. Manag. Rev.*, vol. 9, no. 1, p. 128, Jan. 1984, doi: 10.2307/258239.

[27] N. O’Regan and A. Ghobadian, “Effective strategic planning in small and medium sized firms,” *Manag. Decis.*, vol. 40, no. 7, pp. 663–671, Sep. 2002, doi: 10.1108/00251740210438490.

[28] J. McCarthy, “What is Artificial Intelligence?,” 2007. [Online]. Available: http://jmc.stanford.edu/articles/whatisai/whatisai.pdf. [Accessed: 20-Feb-2021].

[29] S. Russel and P. Norvig, *Artificial Intelligence - A Modern Approach*. Englewood Cliffs, New Jersey: Prentice Hall, 1995.

[30] N. Kühl, M. Goutier, R. Hirt, and G. Satzger, “Machine learning in artificial intelligence: Towards a common understanding,” *arXiv*, 2020, doi: 10.24251/hicss.2019.630.

[31] J. R. Searle, “Minds, brains, and programs,” *Behav. Brain Sci.*, vol. 3, pp. 417–457, 1980.

[32] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.

[33] P. Khan *et al.*, “Machine Learning and Deep Learning Approaches for Brain Disease Diagnosis: Principles and Recent Advances,” *IEEE Access*, vol. 9, no. February, pp. 37622–37655, 2021, doi: 10.1109/ACCESS.2021.3062484.

[34] M. Bojarski *et al.*, “End to End Learning for Self-Driving Cars,” 2016.

[35] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, “Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection,” *Int. J. Rob. Res.*, vol. 37, no. 5, pp. 421–436, 2018, doi: 10.1177/0278364917710318.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[37] L. Deng, G. Hinton, and B. Kingsbury, “New types of deep neural network learning for speech recognition and related applications: An overview,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 8599–8603, 2013, doi: 10.1109/ICASSP.2013.6639344.

[38] A. Géron, *Hands-on Machine Learning*, vol. 53, no. 9. 2019.

[39] X. He, K. Zhao, and X. Chu, “AutoML: A survey of the state-of-the-art,” *Knowledge-Based Syst.*, vol. 212, no. Dl, 2021, doi: 10.1016/j.knosys.2020.106622.

[40] M.-A. Zöller and M. F. Huber, “Benchmark and Survey of Automated Machine Learning Frameworks,” 2021.

[41] Q. Yao *et al.*, “Taking the human out of learning applications: A survey on automated machine learning,” *arXiv*, pp. 1–20, 2018.

[42] R. Gupta and C. Pathak, “A machine learning framework for predicting purchase by online customers based on dynamic pricing,” *Procedia Comput. Sci.*, vol. 36, no. C, pp. 599–605, 2014, doi: 10.1016/j.procs.2014.09.060.

[43] M. H. Huang and R. T. Rust, “A strategic framework for artificial intelligence in marketing,” *J. Acad. Mark. Sci.*, pp. 30–50, 2020, doi: 10.1007/s11747-020-00749-9.

[44] A. A. Mahmoud, T. Al Shawabkeh, W. A. Salameh, and I. Al Amro, “Performance Predicting in Hiring Process and Performance Appraisals Using Machine Learning,” *2019 10th Int. Conf. Inf. Commun. Syst. ICICS 2019*, pp. 110–115, 2019, doi: 10.1109/IACS.2019.8809154.

[45] F. Pittino, M. Puggl, T. Moldaschl, and C. Hirschl, “Automatic anomaly detection on in-production manufacturing machines using statistical learning methods,” *Sensors (Switzerland)*, vol. 20, no. 8, 2020, doi: 10.3390/s20082344.

[46] S. Raschka, *Python Machine learning, deep learning, scikit-learn, and tensorflow*, vol. 31, no. 3. 2017.

[47] C. Shearer, “The CRISP-DM Model: The New Blueprint for Data Mining,” *J. Data Warehous.*, vol. 5, no. 4, pp. 13–22, 2000.

[48] A. Famili, W. Shen, R. Weber, and E. Simoudis, “Data preprocessing and intelligent data analysis,” *Intell. Data Anal.*, vol. 1, no. 1–4, pp. 3–23, 1997, doi: 10.1016/s1088-467x(98)00007-9.

[49] S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*, vol. 72. Cham: Springer International Publishing, 2015.

[50] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2012.

[51] X. Yan, C. Zhang, and S. Zhang, “Toward databases mining: pre-processing collected data,” *Appl. Artif. Intell.*, vol. 17, no. 5–6, pp. 545–561, 2003, doi: 10.1080/713827171.

[52] F. Lazzeri, T. Archer, C. J. Gronlund, J. Martens, and P. Lu, “Machine Learning Algorithm Cheat Sheet for Azure Machine Learning designer,” *Microsoft*, 2020. [Online]. Available: https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-cheat-sheet. [Accessed: 28-Mar-2021].

[53] C. Ezenkwu, S. Ozuomba, and C. Kalu, “Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services,” *Int. J. Adv. Res. Artif. Intell.*, vol. 4, no. 10, pp. 40–44, 2015, doi: 10.14569/ijarai.2015.041007.

[54] A. Chouldechova and A. Roth, “The frontiers of fairness in machine learning,” *arXiv*, pp. 1–13, 2018.

[55] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” *arXiv*, 2019.

[56] S. Caton and C. Haas, “Fairness in machine learning: A survey,” *arXiv*, pp. 1–33, 2020.

[57] F. D. Davis, “Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology,” *MIS Q.*, vol. 13, no. 3, p. 319, Sep. 1989, doi: 10.2307/249008.

[58] Y. Lee, K. A. Kozar, and K. R. T. Larsen, “The Technology Acceptance Model: Past, Present, and Future,” *Commun. Assoc. Inf. Syst.*, vol. 12, no. December, 2003, doi: 10.17705/1cais.01250.

[59] M. Fishbein and I. Ajzen, *Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research*. Addison-Wesley Pub Co, 1975.

[60] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw, “User Acceptance of Computer Technology: A Comparison of Two Theoretical Models,” *Manage. Sci.*, vol. 35, no. 8, pp. 982–1003, 1989, doi: 10.1287/mnsc.35.8.982.

[61] S. F. S. Alhashmi, S. A. Salloum, and S. Abdallah, “Critical Success Factors for Implementing Artificial Intelligence (AI) Projects in Dubai Government United Arab Emirates (UAE) Health Sector: Applying the Extended Technology Acceptance Model (TAM),” *Adv. Intell. Syst. Comput.*, vol. 1058, pp. 393–405, 2020, doi: 10.1007/978-3-030-31129-2_36.

[62] E. Westra, *Modular Programming with Python*. 2016.

[63] S. Coleman, R. Göb, G. Manco, A. Pievatolo, X. Tort-Martorell, and M. S. Reis, “How Can SMEs Benefit from Big Data? Challenges and a Path Forward,” *Qual. Reliab. Eng. Int.*, vol. 32, no. 6, pp. 2151–2164, 2016, doi: 10.1002/qre.2008.

[64] S. Madhavaram and S. D. Hunt, “The service-dominant logic and a hierarchy of operant resources: Developing masterful operant resources and implications for marketing strategy,” *J. Acad. Mark. Sci.*, vol. 36, no. 1, pp. 67–82, 2008, doi: 10.1007/s11747-007-0063-z.

[65] K. Falk, *Practical Recommender Systems*. Manning Publications Co., 2019.

[66] F. O. Isinkaye, Y. O. Folajimi, and B. A. Ojokoh, “Recommendation systems: Principles, methods and evaluation,” *Egypt. Informatics J.*, vol. 16, no. 3, pp. 261–273, 2015, doi: 10.1016/j.eij.2015.06.005.

[67] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T. Chua, “Neural Collaborative Filtering,” in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 173–182, doi: 10.1145/3038912.3052569.

[68] H. Abdollahpouri, R. Burke, M. Mansoury, and B. Mobasher, “The unfairness of popularity bias in recommendation,” *arXiv*, 2019.

## Declaration in Lieu of Oath

I hereby declare, under oath, that this bachelor thesis has been my independent work and has not been aided with any prohibited means. I declare, to the best of my knowledge and belief, that all passages taken from published and unpublished sources or documents have been reproduced whether as original, slightly changed or in thought, have been mentioned as such at the corresponding places of the thesis, by citation, where the extent of the original quotes is indicated. The paper has not been submitted for evaluation to another examination authority or has been published in this form or another.

Place, Day / Month / Year Signature

Schwaz, 29 / 05 / 2021

| | |
|---|---|
| Signiert von: | Marijan Divkovic |
| Datum: | 29.05.2021 22:07:59 |

Dieses mit einer qualifizierten elektronischen Signatur versehene Dokument hat gemäß Art. 25 Abs. 2 der Verordnung (EU) Nr 910/2014 vom 23. Juli 2014 ("eIDAS-VO") die gleiche Rechtswirkung wie ein handschriftlich unterschriebenes Dokument.

**Dieses Dokument ist digital signiert!**

**Prüfinformation:**
Informationen zur Prüfung der elektronischen Signatur finden Sie unter: www.handy-signatur.at

www.a-trust.at

TRUST einfach sicher