jamk

Building Trustworthy AI

Main questions, possible solutions and a case study as example

Satu Korhonen

Master's thesis
January 2022
Artificial Intelligence
Master of Engineering in Artificial Intelligence and Data Analytics

jamk | Jyväskylän ammattikorkeakoulu
University of Applied Sciences

jamk

Description

Korhonen, Satu

Building Trustworthy AI. Main questions, some solutions and a case study as example

Jyväskylä: JAMK University of Applied Sciences, January 2022, 123 pages.

Artificial Intelligence. Master of Artificial Intelligence and Data Analytics. Master's thesis

Permission for web publication: Yes

Language of publication: English

Abstract

Recent years has seen a surge of ethical guidelines from companies and institutions concerning artificial intelligence and regulation is fast approaching. The purpose here was to take a closer look at the research done in trustworthy AI utilizing one framework, that of the EU, and develop a set of questions that would aid in developing trustworthy AI solutions with the goal to realize the potential benefits of AI while safeguarding individuals and the society against the potential issues involved. A second goal was to utilize this set of questions in developing a proof-of-concept phase execution of trustworthy AI for Aveti Learning as well as evaluate its trustworthiness and identify directions for further development. The data had problems especially in completeness, but a K-Means cluster algorithm followed by a Random Forest classifier was developed to allow for Aveti's mentors to find students in need of help. The Random Forest algorithm was deployed as a REST API app utilizing Flask. Also, security features such as a rate limiter was implemented. A failsafe method was created in case of environmental difficulties and can be incorporated into the learning platform. The questions created and adopted served to focus the development on all aspects of trustworthiness and seem to be a useful tool in creating more trustworthy AI solutions.

Keywords/tags (subjects)

Artificial Intelligence, machine learning, trustworthy AI, reliable AI, ethical AI, robust AI, safe AI

Miscellaneous (Confidential information)

All parts of this thesis are public.

Contents

Figures

Code Blocks

Abbreviations

AI	Artificial Intelligence
AIA	Artificial Intelligence Act
ALE	Accumulated Local Effects
API	Application programming interface
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
CTE	Common Table Expression
DI	Disparate Impact
DDoS	Distributed denial of service
EDA	Exploratory Data Analysis
EBM	Explainable Boosting Machines
EU	European Union
GAM	Generalised Additive Models
GA2M	Constrained Generalised Additive 2 Model
GDPR	General Data Protection Regulation
GLM	Generalised Linear Model
HTM	Human-Machine Teaming
ICE	Individual Conditional Expectation

IP	Internet Protocol
LIME	Local Interpretable Model-agnostic Explanation
ML	Machine Learning
PCA	Principal Component Analysis
PDP	Partial Dependence Plot
REST	Representational state transfer
SHAP	Shapley Additive exPlanations
SQL	Structured Query Language
SLIM	Super-sparce liner integer models0
UN	United Nations
US	United States
XAI	Explainable machine learning
XNN	Explainable neural network

1 Introduction

Powerful algorithms operating on computers capable of handling massive amounts of data are affecting more and more areas of our lives. Algorithms and automatic processes utilizing them make decisions about our loans and set insurance premiums, grant admission to universities, assign social benefits, review job applicants' resumes, detect tax evasion, money laundering, drug trafficking, smuggling, and terrorist activities as well as predict the risk of recidivism. They also steer answers to our information queries and choose what advertisements to show us. (de Laat, 2017; Kumar et al., 2020.) With the increasing commoditization of computer vision, speech recognition, machine learning and machine translation systems (Stoica et al., 2017), many tasks and processes previously performed by humans are automated thereby introducing new capabilities and functionalities that weren't previously possible (Oseni et al., 2021). AI-based technologies are becoming quite pervasive and impacting more areas of our lives (Pery et al., 2021).

The promise of artificial intelligence is substantial. According to the European Commission (2020) artificial intelligence may enable humans to develop intelligence not yet reached, opening the door to new discoveries, and helping to solve some of the world's biggest challenges from treating chronic diseases, predicting disease outbreaks, or reducing fatality rates in traffic accidents to fighting climate change, or anticipating cybersecurity threats. These technologies can bring many benefits by improving the safety of products thus making them less prone to certain risks and reducing accidents caused by human error. (European Commission, 2020a) Artificial Intelligence, or AI for short, can help promote gender balance, solve socio-economic challenges, tackle climate change, rationalize the use of natural resources, enhance our health, mobility and production processes and support how we monitor progress against sustainability and social cohesion indicators. (AI HLEG, 2019; Oseni et al., 2021) Furthermore, it can help companies make or save tremendous amounts of money while delighting customers on an unprecedented scale. (Simpson Rochwerger & Pang, 2021)

The promise of AI is therefore quite substantial. However, while these algorithms may determine our lives and change our societies (AI HLEG, 2019), their outcomes suffer from some notable defects. As Simpson Rochwerger and Pang state (2021), machine learning is an extremely powerful technology and extremely easy to use irresponsibly. de Laat (2017) describes that outcomes for the individual may be unjust or differ arbitrarily from one algorithm to the next and on a collective

level the outcomes may be biased against some group or another. Furthermore, this algorithmic decision making all too often remains opaque as the rules, explanations and clarifications of the decisions and processes are often not offered. (de Laat, 2017) This opaqueness often called a black box hide the data, algorithms, and assumptions from view. This is especially problematic when the judgements made are wrong, biased, or even destructive. As Pasquale (2015) states, faulty data, invalid assumptions, and defective models can't be corrected when they are hidden. (Pasquale, 2015)

The European Commission launched a Consultation on Artificial Intelligence in 2020, where citizens and stakeholders could provide their feedback on the topic. Of the respondents, 90% were concerned that AI may breach fundamental rights, 87% that the use of AI could lead to discriminatory outcomes, 82% that it may endanger society, 78% that the actions taken cannot be explained and 70% that AI is not always accurate. (European Commission, 2020b) When an automated decision system makes hundreds of thousands of decisions, even small mistakes can cascade into life-changing reclassifications (Pasquale, 2015). A growing body of evidence shows that AI models can embed human and societal biases and deploy them at scale and there is a growing undercurrent of pervasive distrust in AI systems (Pery et al., 2021) because, when it fails, the results can be devastating (Simpson Rochwerger & Pang, 2021).

This growing prominence of algorithmic risks can be attributed to firstly, their pervasiveness, prevalence, and integral nature to business processes across industries and functions. The algorithms, secondly, are becoming more powerful and the responsibility entrusted upon them is increasing as well. Thirdly, they are becoming opaquer, and their monitoring is increasingly hard. Finally, algorithms are also becoming targets for hacking. (Krishna et al., 2017.) These changes have been made possible by unprecedented levels of data and computation, by methodological advances in machine learning, by innovations in systems software and architectures, and by the broad accessibility of these technologies (Stoica et al., 2017).

The failure or unethical use, whether intentional or not, can result in serious repercussions for companies. Potential consequences for companies include lawsuits, regulatory fines, revenue loss, angry customers, embarrassment, reputation damage, destruction of shareholder value and the trust of its' customers. For individuals, depending on the algorithm and use case, a failure in AI can

be anything from a minor irritation up to being a life and death situation. (Ammanath, 2020; Saif & Ammanath, 2020) For societies the repercussions are no less. Furthermore, the realization of the unwanted consequences of the use of AI systems can result in an unwillingness to utilize the technology which could prevent the realization of the potentially vast benefits to society and economy. (AI HLEG, 2019)

Before going further, it is necessary to define what is meant by artificial intelligence in this thesis. Vähä-Sipilä et al. (2021) defines artificial intelligence (AI) as actions undertaken by machines that in some way resembles human intelligence. However, the artificial intelligence referred to here is a long way from human intelligence as it is not generalizable but instead focuses on solving only very narrow problems. (Vähä-Sipilä et al., 2021) The idea of machine learning (ML), as artificial intelligence is often referred as, is not new. According to Oseni et al. (2021), the first set of machine learning algorithms were introduced in the 1970s (Oseni et al., 2021). However, the rise in computational ability has brought with it the ability to crunch large amounts of data and utilize these algorithms in a way that allows these algorithms to increasingly offer effective solutions to problems previously thought to be unable to solve with computers.

Machine learning deals with narrow problems. Examples of such problems are forecasting a specific statistic, detecting anomalies such as in credit fraud, spam filtering, grouping some data into different groups et cetera. Machine learning is often divided into three different groups. The first group is supervised learning, where there is data that contains information on the label or statistic one is trying to predict. Supervised learning is further divided into regression and classification, where regression attempts to predict a statistic and classification attempts to predict a label. The second group is unsupervised learning, that is most often used to cluster data into groups or reduce dimensions in a large dataset. Finally, the third group is reinforcement learning, that is often used in games where the model learns through trial and error to succeed in a narrow task. (Oseni et al., 2021)

The fundamental components of machine learning are data, task, model, and features. Firstly, the data is possibly the most important component. For instance, to predict whether an email is spam or not, the machine learning model needs to be trained with samples of spam email and email that is not spam. The data used to train the model is often divided into train and test data. The train

data is used in the training to find a model that fits the data. Once such a model is found, the test-data is used to see whether this model can predict previously unseen data or whether it only works on data that it used for learning. The goal is for the model to work on unseen data. The second important component is the task or problem. The problem needs to be narrow enough and such that it can be solved with the data used. Otherwise, more data will need to be gathered. The third component is the model, which is a mathematic equation that best suits the data. Many models are suited to solve only a small number of tasks, but more models are developed all the time allowing for a great versatility on the choice of models and tasks. The fourth component are the features that simply are the characteristics of the data that simplify the learning of patterns between the input data and output data. These features can be discovered through statistical methods before modelling or, for instance, in deep learning they can be discovered by the neural network utilized in solving the problem. (f.ex. Oseni et al., 2021)

The performance of machine learning solutions is often determined by performance testing. For supervised learning, this can be the number of errors the model makes on unseen data. The goal is to minimize the count of these errors. (Scantamburlo et al., 2020) Most problems with AI solutions are similar to those in any other computer solution, but learning from data brings with it some novel issues, which we will go into shortly. (Vähä-Sipilä et al., 2021) For some time, the focus of AI research is to minimize these errors and the quality of the machine learning system was measured by the quality of the model's predictions (Logrén, 2020).

The incorporation of AI into large portions of human life has shown that technical performance is not enough. The adequacy of AI solutions depends, instead, on a broader set of considerations looking at different aspects of the performance in the environment they are embedded in. The focus has shifted, in the words of Scantamburlo et al. (2020) from "performance to accountability, from advances in accuracy and speed of computation to the protection of human rights and democratic values" (Scantamburlo et al., 2020). Scantamburlo et al. continue that there have been more than hundred declarations of AI principles from governments, organizations, companies, and initiatives aimed at providing normative guidance on ethical, rights-respecting and socially beneficial development and use of AI technologies. In their research, they identified eight key themes, that are privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility and promotion of human

values (Scantamburlo et al., 2020). Euijong Whang et al. (2021) identified five main topics that were fairness, robustness, explainability, transparency and accountability (Euijong Whang et al., 2021). Logrén's (2020) list of important aspects included machine learning service quality, interpretability, fairness (Logrén, 2020). Jobin et al. (2019) discovered a global convergence on five themes that were transparency, justice and fairness, non-maleficience, responsibility and privacy (Jobin et al., 2019). The wording and listed attributes of what is often referred to as trustworthy AI or reliable AI varies. Further, while there can be seen to be a cohesive idea behind these different focal points, there is substantive divergence in relation to how these principles are interpreted, why they are important, what they pertain to and how they should be implemented (Jobin et al., 2019). Further still, while research efforts into reliable AI have intensified, the practices are still often insufficient, inefficient and scattered as the practices developed do not adequately address the challenges of context-dependency, lack the ease of use and completeness, address only silo disciplines or single process steps or particular problems. (Scantamburlo et al., 2020)

Despite this, the need of trustworthiness in AI solutions is becoming critical as machine learning becomes widespread in our everyday lives. Companies such as Google, Microsoft and IBM publicly state that AI not only needs to be accurate, but also used and developed, evaluated, and monitored for trust. The model needs to be accurate, but it also needs to be checked for discrimination and altered should discrimination be found. It needs to be resilient to noisy data and be able to cope with poisoned data. The decisions made need to be understandable and explainable. Finally, it also needs to be usable. These demands not only affect the training stage, but all steps in the end-to-end machine learning pipeline including data collection, data cleaning, data validation, model training, model evaluation, model management and finally model serving and using. (Euijong Whang et al., 2021) Furthermore, being transparent about the AI solution is important as it enables identification, auditing and oversight as well as holding those responsible for account, should the need arise. (Singh et al., 2019)

These are daunting challenges, but we need AI systems that make timely and safe decisions in unpredictable environments, that are robust against sophisticated adversaries and failures, that can process ever increasing amounts of data across organizations and individuals without compromising confidentiality. (Stoica et al., 2017) AI utility needs to be balanced with the fairness and beneficial nature of the outcomes as well as other ethical and legal issues. (Pery et al., 2021)

For companies, there is a need to identify and manage AI risks effectively (Saif & Ammanath, 2020). When AI comes into play in the toolbox of the company, there is a need to review and update the risk practices that manages these new risks. Also, it may be necessary to write new policies for instance on fairness. Furthermore, there is a need for the executives especially, but also others, to understand the risks involved and how they can be minimized, managed, and monitored. To start with, companies need to develop a set of clear and consistent assessment criteria to apply to all use cases. A standard set of questions help companies understand which risk areas require more focus in each application. There should be established testing and approval processes, quality assurance metrics and regular review of AI applications' performance. Further, any variation to existing algorithms should be documented and any significant change be subject to rigorous and documented testing. (Bigham et al., 2018) Finally, there needs to be multidisciplinary dialogue and a diverse range of perspectives including different domains of expertise in this process. (Scantamburlo et al., 2020)

The purpose of this thesis is to, firstly, take a deeper look at the research and discussion on trustworthy AI and find the main questions that need answering in any or most AI solutions, although their importance will vary depending on the use case. This list of questions aims to be an aid to those developing AI solutions to enable them to think about the issues involved and find solutions that work in their specific use case. To succeed in this task, it is necessary to first look at trustworthy AI from a business viewpoint followed by focusing on the selected framework of trustworthy AI developed by the European Union. Thirdly, a deeper look is taken at the ethics and robustness and the topic of the legality of AI solutions is delimited outside the scope of this thesis. Finally, the focus is on looking at each aspect of ethicality and robustness individually and finding the issues and possible solutions in that area and, also, combining them with the machine learning solution development life cycle.

Secondly, the goal of this thesis is to build and deploy an AI solution utilizing these main questions to build as trustworthy of an AI solution as possible in the selected use case and find areas of improvement. The company chosen is Aveti Learning. The focus of this company is to bring better education to the rural and poor area of India with the help of computers. The goal is to provide personalized learning in areas such as reading, math, science, and English, that allow for the students

to reach new heights previously thought to be unattainable to them due to hardship in their situation. The chosen machine learning problem was to identify those students that firstly, need most help, and secondly, are most likely to benefit from extra help.

Finally, in conclusion of this introduction, a quote from the trustworthy AI guidelines of the European Union to apply to this thesis as well:

> *Nothing in this document shall be construed or interpreted as providing legal advice or guidance concerning how compliance with any applicable existing legal norms and requirements can be achieved. Nothing in this document shall create legal rights nor impose legal obligations towards third parties. We however recall that it is the duty of any natural or legal person to comply with laws – whether applicable today or adopted in the future according to the development of AI.*

This thesis proceeds on the assumption that all legal rights and obligations that apply to the processes and activities involved in developing, deploying and using AI systems must be duly observed. (AI HLEG, 2019)

2 Trustworthy AI from a business viewpoint

2.1 From silos to data ecosystems

In as much as there are failures and problems with the use of machine learning algorithms in business solutions, Bhattacharya states (2020) that they still play a vital role in the development of cognition and the provision of solutions to business problems, processes, and decision-making (Bhattacharya, 2020). Data is combined from different systems to data warehouses and datalakes. It is aggregated and curated and then used as material for developing AI systems. Data is also shared across companies, as businesses leverage third-party data to augment their AI-powered services. This all is a step away from data silos, where each system and each company uses these siloed data solutions, to a data ecosystem, where AI solutions learn and make decisions using data owned by different organizations creating outputs that can be new data assets, or heavily data-influenced assets, including machine learning models, which are used to improve or automate decision making either within the organization responsible for the data creation or shared with a

partner or third party organization. (Barclay et al., 2019; Stoica et al., 2017) Such solutions can be for instance predicting quality issues and failures, or developing targeted marketing campaigns, or supporting workforce planning, as Krishna et al. (2017) depict in Figure 1 below.

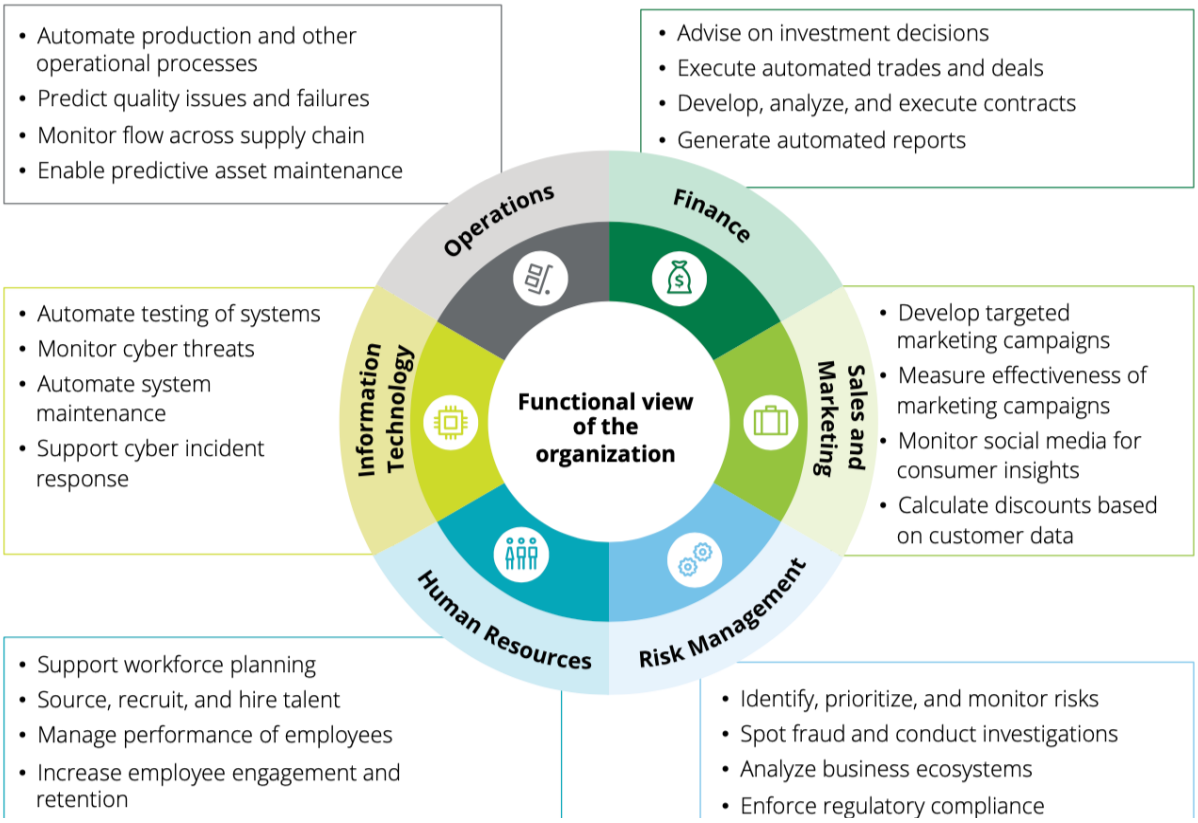


Figure 1. Algorithm use across business functions. Source: Krishna et al. 2017

The

as-

ns

need to be developed and designed in a way that the machine learning models can be trained on datasets owned by different business functions or companies without compromising their confidentiality. Further, the data systems need to be able to demonstrate the provenance and authenticity of the data and knowledge used. (Barclay et al., 2019; Stoica et al., 2017)

2.2 Risks involved

There has been several of high-profile incidents reflecting the risk and the difficulty in detecting bias and unfairness in machine learning models. Kumar et al. (2020) lists Tay, the Microsoft bot that turned racist and was shut down within 16 hours of launch, AI-based hiring tool developed and used by Amazon that was clearly biased against women, and Apple Credit Card that offered smaller lines of credit to women than men. (Kumar et al., 2020) Facebook's experimentation with

its algorithm have helped to amplify fake stories and allowed rumors to spark violence in Sri Lanka, Myanmar and the Philippines (Stevenson, 2018). One lesson to be learned from these is that if these can happen to the biggest technology firms in the world, there is a risk to every institution in ignoring the risk of bias in AI solutions. It is also possible, as Kumar et al. (2020) state, that there might be several issues in machine learning models that have gone unnoticed for some time. These issues bring with them several risks to companies, the users of the solutions and to societies as a whole. (Kumar et al., 2020) As these machine learning algorithms operate at faster speeds in fully automated environments, they can become increasingly volatile as algorithms interact with other algorithms and the risks can quickly get out of hand (Krishna et al., 2017).

The immediate fallout of these algorithmic risks can include inappropriate and potentially illegal decisions. For instance, in Finance, inaccurate financial reporting can result in regulatory penalties and shareholder backlash, or in Sales and Marketing where algorithms can discriminate against certain groups of customers in product pricing, offerings and ratings. Algorithmic risks to organizations are for instance reputation risks, which can actualize if the various stakeholders believe that the workings of the algorithm aren't aligned to the ethics and values of the organization or if the algorithm is designed to covertly manipulate any shareholder group or regulators. Another risk is a financial risk, which can actualize in faulty strategic decision making due to faulty algorithmic suggestions. This can also lead to strategic risks that can leave a company at a competitive disadvantage. Thirdly, there are operational risks especially when automating supply chain and other operational areas, where errors can result in significant operational disruption. Fourth type of risk are regulatory risks, which can actualize when algorithms make decisions that violate the law, circumvent existing rules and regulations, or discriminate against any group. Finally, there are also technology risks as the wide-scale use of advanced algorithms can open new points of vulnerability for IT infrastructure. (Kumar et al., 2020)

Financial risk is also present in developing AI solutions as the R&D needed can result in a solution that cannot be put into production. Simpson Rochwerger and Pang (2021) state that only 20% of AI pilots make it to production. The other 80% fail. Reasons behind these failures include not picking the right problem, not having a clear strategy, not having the right team, not creating a sustainable data infrastructure, or neglecting security or ethical considerations. Also, pilots can fail if their success isn't measurable, or their goals are not realistic or achievable, or don't address a

business need directly. After the adoption of best practices in these areas, Simpson Rochwerger and Pang have been able to increase their percentage from 20% to 67%. (Simpson Rochwerger & Pang, 2021)

Many of the risks in machine learning systems come down to the point that it is difficult to explain the logic behind the decisions or predictions. The decision boundary of these algorithms is often so complicated and multifaceted, that human intuition cannot comprehend the logic and math involved. It is also quite easy to create adversarial examples that cause them to make wrong or unexpected decisions. The systemic risks come down to the point where the machine learning model interacts with other systems. These interfaces are often quite complicated, and each contributor are difficult to analyze and test thoroughly for issues. (Vähä-Sipilä et al., 2021)

Further, Barclay et al. (2019) discuss a disconnect that can result from increased adoption and deployment of machine learning models into business, healthcare, and other organizational processes. The disconnect is between the developers of the solutions and other stakeholders and Barclay et al. claim that it is inevitable as the models begin to be used over several years or are shared among third parties and it will become increasingly difficult for users to maintain ongoing insight into the suitability of the parties who created the model, or the data that was used to train it, or in fact the method and logic of operation of the algorithm. This is especially problematic then regulations change, and once acceptable standards become outdated, or data sources discredited as biased or corrupted. (Barclay et al., 2019) Without appropriate insight and governance, it can be quite difficult to manage the risks involved.

2.3 Risk management in the AI era

As the risks are quite clear, it is imperative that they are appropriately and effectively managed. As Krishna et al. (2017) state, only then can an organization harness the power of these algorithms to expand its value proposition and bring added efficiency and effectiveness to the development and delivery of products and services in the marketplace. By effectively managing their risks, organizations can leverage this technology to accelerate corporate performance. (Krishna et al., 2017)

Risk management is not new to businesses. However, the risk practices need a review, and update to manage the risks of AI solutions. Also, some risk management practices need to happen at more

frequent intervals. As Bingham et al. state (2018), it is more a matter of enhancing the existing processes to consider the new challenges and fill the necessary gaps. The risk assessment needs to be revisited periodically to assess whether the risk profile of an AI solution has changed. Further, there should be an established and documented testing and approval process, quality assurance metrics and regular review of AI applications' performance and all algorithms should be subject to periodic re-validation. Furthermore, firms should have a clear and full overview of all AI applications deployed throughout their organization including their relevant owners and the key compliance and risk controls in place. Also, any significant change in the solutions should be rigorously tested and the process documented. (Bigham et al., 2018).

Krishna et al. (2017) have created a framework for algorithmic risk management as seen in Figure 2 below. There are three different components. First, there is strategy and governance, which requires companies to create an algorithmic risk management strategy and governance structure to manage technical and cultural risks. The components of this are the principles, policies and standards involved, the roles and responsibilities of each party, the control processes, and procedures. Further, the appropriate personnel selection and training should be documented. Finally, providing transparency and processes to handle inquiries and problem reports can help organizations use algorithms responsibly. The second aspect is the design, development, deployment, and use of algorithms. This entails that the previously stated governance structure and principles and policies guide each stage of the developmental life cycle from data selection to algorithm design, to integration to other systems and to actual live use in production. The third aspect of the framework is monitoring and testing where data inputs, their workings and outputs need to be assessed. Also, objective reviews are suggested by internal and external parties. (Krishna et al., 2017)

Simpson Rochwerger and Pang (2021) also suggest being able to calculate a baseline performance This type of framework is needed for each problem that is approached with AI methods. This describes the performance currently. This allows the calculation of the return of investment of each solution. Further, it is necessary to define what success looks like before starting development. The metric chosen, that is optimized in the algorithm development and training, needs to be chosen beforehand and the desired level of this metric needs to be chosen so that the AI solution out-

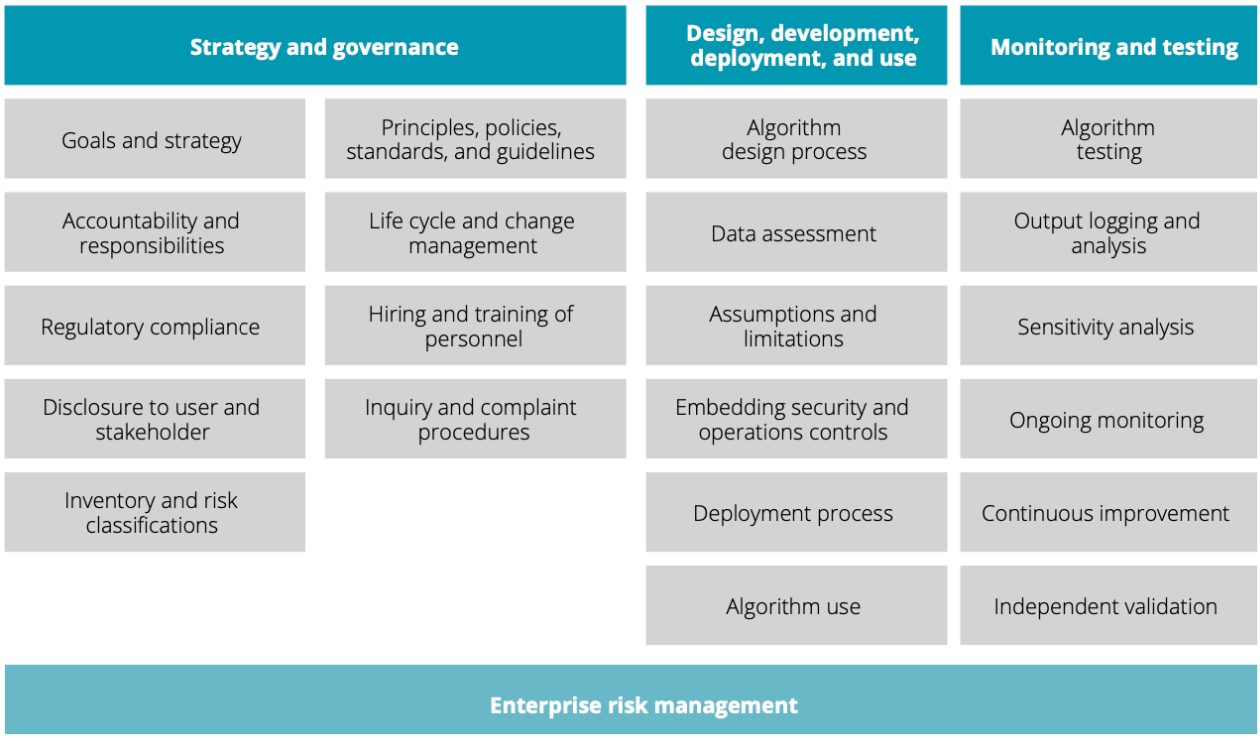


Figure 2. Framework for algorithmic risk management. Source: Krishna et al. 2017

solution to have the best possible chance of success, connect the dots between the output of the machine learning project and business value. Further, building AI systems responsibly and with good data management from the beginning, creates machine learning systems that are both more adaptable and more successful over time. (Simpson Rochwerger & Pang, 2021)

2.4 Regulation is coming

There is already regulation in place that supports the trustworthiness of AI solutions. Regulation and laws concerning product safety or liability as well as discrimination is in effect and already regulates AI solutions as well as any other product or solution. (AI HLEG, 2019) Also, the European Union's General Data Protection Regulation (GDPR) and the US government's Algorithmic Accountability Act have affect what can be done and how (Rabiul Islam et al., 2021). However, it is reasonable to expect that the level of scrutiny and regulation will increase in the future (Bigham et al., 2018) and will cover AI solutions no matter of the technology in which they were created (Vähä-Sipilä et al., 2021).

On April 20th 2021 the European Commission released the proposal for the regulation of artificial intelligence. The proposed Artificial Intelligence Act (AIA) takes a risk-based approach to regulating

AI by focusing on the use cases and categorizing them into three categories based on a combination of factors that include the intended purpose, the number of impacted persons, and the potential risk of harms. Based on these factors, systems that use subliminal techniques that cause physiological or psychological harm, exploit vulnerable groups, effectuate social scoring by public authorities that may result in discrimination or unfavorable treatment, and remote biometric systems used by law enforcement in public areas, subject to well-defined exceptions, are prohibited. The second group is considered high risk, which are solutions used in critical infrastructure such as education, human resources, essential private and public services, law enforcement, migration, asylum and border control management, and administration of justice and democratic processes. Lastly, all other uses are considered low risk. (Pery et al., 2021)

The proposed AIA would apply to all providers. Responsibility is assigned to users, importers, distributors, and operators who make use of or make substantial modifications to the functionality and performance of AI systems and cover all use where the system users are in the EU or the output of the systems is used in the EU.(Pery et al., 2021)

The AIA sets forth a comprehensive legislative mandate to ensure fairness in the application of AI systems that safeguard fundamental human values and promotes socio-economic rights, such as obligation to implement appropriate risk management throughout the entire lifecycle of AI systems, rigorous data governance processes, technical documentation and record-keeping processes to enable monitoring of compliance, transparency that enables full interpretation of outputs and human-in-the-loop oversight. Further, all systems will be required to implement a range of processes to ensure full transparency into and accountability for AI systems such as conformity assessment and certification processes, auditability including accessible event logs and explainability. Compliance with AIA constitutes several interdependent steps. In step one, R&D teams develop and bring to market AI systems in accordance with the risk classification system. If the solution is considered high-risk, then a priori conformance assessment must be undertaken before the solution may be placed on the market. In step two, legal and compliance teams must institute compliance measures in accordance with the proposed regulation to ensure adherence to data governance, accountability, transparency, robustness, and cybersecurity provisions. In step three, data science teams must undertake continuous monitoring of AI systems and collect data on the system's operation and take corrective action if needed. In step four, customer-facing functions

are responsible for providing clarity and certainty as to the expected AI system inputs and outputs in a way that users are informed that they are interacting with an AI system, augmented with human oversight, who monitors operation and can override and reverse the output, if needed. Finally, in step five, auditable and traceable documentation is created about data procedures for data management, analysis, labeling, storage, aggregation, retention, and, also, of serious incidents. (Pery et al., 2021)

The AIA incorporates an enforcement mechanism that surpasses the fines under the GDPR. For instance, for the supply of incorrect, incomplete, or misleading information to the authorities, fines up to 10 million euros or 2% of the total worldwide annual turnover can be given. For non-compliance with any other AIA requirement or obligation, fines up to 20 million euros or 4% of the total worldwide turnover can be given. Finally, fines up to 30 million euros or 6% of the total worldwide annual turnover can be given for violations of prohibited practices. (Pery et al., 2021)

While this chapter has focused on the EU, regulation is on the horizon for the US (Vought, 2020) and China as well. China, for instance, has proposed regulating recommender algorithms. The provisions apply to search filters and personalized recommendation algorithms such as used in social media feeds, content services and online stores. It also regulates dispatching and decision-making algorithms such as used in gig work platforms, and generative or synthetic-type algorithms used in content generation. (Sapni & Mihir, 2021)

3 The concept of trustworthy AI and steps to achieve it based on previous research

3.1 In search for trustworthiness: EU framework for Trustworthy AI

The goal of The EU framework for Trustworthy AI is to support AI development, deployment, and use in a way that ensures that everyone can thrive in an AI-based world and to build a better future while being globally competitive. The decisions made by AI systems need to be in line with human rights and democratic values that should not be compromised, and the systems need to be

able to act accordingly and have suitable accountability processes in place to ensure this. Development, deployment, and use, as well as the competitive edge, is supported by embedding Trustworthy AI in solutions. The framework consists of three aspects, which are lawful, ethical and robust AI. (AI HLEG, 2019)

The framework is based on the fundamental rights enshrined in the EU treaties, the EU Charter and international human rights law. These rights can be understood as rooted in respect for human dignity and referenced by freedom of the individual, equality and solidarity, citizens' rights, and respect for democracy, justice, and the rule of law. AI systems can enable the fulfillment of these rights. For instance, they can help individuals track their personal data or increase access to education. AI solutions can also hamper fundamental rights by, for instance, discriminating against vulnerable groups. Second important aspect is the need to pay attention to vulnerable groups, such as children, persons with disabilities etc., or situations characterized by the asymmetry of power, such as between employers and employees, or between businesses and customers. Thirdly the basis of the framework is the acknowledgement that while having the possibility to bring substantial benefits to individuals and societies, AI systems also pose risks and possible negative effects. Hence, all adequate measures should be adopted to mitigate these risks when appropriate, proportionately to the magnitude of the risk, and throughout the AI solutions' lifecycle. (AI HLEG, 2019)

From the afore mentioned fundamental rights arise four ethical principles, or ethical imperatives, that must be respected in the development, deployment, and use of AI systems. These imperatives are respect for human autonomy, prevention of harm, fairness, and explicability. Firstly, human autonomy, refers to the need for humans interacting with AI systems to maintain and even augment and enhance full and effective self-determination over themselves, their cognitive, social, and cultural skills, and be able to partake in the democratic process. AI systems should not unjustifiably subordinate, coerce, deceive, or manipulate humans. This also means securing human oversight in work processes of AI systems. The second ethical imperative is prevention from harm meaning that AI systems should never cause or exacerbate harm or adversely affect human beings. This entails the protection of human dignity as well as mental and physical integrity and the protection of the natural environment, and all living beings. This means that AI systems, and the environments they operate in, must be safe and secure. They need to be technically robust and

not open to malicious use. The third ethical imperative is fairness. This means the equitable and just distribution of both benefits and costs. It also means ensuring that individuals and groups are free from unfair bias, discrimination, and stigmatization. If possible, AI systems should increase social fairness. Additionally, fairness implies that AI practitioners should respect the principle of proportionality between means and ends and consider carefully on how to balance competing interests and objectives. Finally, a dimension of fairness is the ability to contest and seek effective redress against decisions made by AI systems. For this to be possible, the entity accountable for the decisions must be identifiable, and the decision-making process explicable. The fourth, and final, ethical imperative is explicability, which is seen to be crucial for building and maintaining users' trust in AI systems. This means that the processes need to be transparent, the capabilities and purpose of AI systems is openly communicated, and decisions explained to those directly and indirectly affected. Without these, the decisions cannot be effectively contested, which is necessary. If an explanation is not possible, such as in so called black box methods, other explicability measures such as traceability, auditability, and transparent communication on system capabilities may be required. The degree to which explainability, as well as the other ethical imperatives, is needed is highly dependent on the context and severity of the consequences if the output is erroneous or otherwise inaccurate. (AI HLEG, 2019)

These four ethical principles are to be realized by seven key requirements for Trustworthy AI. The way in which AI solutions adhere to these should be transparently and clearly communicated with stakeholders enabling realistic expectations about the capabilities and limitations of the system. Also, any fundamental tensions existing between these requirements, the solutions to them, and any trade-offs done, should be documented, and communicated. These requirements pertain throughout the AI system's entire life cycle and their importance depends on the specific use case. (AI HLEG, 2019)

The first requirement is human agency and oversight. This means that users should be able to make informed autonomous decisions regarding AI systems, be informed that they are interacting with one, and be given tools and knowledge to comprehend the AI system to a satisfactory degree and be able to interact with it, as well as able to challenge the system. Human oversight is needed to ensure that the system does not undermine human autonomy or cause other adverse effects

and the more autonomous the system, the stricter oversight and governance is needed. The second requirement is technical robustness and safety, which means that the system is developed with a preventative approach to risks and in a way that they reliably behave as intended, fail in a controllable and predictable manner for instance in case of a cyberattack, resume operations after a forced shut-down, minimize unintentional and unexpected harm, and prevent unacceptable harm and produce reproducible results with a range of inputs and in a range of solutions. They also need to make correct judgements as well as indicate the level of uncertainty of the decision. (AI HLEG, 2019)

The third requirement concerns privacy and data protection, which must be guaranteed throughout the AI systems life cycle including both the initial data as well as the data provided during the use of the system. Also, it must be ensured that the data will not be used unlawfully. The data needs to be correct and of high quality as that is paramount to the performance of the AI system. If the gathered data contains biases, inaccuracies, or errors, these need to be addressed prior to training the machine learning model. The processes and data sets used must be tested and documented at each step and access to data needs to be governed. (AI HLEG, 2019) Special care needs to be taken when combining data from separate sources as the consent given to each individual data source may not cover the result as the type and depth of the combined data may be entirely unpredictable for the data subject and infringe upon data protection and privacy. (Hildebrandt, 2020; The Committee of experts on internet intermediaries (MSI-NET), 2018)

The fourth requirement is transparency, which is divided into traceability, explainability, and communication. Traceability means that the data sets and processes leading to the AI system's decision, as well as the decision or prediction itself, should be documented to the best possible standards. This increases transparency, explainability, and auditability. Explainability concerns the ability to explain both the technical processes of an AI system and the related human decisions in a way that can be understood and traced by human beings. Such an explanation should be timely and adapted to the expertise of the stakeholder concerned. Finally, a description of the AI system, its level of accuracy as well as limitations, should be communicated to stakeholders. Individuals

should be made aware when they are interacting with an AI system and have the option to choose human interaction instead. (AI HLEG, 2019)

The fifth requirement is diversity, non-discrimination, and fairness, which means that identifiable and discriminatory bias should be removed in the data collection phase where possible and/or handled in later stages of the AI development cycle. (AI HLEG, 2019) A simple solution would be to remove, or not even collect, variables referring to protected aspects such as gender or ethnicity. This solution, however, creates an issue as it is then very difficult to check the data for biases (Criado Perez, 2019). Further complexities come from proxies. While no variable in the data in and of itself describes the protected aspects, there may be proxies that correlate very highly with them and sustain the discriminatory pattern (Hildebrandt, 2020; The Committee of experts on internet intermediaries (MSI-NET), 2018). In any case, discrimination based on age, origin, nationality, citizenship, language, religion, opinion, political activity, family, health, disability, sexual preference and other personal attributes is strictly prohibited (Yhdenvertaisuuslaki, 2015). In some cases, positive discrimination is allowed, if the goal is to prevent or compensate for disadvantage (Neuvoston Direktiivi 2000/43/EY, 2000). Stakeholder participation in all stages is highly encouraged as well as hiring from diverse backgrounds. Especially in business-to-consumer domains, systems should be user-centric and designed in a way that allows all people to use them regardless of age, gender, abilities, or other characteristics. (AI HLEG, 2019)

The sixth requirement is societal and environmental well-being. AI systems should benefit all human beings, the broader society, other sentient beings, and the environment. Sustainability and environmental friendliness, as well as social impact, should be assessed and monitored. Also, the impact from a societal perspective, on institutions and democracy, should be assessed. The seventh, and final, requirement is accountability. This means that the complete AI system, the model, the data, and the design processes, is auditable. An impact assessment both prior to and during the life cycle can be helpful and should be in proportion to the risk the AI system poses. All trade-offs made with these requirements should be addressed and properly documented and accessible

methods should be created to ensure adequate redress should things go wrong. (AI HLEG, 2019) A summary image of the framework can be found in Appendix 1.

3.2 Lawful AI

AI systems do not operate in a lawless world and several legally binding rules at European, national, and international level already apply or are relevant to the development, deployment, and use of AI systems today. This area in its breadth is outside the boundaries of this thesis. However, a few words about the topic are in order.

AI HLEG (2019) lists the following legal sources, which are to be considered when designing an AI system. These are: EU primary law namely the Treaties of the European Union and its Charter of Fundamental Rights, EU secondary law such as the General Data Protection Regulation (GDPR), the Product Liability Directive, the Regulation on the Free Flow of Non-Personal Data, anti-discrimination Directives, consumer law and Safety and Health at Work Directives, the UN Human Rights treaties and the Council of Europe conventions such as the European Convention on Human Rights, and numerous EU Member State laws. Besides horizontally applicable rules, various domain-specific rules exist that apply to AI applications. An example of such domain-specific rules is for instance the Medical Device Regulation in the healthcare sector. (AI HLEG, 2019)

3.3 Ethical AI

3.3.1 Basics of AI ethics

The field of AI ethics has emerged largely due to concerns from individuals, society and industry as it is recognized that things could go really wrong if AI is implemented without due regard and consideration for its potentially harmful impacts (Leslie, 2019; Rességuier & Rodrigues, 2020). The principal motivation that has driven the development of applied AI ethics can be deduced from the definition of artificial intelligence. Leslie (2019) quotes Marvin Minsky, an AI pioneer, that defined

AI as follows: “Artificial Intelligence is the science of making computers do things that require intelligence when done by humans”. When humans do things that require intelligence, such as decide on loan premiums or recommendations, we, according to Leslie, hold them responsible for the accuracy, reliability, and soundness of their judgements, and that these decisions are fair and reasonable. However, an algorithmic process can neither be directly responsible nor immediately accountable for the consequences of their actions as they are not morally accountable agents. (Leslie, 2019) The Committee of experts on internet intermediaries (2018) ask, who, then is responsible when for instance human rights are infringed by an AI solution? Is it the person who programmed the algorithm, the operator of the algorithm, the company? (The Committee of experts on internet intermediaries (MSI-NET), 2018). This is the driving motivation behind AI ethics.

Rességuier and Rodrigues (2020) argue that AI ethics, as it is currently used, and sometimes accused of, as toothless, is used as a softer version of the law and misused as a replacement for regulation. They further argue that Silicon Valley support the development of AI ethics as a way of avoiding legally enforceable restrictions of controversial technologies. (Rességuier & Rodrigues, 2020) Paul Nemitz, a principal advisor in the European Commission and a Member of the Data Ethics Commission of the German Government and of the World Council on extended Intelligence, is on the same lines. He states that big companies invest in ethical artificial intelligence and in self-regulation to avoid regulation but that is not enough as ethics code lacks democratic legitimacy and cannot be enforced. He continues to argue that AI cannot and will not serve the public good without strong rules in place. (Nemitz, 2018).

Using ethics to avoid regulation, however, is not what ethics is for. The objective of ethics is not to impose particular behaviors, and to ensure that these are complied with. That is the role of regulation. Rességuier and Rodrigues (2020) argue that ethics is primarily a form of attention, a continuously refreshed and agile attention to reality as it evolves and as such is a powerful tool against cognitive and perceptive inertia that hinders us from seeing what is different from before or in different contexts and calls for a change in behavior, such as new regulation. It helps us to notice small changes as they unfold. In the field of AI, they continue, these changes are for instance increasing dependency on technology, the deployment of biased systems that lead to discrimination towards women and minorities, and deepening surveillance of governments and private companies. Ethics helps us look at concretely at how the world changes and to see if some developments

need to be resisted, or regulated, when their negative impacts outweigh their benefit. Ethics is about watchfulness and investigation, or digging behind what seems to be settled, and questioning what may be obvious. This way we may make sure that the systems we deploy, and use, do not go against dearly held norms and values. (Rességuier & Rodrigues, 2020) It means questions about, for instance, whose well-being is being optimized for and by which actors, what is meant by fairness and for whom, what bias is deemed as necessary to fix, what measures users will get to control the data they share, and when does a dataset need to be augmented and how to create a more balanced dataset (f.ex. Jobin et al., 2019; Rességuier & Rodrigues, 2020). This chapter focuses on questions of specific importance in designing, developing, and deploying AI solutions so that their benefits will outweigh their potential harms.

3.3.2 Transparency, Interpretability, Explainability

Transparency, interpretability and explainability are all interconnected aspects that focus on the ability to understand the AI system by different stakeholders. While the literature often uses interpretability and explainability interchangeably, and sometimes combines them with transparency, in this thesis the terms are defined as follows. Interpretability of AI provides insight into the process between inputs and outputs and helps to understand how and why those inputs become outputs. Explainability goes one step further by providing different stakeholder groups a view into that understanding. It also allows a wider perspective into the data gathering, labeling, processing, and deploying, while interpretability focuses more often on the model, and its' internal workings, itself. Finally, transparency of an AI system is the degree to which these processes and understandings are shared with stakeholders. All of these are connected with the ability to justify decisions, understanding how an AI system reached its conclusion, making sure it does what it is supposed to do, understanding what it does when confronted with unfamiliar circumstances and anomalies, and key to building trust between the AI solution and humans (Jansen Ferreira & Monteiro, 2021; Leslie, 2019; Pery et al., 2021; Rabiul Islam et al., 2021; Rosenfeld & Richardson, 2019). We will first focus on interpretability, secondly on explainability, and thirdly on transparency.

There are many ways to create an interpretable AI system, where it is possible to understand what is does. One approach is to use inherently interpretable models. Interpretable models we will briefly look at from the viewpoint of interpretability are linear and logistic regression, as well as

generalized linear models and generalized additive models, decision tree-based models, decision rules, naïve Bayes classifier and K-nearest neighbor. (Leslie, 2019; Molnar, 2021; Rabiul Islam et al., 2021) More information on each of these can be found both online and in nearly all machine learning textbooks. The list of interpretability methods covered in this thesis is not all-encompassing as the research and development of interpretable methods is ongoing.

In linear regression, the predicted target consists of the weighted sum of input features. The weights offer a medium of explaining the importance of each feature when the number of features is small. Logistic regression, as an extension of linear regression to classification problems, models the probabilities for classification tasks. Probability is given as a number between 0 and 1, where the weight provides an indication of the direction of the influence and the factor of influence between classes. Generalized linear models (GLMs) and generalized additive models (GAMs) help in situations where the target outcome does not follow a Gaussian distribution or there is interaction between features. Both situations are problematic for linear and logistic regression. However, these models are more complex due to the added interaction and less interpretable. (Molnar, 2021; Rabiul Islam et al., 2021)

Decision tree-based models split the data multiple times based on a cutoff threshold at each node until it reaches a leaf node. It works even when the relationship between input and output is not linear and when there is interaction between features. It is also quite interpretable as the path from the root node to the leaf node tells how the decision took place. However, slight changes in input can have a big impact on the predicted output and multiple different kinds of trees can be developed for the same problem. Also, the mode nodes or depth of the tree, the more challenging it becomes to interpret. Decision rules are simple if-then-else conditions that are straightforward to interpret, but mostly limited to classification problems and inadequate in describing a linear relationship. Naïve Bayes classifier is based on the Bayes Theorem, where the probability of classes for each of the features is calculated independently and assumes strong feature independence. K-nearest neighbor uses the k-amount of nearest neighboring data points for prediction and interpretability can be sought at looking at the nearest data points. (Molnar, 2021; Rabiul Islam et al., 2021) Interpretable models are also continuously developed. Hall et al. (2019) mention newer, highly interpretable machine learning modeling techniques such as explainable neural network

(XNNs), explainable boosting machines (EBMs, GA2Ms), monotonically constrained GBMs, scalable Bayesian rule lists and super-sparse linear integer models (SLIMs). (Hall et al., 2019)

If using these models does not offer a sufficiently high degree of accuracy, or they are otherwise unsuited to the problem, or, indeed, the data is sufficiently complex that these models become difficult to understand, another possibility is to use model-agnostic interpretation methods. They also help in situations where feature transformations are used, which can diminish the interpretability of inherently interpretable models. They are of specific importance in domains where trust, user-confidence and public acceptance are critical for the realization of optimal outcomes. These model-agnostic methods that are used post hoc, after the modeling, can also be used in combination with interpretable models if there is need to add different viewpoints to the interpretation. With model-agnostic methods the developers are free to use any model or combination of models they choose and apply the explanation of their choice. Hence, these offer quite a bit of flexibility. They offer a way to peer into the black box and reverse engineer explanatory insight. However, they can fail to accurately represent certain areas of the model's feature space. These methods are divided into local and global methods. Local methods enable the interpretability of individual cases by focusing on single data points, or neighborhoods in its feature space, or smaller sections of the model. Global methods explain the model's behavior on the entire dataset and across predictions or classifications. There is, however, a tradeoff between the need for a global explanatory model to be simple to be understandable and complex to capture the intricacies of the mapping function of inputs to outputs. (Leslie, 2019; Molnar, 2021; Rabiul Islam et al., 2021)

First three model-agnostic methods are PDP, ICE plots and ALE plots. Partial Dependence Plot (PDP) is a global method that works when the number of features is two and they are independent of each other. It can show whether the relationship between the target and a feature is linear, monotonic, or more complex. It is intuitive and easy to understand. Individual Conditional Expectation (ICE) is a local method that focuses on each instance in the dataset and shows how the instance's prediction changes when a feature changes. PDP is basically the average of all the lines of an ICE plot. It is possible to vary the feature of interest and the plot is very intuitive. Also, they can uncover heterogenous relationships. The disadvantages are that they can focus on only one feature at a time and the plots can become overcrowded. Accumulated Local Effects (ALE) is also a

global method and describes how features influence a prediction on average. It works by highlighting the effects of specific features on the prediction by partially isolating the effects of other features. The plot works also when the features are correlated. The ALE plots are centered at zero which makes their interpretation easy and conditional on a given variable. While ALE plot works on correlated features, it can be difficult to read. (Molnar, 2021)

Another local interpretative strategy seeks to explain feature importance in a single prediction or classification by perturbing input variables. The Local Interpretable Model-agnostic Explanation (LIME) works by fitting an interpretable model to a specific prediction or classification by sampling data points at random around the target and then using them to build a local approximation of the decision boundary that can account for the features which figure prominently in the specific prediction or classification under focus. Another significant local interpretive strategy is Shapley Additive exPlanations (SHAP). It uses a game theory to define a 'shapley value' for a feature of concern that provides a measurement of its influence on the underlying model's prediction. The shapley value is calculated a feature by averaging its marginal contribution to every possible prediction for the instance under consideration. It calculates the marginal contribution for the relevant feature for all possible combinations of inputs in the feature space of the instance and produces the complete distribution of the prediction for the instance. (Leslie, 2019)

Another approach is to use example-based explanations, which use instances from the dataset to explain the behavior of the model and the distribution of the data in a model agnostic way. The counterfactual method indicates the required change in the input side that will have significant changes in the output, like reversing the prediction. They can explain individual predictions. However, a possible problem comes from the so called Rashomon effect, where each counterfactual explanation tells a different story to reach a prediction and there may be multiple true counterfactual explanations and the challenge therefore is to choose the best one. This method does not require access to data or models. It also offers a way to provide affected stakeholders with actionable recourse and practical remedy. It allows the stakeholders to see what input variables of the model can be modified so that the outcome can be altered to their benefit. A downside is that it does not work well for categorical variables with many values. A second example-based explanation is adversarial technique that can flip the decision by using counterfactual examples to fool the model to make a false prediction. These can help the developers to discover hidden vulnerabilities

as well as to improve the model. A third method is influential instances, which are data points from the training set that are influential for prediction and parameter determination of the model. However, while it helps to debug the model and understand its behavior better, determining the right cutoff point to separate influential from non-influential instances is challenging. (Leslie, 2019; Rabiul Islam et al., 2021) Figure 3 demonstrates several methods to improve interpretability and their characteristics.

Method	Approx.	Inherent	Post/Ante	Agnos./Spec.	Global/Local
Linear/Logistic Regression	No	Yes	Ante	Specific	Both
Decision Trees	No	Yes	Ante	Specific	Both
Decision Rules	No	Yes	Ante	Specific	Both
k-Nearest Neighbors	No	Yes	Ante	Specific	Both
Partial Dependence Plot (PDP)	Yes	No	Post	Agnostic	Global
Individual Conditional Expectation (ICE)	Yes	No	Post	Agnostic	Both
Accumulated Local Effects (ALE) Plot	Yes	No	Post	Agnostic	Global
Feature Interaction	No	Yes	Both	Agnostic	Global
Feature Importance	No	Yes	Both	Agnostic	Global
Global Surrogate	Yes	No	Post	Agnostic	Global
Local Surrogate (LIME)	Yes	No	Post	Agnostic	Local
Shapley Values (SHAP)	Yes	No	Post	Agnostic	Local
Break Down	Yes	No	Post	Agnostic	Local
Counterfactual explanations	Yes	No	Post	Agnostic	Local
Adversarial examples	Yes	No	Post	Agnostic	Local
Prototypes	Yes	No	Post	Agnostic	Local
Influential instances	Yes	No	Post	Agnostic	Local

Figure 3. Comparison of different interpretability methods from a set of key perspectives (approximation or actual values; inherent explainability or not; post-hoc of ante-hoc; model-agnostic or model specific; and global or local). Source: Rabiul Islam et al. 2021

The field of interpretable or explainable machine learning (XAI) has been predominantly algorithm-centered and focused on the model instead of the audience (Ehsan et al., 2021). However, explanation from the model and AI system needs to be comprehensible by the user, and there might be some supplementary questions to be answered for a clear explanation. (Rabiul Islam et al., 2021) Explanations are socially situated human to human interactions (Ehsan et al., 2021) and should be delivered in a recipient friendly manner and in plain language (Rabiul Islam et al., 2021; Smith, 2019). Explanations play a central role in sense-making, decision-making, coordination and provide necessary delineations of reasoning and justification of one's thoughts and actions. Technical transparency is not always understandable for the end user. Explainability requires the ability to answer users questions and the ability to be audited. (Ehsan et al., 2021)

Explaining is a process by which an explainee and an explainer achieve common ground and understanding and therefore must be understood by the explainee to be effective. It needs to enable the users to explain to themselves what is happening and why and can be accompanied for instance by instructions, tutorial activities, comparisons, and exploratory interfaces to succeed. Multiple kinds of information are often necessary as they complement each other. Explanations benefit from contrasts, comparisons, and counterfactuals in understanding the boundary conditions of a system. These explanations are especially important when the user is surprised, or their expectations are contradicted. This triggers a need for explanation that may be unnecessary until this time. (Mueller et al., 2021)

While explanation and understanding can sound to be beneficial, explainable machine learning can be misused. Hall et al. (2019) offer four guidelines to avoid unintentional misuse and identifying intentional abuse of explainable machine learning. First guideline is to use explanations to enable understanding. It is important here to differentiate understanding from trust as one can understand a system and not trust it enough to use it. Explanations, however, typically increase trust in models as a side-effect when they are otherwise acceptable to users by various criteria. Hall et al. state that debugging and testing methods should be used to directly promote trust. The second guideline is to learn how explainable machine learning can be used for nefarious purposes. Explaining the model can enable hacking or stealing the model or the data through public endpoints. When explanations are used, the system needs to be tested for vulnerabilities to model stealing, inversion and membership inference attacks. Also providing explanations along with predictions eases attacks that can compromise sensitive training data. This needs to be accounted for. The third guideline is the augment surrogate models with direct explanations and basically combining approaches in a way to enable accuracy of explanation as well as its understandability. The fourth guideline is to use highly interpretable models for life- or mission-critical machine learning solutions. (Hall et al., 2019)

Transparency is a term that is most prevalent in literature, according to Jobin et al. (2019). References to transparency comprise of efforts to increase explainability, interpretability, or other acts of communication and disclosure. It is seen to minimize harm, improve AI solutions, and to foster trust. (Jobin et al., 2019) Barclay et al. state that a system offering good levels of visibility to its internal workings is more likely a system affording transparency and accountability and more likely

to give better assurance on their quality and trustfulness further into the future and hence a provide a better return on investment for the developers and users. (Barclay et al., 2019)

The sources in the research conducted by Jobin et al. vary greatly in what should be communicated. Possibilities include but are not limited to the use of AI, source code, data use, data itself, limitations, laws, responsibility for AI, investments in AI and possible impact. There is also variation for whom should there be transparency and how. (Jobin et al., 2019) There are several possible recipients of disclosure, such as intermediate bodies with oversight role, affected individuals, public in general and so on. (de Laat, 2017) de Laat (2017) further distinguished several stages of decision-making for which there should be transparency. First the data is collected, then processed and used to develop a model, and finally that model is used for decision-making. All in all, this means that transparency has several gradations. (de Laat, 2017) And while it may seem tempting to treat AI solutions as a black box from where a clear understanding of its working is impossible to acquire, this is no longer an option as the decisions and processes that rely on AI increase both in number and importance (Saif & Ammanath, 2020).

Transparency, as a term, means the quality an object has when one can see clearly through it, and the quality of a situation or process that can be clearly justified and explained because it is open to inspection and free from secrets. Transparency as a principle of AI ethics encompasses both meanings, according to Leslie (2019). Transparency means the ability to know how and why a model performed the way it did and the rationale behind its decision or behavior. On the other hand, it involves the justifiability of both the processes that go into its design and implementation and of its outcome and therefore the soundness of the justification of its use. (Leslie, 2019)

Leslie (2019) identifies three critical tasks for designing and implementing transparent AI. First of these is process transparency. This includes creating and maintaining the governance of the solution including relevant team members and their roles, relevant stages of the workflow where intervention is necessary, explicit timeframes for follow-ups and re-assessments and clear and well-defined protocols for logging activity and for instituting mechanisms to assure end-to-end auditability. The company should have a clear documentation on who did what and when and who should do what in the future. This is necessary to demonstrate that trustworthiness is considered in operative end-to-end design and implementation. The second and third task concern outcome

transparency. The second task is that one should be able to show in plain and understandable language also to non-specialists how and why a model performed the way it did in a specific decision-making or behavioral context. The third task is to justify the outcome and demonstrate that a specific decision or behavior is ethically permissible, non-discriminatory, and worth of the public trust. To complete task three, it is necessary to take the explanation created in task one as a starting point and weight that against the justifiability criteria adhered to throughout the design and use of the AI solution. (Leslie, 2019) Barclay et al. (2019) suggests that companies should have a way to qualitatively rank the transparency of their algorithms. This enables them to follow the suitability of the models and to monitor their on-going confidence in the suitability of the model over several years. One such example is in Figure 4, although they state that ideally progress will be made towards determining measurable and objective criteria.(Barclay et al., 2019)

Score	Quantity	Freshness	Accuracy
1	Sparse or insufficient information	Never updated	Demonstrably inaccurate
2	Some information missing	Out-of-date	Believed to be inaccurate
3	Sufficient to gain confidence	Updated when changed	Believed to be accurate
4	Sufficient to validate	Real-time validation	Evidenced and verifiable

Figure 4. Example of a scale to judge documentation on each AI solution. Source: Barclay et al. 2019

A final note on transparency is that there is a debate as to how much transparency is suitable. For instance, full transparency on raw data creates a clear problem from the viewpoint of privacy. If algorithms are shared, it is possible to game the system, which could make the algorithm useless in some use cases. Transparency can also hurt companies' competitive edge. Further, transparency needs can affect model choice and create a tension to the accuracy of the model. It is therefore not advisable to be fully transparent, except to intermediate parties with oversight authority. However, fully opaque to the public is also not advised. (de Laat, 2017) The level of transparency required in law usually involves high-level information about what is happening with the data, what the systems are doing, the risks involved, and what entities are involved rather than the exact specifics of how they function (Singh et al., 2019). The companies utilizing AI should therefore have full, clear, and up-to-date documentation and a clear plan on the levels of transparency to different stakeholder groups.

3.3.3 Diversity, non-discrimination and fairness

AI systems might be a powerful tool to expose bias, unfairness, and other problems (Jansen Ferreira & Monteiro, 2021). They may also reproduce, reinforce, and amplify patterns of marginalization, inequality, and discrimination (Leslie, 2019). Fairness has gained explosive interest in the past decade. It was popularized by an ProPublica report on COMPAS software used in the US courts to predict a defendant's risk of reoffending. In this report, it was detailed how the software overestimated black people's recidivism risk compared to white people. (Euijong Whang et al., 2021; Hildebrandt, 2020) Face recognition systems trained primarily on white faces have been discovered unable to recognize darker skin. Similarly, a voice-to-text system trained on American English may not recognize other accents or forms of English. (Smith, 2019)

Fairness in an elusive concept. It can be defined as the absence of any prejudice or favoritism towards an individual or a group based on their traits in the context of decision-making. (Kumar et al., 2020) The concept of fairness is somewhat amorphous. It may be influenced by cultural, sociological, economic, and legal considerations. Pery et al. (2021) ask in their research, what ought to be fair and who defines this. For instance, unequal distribution of opportunity may require the application of distributive fairness, or affirmative action, that levels the playing field as equality does not necessarily result in the fairness of the outcome. Also, it is necessary to understand that minority groups that are typically the victims of algorithmic bias are rarely given the option to participate in the design, development, and deployment of these systems. (Pery et al., 2021)

The term bias has a historical meaning in machine learning that differs from how the term is used in everyday situations. Bias in machine learning refers to the necessary preference of certain functions over others. Too low inductive bias may lead to the model to overfit. Too high inductive bias leads to the model working badly both in the training data and for new data. (Hellström et al., 2020) However, what we are focusing on here, is unwanted sociological bias that encompasses several forms of discrimination including overt discrimination, disparate treatment, disparate impact (DI), and unintentional discrimination. A model is said to be biased here if group membership, or membership in a subset of a group, is not independent of the likelihood of a favorable outcome. (Hall et al., 2019) Biased machine learning models, or AI systems, may result in making unfair or biased decisions that negatively impact the discriminated individuals. It could also result in finan-

cial and legal issues as well as reputational damage and even push a firm to insolvency. In the jargon of machine learning, the variables that can lead to discrimination are called sensitive attributes. Gender, religion, political affiliation, age, and ethnicity are examples of such sensitive attributes. It is tempting to remove such sensitive attributes, but it is well known that discarding, massaging, or transforming the sensitive attributes does not necessarily remove the unfairness and discrimination due to proxies that carry the discrimination even in the absence of the sensitive attributes. (Hellström et al., 2020; Kumar et al., 2020)

Bias can creep into the machine learning system from various sources. Humans are deeply involved in all parts of the machine learning process. The data can be flawed or inappropriate. The decisions made on the processing of that data can produce bias. All stages of the process can induce a biased system. Human error, prejudice, and misjudgment can enter the lifecycle and create biases at any point in the project. (Leslie, 2019) We will look at some different types of bias next to get a clearer view on possible sources of it.

Specification bias denote bias in the choices and specifications of the designers of what constitutes the input and output in a learning task. Biased choices in these may negatively affect performance and, also, systematically disadvantage protected classes in systems building on these choices. (Hellström et al., 2020) Sampling bias, also called selection bias, population bias or representation bias, occurs when there is an underrepresentation or overrepresentation of a segment of the population. It can result, for instance, in the sampling method reaching only a portion of the population, or it can be self-selection bias as in the case with online surveys about computer use. Sampling bias can result in the model performing badly in general or for a certain demographic group. (Hellström et al., 2020; Pery et al., 2021; Suresh & Guttag, 2020) Aggregation bias is a result of inappropriately combining distinct populations in the data. It is worth considering if a single model can suit all sub-groups of a heterogenous data. Measurement bias occurs when choosing, collecting, or computing features. It can arise in several ways. The measurement process can vary across groups. For instance, more stringent measurement in one group can reveal more errors observed in that group. The quality of data can vary across groups. Also, we use proxies to be able to measure the thing which we are interested in. For instance, the use of pain medication can be a proxy for people in pain. The proxy can be misleading or an oversimplification creating measurement bias. (Suresh & Guttag, 2020) Uncertainty bias refers to classification tasks where the threshold

needs to be set. It is usually set manually and may create a bias against underrepresented demographic groups as less data leads to higher uncertainty. Model bias or algorithmic bias refers to bias as it appears and is analyzed in the final model. If the model is used to predict the biased world as it is, then model bias may not be a problem if it correctly predicts a biased outcome of a biased system. (Hellström et al., 2020) However, if this is not the desired situation, then it may be a case of historical bias, that can occur even in perfectly measured and sampled data if the current state of the world is not the wanted outcome. Historical bias is a misalignment between the current state and the desired state. (Suresh & Guttag, 2020) Inherited bias can occur when the output of one machine learning model is used as an input of another model. If the first model is biased, then the second model inherits this bias. (Hellström et al., 2020) Evaluation bias occurs during model iteration and evaluation when the testing or external benchmark populations do not equally represent the various parts of the use population. It can also occur if the use of the performance metric is not appropriate for the use case of the model. Deployment bias occurs when a system is used or interpreted in inappropriate ways. (Suresh & Guttag, 2020) Decision-automation bias also known as the technological halo effect can occur if the users of automated decision-support systems are hampered in their critical judgement because of their faith in the perceived neutrality and objectivity of the AI system. Automation-Distrust bias can occur at the other extreme where users will disregard evidence-based reasoning due to their distrust or skepticism about AI technologies. (Leslie, 2019) Finally, human cognitive biases are systematic patterns in human judgement that can affect every stage of the design, development, deployment, and use of the AI system. There are more than 190 types of cognitive biases mentioned in the Wikipedia, according to Hellström et al. (2020) suggesting caution when claiming that a machine learning system is unbiased. Figure 5 contains a depiction of a machine learning project process and some possible sources of bias in each stage.

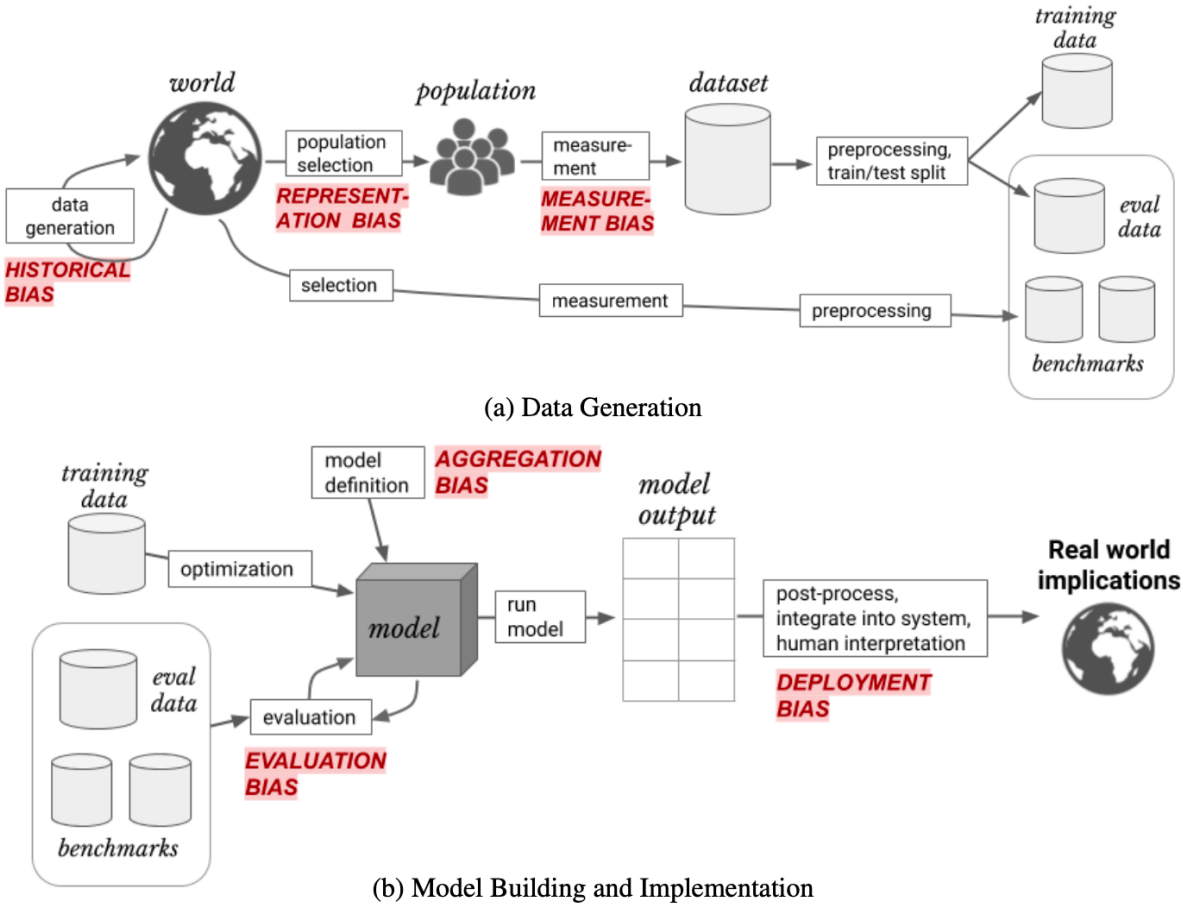


Figure 5. Some sources of bias in the workflow of Machine learning systems. Source: Suresh & Guttag 2020

As fairness is something that requires a definition, the principle of discriminatory non-harm is a minimum requirement of fairness. It means prioritizing the mitigation of bias and the exclusion of discriminatory influences and ensuring that AI systems do not generate discriminatory impacts on affected individuals and communities. This entails that the models are trained and tested on properly representative, relevant, sufficient, timely as well as recent, accurate, and generalizable datasets creating data fairness. A data factsheet can be created and maintained diligently through-out the design and implementation lifecycle to secure data quality, bias-mitigation aware practices and optimal auditability. Secondly, it means that the model architecture does not include target variables, features, and processes which are unreasonable, morally objectionable, or unjustifiable creating design fairness. Choices in the data preprocessing stage as well as feature determination, model building and hyperparameter tuning must be made in a fairness aware manner. Thirdly it requires that the outcomes of the AI system do not have discriminatory or inequitable impacts on

the lives of the users creating outcome fairness. The determination of outcome fairness should depend on the specific use case and the technical feasibility of incorporating the chosen criteria into the construction of the AI system. Finally, it means that the users are sufficiently trained to implement the models responsibly and without bias creating implementation fairness. To accomplish these, it is necessary to identify fairness and bias mitigation dimensions in each stage of the design, development, deployment, and use of the AI system. It requires scrutinizing the potential risks involved and taking action to correct any identified problems. (Leslie, 2019)

There are technical fairness techniques used to diagnose and remediate unwanted social bias in ML models. They can be generally divided into three categories. Fair exploratory data analysis and pre-processing try to transform the data so that the underlying discrimination is removed. Fair in-processing techniques try to modify and change learning algorithms to remove discrimination during the training process. Fair post-processing is performed on a trained model by accessing a hold-out dataset and adjusting for any discrimination discovered. Each method has its own pros and cons. (Kumar et al., 2020) Most of them require significant amounts of effort to deploy. Preprocessing strategies are applicable to any model but require changes in the training data to remove bias. In-processing techniques perform well, but usually propose a new model training algorithm that replaces the existing one. (Euijong Whang et al., 2021)

Other methods to increase fairness is to train the implementors, when they are used, with basic knowledge about machine learning and its limitations. The user-system interfaces should be designed to encourage active user judgement and situational awareness. Fairness policies should be developed, and the AI solutions measured and monitored according to them. (Bigham et al., 2018; Leslie, 2019; Nair et al., 2020) The teams creating AI solutions should be diverse in terms of gender, culture, age, professional background, and skillset. (AI HLEG, 2019; Smith, 2019)

3.3.4 Human agency and oversight

Trustworthy AI starts with the principle of human agency and autonomy (Pery et al., 2021). This brings with it some clear assumptions and design principles. As Smith (2019) elaborates, AI systems must be built in ways that ensure humans are always in ultimate control and responsible for all the system does. This includes the decision, classifications, and predictions made by the system. This is especially important in situations such as judicial, medical, financial, and recruiting decisions

as well as reputational situations. It is very important also regarding government and public sector applications that affect broad populations. The decisions made by the system need to be appealable to a human. There needs to be ways to override the made decisions. The system needs to be monitored and when unexpected results surface, this monitoring should increase. Further, if the system is unknowable and a black box, it should, according to Smith (2019), be turned off. People interacting with the AI system need to be able to easily discern when the AI system is taking action. The system needs to present and explain data sources, their providence, and the training method both in technical and plain language. Further, updates should be scheduled so that people using the system can anticipate if the update will affect their work. (Smith, 2019) Bigham (2018) continues that there needs to be a clear point where the AI solutions hands control over to humans, when the algorithm cannot produce an output within the predefined risk tolerances (Bigham et al., 2018). These guidelines, as well as the others presented in this thesis, need to be applied to the specific use case and their need be in relation to the risk they pose to humans.

There is, based on Jansen Ferreira and Monteiro (2021), a more profound question to be answered by the designers of the AI system and that is of the role the AI system plays in the human-AI interaction and the desired type of this interaction. An AI system can be an assistant, critic, a second opinion, a collaborator, source of information, and an expert. It can empower individuals and help them make better decisions as well as boost their analytic and decision-making abilities. It can provide a different way to assess and classify large amounts of diverse data and show relationships within that data. Best results, based on Jansen Ferreira and Monteiro, are achieved from a partnership between AI and people, enabling new ways for the human brain to think and computers to process data. This relationship between an AI system and a human needs a different approach than a human-to-human collaboration. When humans begin to collaborate, their relationship starts with a mental model based on their shared humanity. Both are aware of the limitations and talents of being human. The collaboration process itself then reveals the particularities of each human. Building an appropriate mental model about the counterpart in collaboration is decisive to build the necessary mental model about that system. The initial human-AI onboarding process can be a way to build an initial impression and the development of appropriate mental models and strategies of use for a human-AI relationship, because people struggle to understand core elements of AI, such as the models and algorithms commonly used in creating predictions and decisions. (Jansen Ferreira & Monteiro, 2021) People are also slow to adopt systems they do not understand and trust (Rosenfeld & Richardson, 2019).

Explanability of the AI system can enable understanding, but different ways of presenting the information and models and structuring the human-algorithm interactions may affect the quality and type of decisions made (Jansen Ferreira & Monteiro, 2021). The type of explainability needed directly depends on, according to Rosenfeld and Richardson (2019), the motivation for the type of human–agent system being implemented and thus directly stems from the first question about the overall reason, or reasons, for why the system must be explainable. For instance, recommendations as well as training and tutoring systems are human-centric and the information provided will need to persuade the person to choose a specific action. If the system is AI-centric, such as knowledge discovery or self-driving cars, the AI system might need to provide information about its decision to help convince the human participant of the correctness of their solution. In both cases, the information provided should build trust to enable the acceptability of the decisions. This, then, creates the need to consider and evaluate how these explanations are generated and presented and if their level of detail matches the system's needs and possible legal considerations. (Rosenfeld & Richardson, 2019) It is also paramount to provide information on the limitations of the system in plain and easily understood language. Usability testing can help determine if the users understand how the AI system works and how to use it responsibly. (Smith, 2019). Smith (2019) introduced a Human-Machine Teaming (HTM) framework for designing ethical AI experiences, that can be found in appendix 2 designed in conjunction with a set of technical ethics to help teams create AI systems that behave as expected, safely, securely, and understandably creating the best possible change for strong human-AI teams.

3.3.5 Accountability

There are two main issues with accountability concerning responsible, or trustworthy, AI projects especially in public sector projects, but also in others. First issue is the accountability gap. The decisions made by the AI system are not self-justifiable as human agents would be. The statistical models and hardware serving them is also not morally responsible as a human would be. This creates an accountability gap. The second issue is related to the complexity of AI production projects. There are many people involved in various parts of the company, or public sector actor, and sometimes also several companies involved. The question then is that who among these parties involved in the production of the system should bear the responsibility if these systems have negative consequences. Should the developers, designers, institutions, or industry be held responsible? Which and by how much? Especially in systems what have been developed by several companies,

but also in others, it is difficult to exercise oversight and to determine where exactly something went wrong, who is responsible, or even to identify the parties involved. This accountability gap can may harm the autonomy and violate the rights of the affected individuals. (Jobin et al., 2019; Leslie, 2019; Singh et al., 2019)

Accountability, as defined by the Committee of experts on internet intermediaries (MSI-NET, 2018), is the principle that the person who is legally responsible of the harm must provide some form of justification or compensation (The Committee of experts on internet intermediaries (MSI-NET), 2018). The entities held accountable are natural and legal persons that is people and organizations, whether for their own actions or actions of people, organizations, or machines under their control (Singh et al., 2019). However, someone can only be accountable if they have a degree of control in causing the harm. This means that they have facilitated or caused the harm or are in position to prevent or mitigate it. Legally speaking accountability manifests itself through liability to provide a remedy. In AI systems, it is not clear who has the necessary degree of control so that liability may be assigned. The developer may not know how the algorithm is used and implemented. The person or team implementing the algorithm may not fully understand what it does. (The Committee of experts on internet intermediaries (MSI-NET), 2018) Also, the context is relevant. In one context an entity may be accountable to end-users. In others, to other system operators, other companies, regulators, courts, or other oversight bodies. The accountability may be a result of a result of statutory obligations, as in data protection, or through contractual relationships. At its core, accountability involves determining liability and, where harm arises, what restitution is owed by who and to whom for that harm. (Singh et al., 2019)

Accountability can be divided into answerability and auditability to help in its implementation. Answerability is the answer to the question who is accountable, and it entails that for each AI system a continuous chain of human responsibility is created throughout the whole AI project lifecycle with no gaps permitted from the first steps of design to use and the outcomes of the system. Answerability also demands explanations and justifications for both the decisions or predictions of the AI system as well as the process behind their production offered by competent human authorities in plain and understandable language. Auditability answers the question how the designers and implementors of the system are to be held accountable. This aspect requires demonstrating the responsibility of the design, use practices, and the justifiability of outcomes. Every step in the

lifecycle of the system needs to be accessible for audit and review, which requires documentation of each step. The deliberate incorporation of both of these elements into the project lifecycle may be called Accountability-by-design. (Leslie, 2019)

Accountability measures can be divided into anticipatory accountability covering accountability during the design and development stages. This is also called ex-ante accountability. The other type is ex-post or remedial accountability, which focuses on remedying any possible harm the system has caused. Of these two, anticipatory accountability should be prioritized as implementing accountability in the design and implementation effectively pre-empts possible harms. However, remedial accountability is no less important for providing necessary justifications for the bearings these systems have on the lives of the stakeholders. Putting in place a comprehensive auditability regimes as a part of accountability measures, and establishing transparent design and use practices, and providing understandable explanations to affected stakeholders, are essential components for remedial accountability. (Leslie, 2019)

At a technical level, accountability is grounded in transparency and control (Singh et al., 2019) throughout every step of the lifecycle of the system (de Laat, 2017). Transparency is often a regulatory requirement for identifying responsibility or liability (Singh et al., 2019). Control of an AI system includes monitoring it for usage, outcomes, accuracy, confidence, and overall analysis (Smith, 2019). Also, effective redress mechanisms for individuals whose rights are infringed are also essential (The Committee of experts on internet intermediaries (MSI-NET), 2018). In all cases, users of the system should be able to do some research themselves on the functioning of the system, should they suspect that something is wrong, and have a way of reporting problems (Smith, 2019). Another possible way to approach accountability is to appoint a person, or a team, in charge of ethics issues relating to the AI systems, who provides oversight and advice (AI HLEG, 2019). Finally, Singh et al. (2019) propose decision provenance as a way to assist accountability considerations in algorithmic systems. Decision providence involves providing information on the nature and contexts of the data flows and interconnections leading to a decision or action, the flow-on effects, and how this information can be used to improve the system and inspect it. It helps expose the decision pipelines by making the inputs and outputs of each stage visible by showing from whom the data comes from or goes to, how the data is processed, and used, as well as any data protection aspects, system configurations, actions of individuals and so on. Decision providence is the history

of the data and the broader view of system behavior and interactions in the system. (Singh et al., 2019)

3.4 Robust AI

3.4.1 Technical robustness of AI solutions

Robustness of a system refers to the ability to produce consistent and reliable output, be able to act if inconsistencies are discovered, and must be available when it is supposed to be available. It needs to produce consistent and reliable outputs also in less ideal conditions for instance when encountering unexpected data. Further, it needs to scale well while remaining robust and reliable, and if it fails, fail in a predictable manner. For this to be possible, processes for handling issues and inconsistencies need to be established. (Leslie, 2019; Saif & Ammanath, 2020) It needs to be built in a professionally acceptable way and to current standards (Smith, 2019). Reliability is also an aspect of robustness and that means that the AI system behaves exactly as its designers intended it to behave. A measure of the robustness is the strength of a system's integrity and soundness of its operation in response to difficult conditions, adversarial attacks, perturbations, and data poisoning (Leslie, 2019). We will look at adversarial attacks and data poisoning in the next subchapter. Now, we focus on issues concerning robustness.

Most risks for the safety and robustness of AI solutions are the same as in more traditional IT systems. However, what sets these solutions apart, is the interpretation of training data, the knowledge contained in the machine learning model, use of transfer-learning and the process of fitting the model. (Vähä-Sipilä et al., 2021) The risks one faces with AI solutions will depend on for instance the sort of algorithms and machine learning techniques used, the type of application the model is going to be deployed in, the provenance of the data, the way the training objective is specified, and the problem domain in which the AI solution is deployed into. Unreliable, unsafe, or poor-quality outcomes are due to, among others, irresponsible data management, negligent design and production processes, and questionable deployment practices. (Amodei, 2016; Leslie, 2019) Lack of transparency around algorithm design, incorrect use of algorithms, and weak governance are also reasons why AI systems are subject to risks due to errors, biases, and malicious

acts (Krishna et al., 2017). Further, the hardware that is used can also cause performance issues (Banerjee & Chanda, 2020). As can be seen, there are many different routes for issues to arise for robustness. Next, we will look at three of them more carefully.

One big area of possible robustness issues is the choice of the performance metric, or objective function. If chosen incorrectly, this can lead to harmful results even with perfect learning and infinite data. (Amodei et al., 2016; Leslie, 2019.) In machine learning the performance metric is what is optimized in the algorithm and should be chosen carefully. Choices include but are not limited to accuracy, precision, and specificity. Accuracy, for instance, is the proportion of the examples for which a correct output is produced. (Leslie, 2019) For instance, in a system predicting fraud attempts, testing for precision which is calculated by dividing the amount of correctly identified positive cases by all predicted positive cases is a better metric than accuracy, where the system can achieve a 99% accuracy in a system where one percent of the data is fraudulent by classifying all as negative. More information on performance metrics can be found in any machine learning textbook. A hundred percent infallibility is not realistic (Vähä-Sipilä et al., 2021) and an acceptable level of the performance metric for production should be decided before development to avoid the slippery slope of almost good enough. The specific use case defines the most suitable performance metric and the suitable level of performance. For instance, domain established benchmarks can help in setting the level of the performance metric. (Leslie, 2019)

Another issue in machine learning projects is concept drift. All machine learning models are built with historical data what have become fixed in the systems' parameters. When this crystallized historical data, known as training data, ceases to reflect the population concerned, the model's mapping function will not be able to transform inputs accurately and reliably into the target output values in an accurate way. These systems become prone to errors. A feedback-loop where new data is used to refresh training data with model retraining can help avoid concept drift. Also monitoring new data can help identify concept drift, or data drift as it is also called. (Leslie, 2019)

A third issue is that many high-performing machine learning models, such as deep neural nets, can be brittle. This means that as they are running in an unpredictable environment, these systems may have difficulty in processing unfamiliar events and scenarios. This can lead them to make unexpected and serious mistakes that may remain unexplainable given the high-dimensionality and

computational complexity of their mathematical structures. In safety-critical applications, such as in automated transportation and medical decisions, these undetectable changes in inputs may lead to significant failures. (Leslie, 2019)

One best practice in the field is to train and test the model on different data set to ensure the validity of the outputs when the model is dealing with new and unseen data. Frequent or ongoing testing as well as statistical analysis of the algorithm should also be conducted to check if the AI solution is performing in line with expectations. Also, it is advised to check the performance against the results of a non-AI system to check for accuracy. Finally, developers need to build a way to stop an algorithm as soon as an error or abnormal behavior is discovered, which needs to also include procedures for business continuity and remediation. (Bigham et al., 2018)

3.4.2 The security of AI solutions

AI systems are currently all too often designed with no consideration for security, according to Oseni et al. (2021), making them very vulnerable to adversarial attacks. The goal of AI security encompasses the protection of the AI system from possible adversarial attacks while maintaining the integrity of the information that constitutes it and having the system continuously functional and accessible to authorized users. The architecture and all the individual parts of the system needs to be protected from unauthorized modification or damage. The data used needs to be kept confidential even under hostile or adversarial conditions. (Leslie, 2019)

The attack surface of an AI system considered to be by Oseni et al (2021) the total sum of vulnerabilities the AI model is exposed to. An AI system deployed in software, however, also faces additional attack vectors due to security risks in the software (Vähä-Sipilä et al., 2021). The attack surface can be described as a list of inputs that an adversary can use to attempt an attack on the system. Adversarial goals can be broadly defined as attacks on the confidentiality, integrity, availability, or privacy of the system. If the attack is focused on confidentiality, the goal is to gather insights about the internals of the model or dataset and to use this information to carry out more advanced attacks. If the attack is focused on integrity, then the goal is to modify the AI logic. If the attack is focused on availability, then the goal is to disable the system's functionality by for instance flooding it to prevent authorized users from accessing it or lead it to making errors. Privacy attacks focus on gaining insight on the data or model. (Oseni et al., 2021)

The attack mainly happens during the training phase or during the testing phase. Attacks during the training phase seek to learn, influence, or alter the performance of the model. The most direct attack is an attempt to read or access the training data. The adversary, providing they have information on the dataset, can inject malicious data to the training dataset to poison the data, manipulate input features or data labels. If the attacker has knowledge of the data and the model, they can carry out a logic corruption attack by altering the learning logic. This last group of attacks is the most advanced and the most difficult to guard against. Attacks during the testing phase are exploratory attacks that do not alter the training process nor influence the learning. Rather, their goals are to discover information about the state of the AI model. These inference attacks rely on information about the model and its use in the target environment. In a white box setting the attacker has knowledge of everything about the model and data. In a black box setting the adversary has no information, but is able to use input to output pairs to infer vulnerabilities in the model. (Oseni et al., 2021, also Vähä-Sipilä et al. 2021) After the model is in use, its decisions, predictions, and classifications can be used to infer assumptions and details about the training data breaching confidentiality (Vähä-Sipilä et al., 2021). Figure 6 contains a framework for analysis of adversarial attacks.

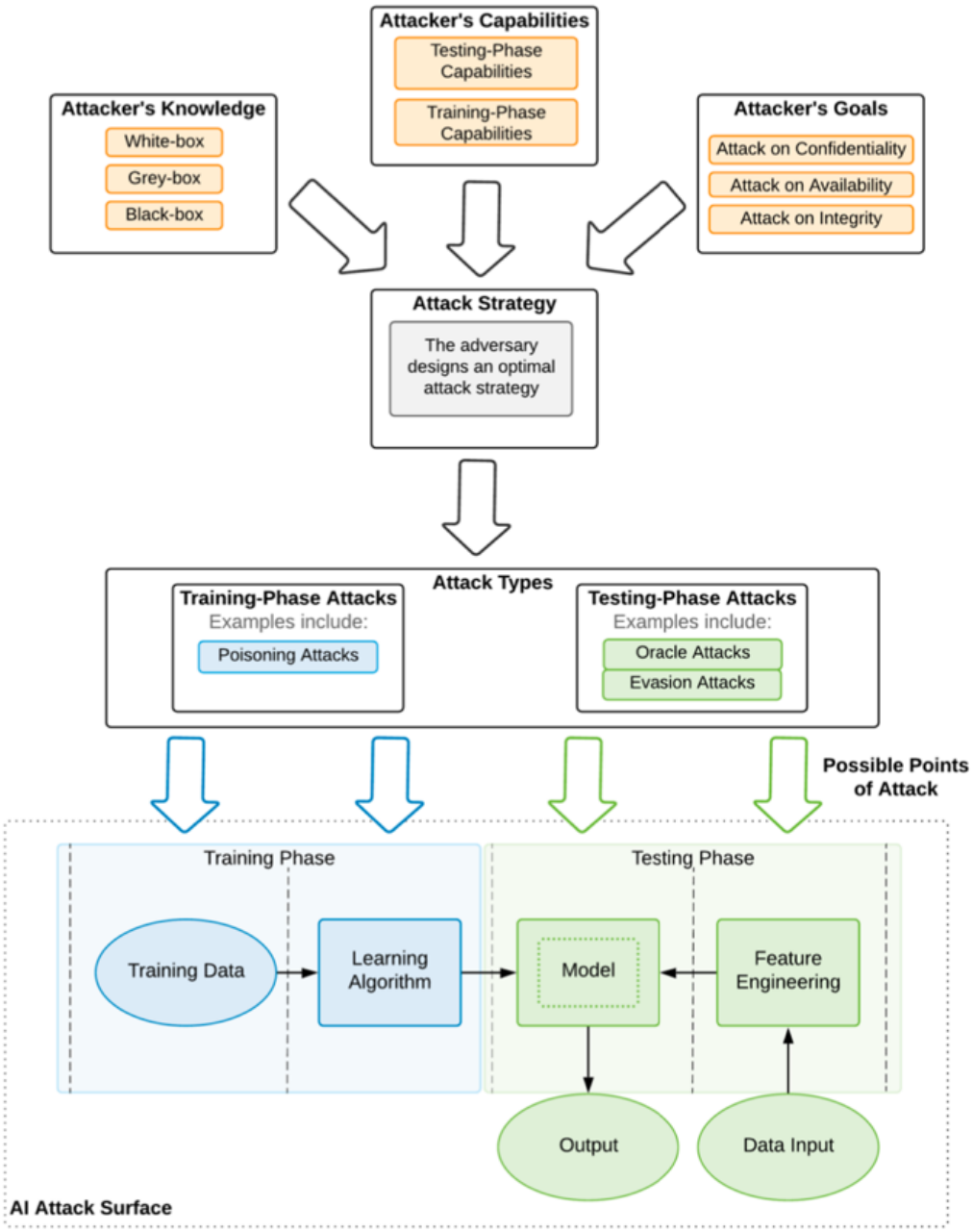


Figure 6. A framework for analysis of adversarial attacks against AI models. Source: Oseni et al. 2021

There are many different types of adversarial attacks. Poisoning attacks are staged at the training stage by injecting or removing samples in the training dataset with the aim of changing the decision boundary of the target model. As this manipulated dataset is used to train, validate, and test a model, it is more prone to misclassifications, systemic malfunction, and poor performance. Further, an adversary can introduce a backdoor into the model that causes it to trigger an error or failure when the maliciously selected inputs are processed. The most common type of poisoning

attack is an error-generic poisoning attack, where the aim is to cause a distributed denial of service attack (DDoS) by producing as many misclassifications as possible. Oracle attacks are exploratory attacks where an adversary uses samples to collect and infer information about the model and the training data. This is easy when the model is served via an API endpoint and the attacker can send an input datapoint into the endpoint and receive the reply. By connecting these input and output pairs it is possible to train a surrogate model that operates much like the original model. Membership inference attacks aim to determine if a given datapoint belongs to the training set to learn the model's parameters. For instance, using this attack form against a model predicting the presence of a disease it is possible to infer if a specific patient, whose data is used in the training, has that disease. Inversion attacks aim at reconstructing training inputs from a model's predictions. Individual datapoints are not revealed, but average representations of each class are illustrating privacy issues especially when models are served via API's. Evasion attacks aim to manipulate input samples at test time to avoid detection and to cause the desired error in use. Deep neural networks have been found to be highly vulnerable to this type of threat. (Oseni et al., 2021; Vähä-Sipilä et al., 2021; also Leslie, 2019) Deep neural networks, luckily, can also represent functions that can resist adversarial perturbations (Oseni et al., 2021; Zhang et al., 2019). The evasion attack can be error-generic when the goal is to mislead classification irrespective of the output class. It can also be error-specific when the aim is to produce a specific type of error. (Oseni et al., 2021)

Oseni et al. (2021) express concern that most of AI technologies are so vulnerable to adversarial attacks. There are many methods to fight back, however. Smith (2019) proposes holding workshops with the, hopefully diverse, development team and stakeholders to identify the full range of harmful and malicious use of the AI system being designed. After they have been identified, then a plan to evaluate and mitigate needs to be created. Also, blind spots in the data need to be discovered. Finally, a plan should be created for mitigation should an attack occur. Having answers, plans, and responsibilities of each actor ready for instances where an attack occurs enables the team responsible for the AI solution to use rationale and clear thinking in response. (Smith, 2019)

There are also technical methods to help prevent and fight adversarial attacks. Their goals are to narrow the attack surface and make attacks through it more difficult. Limiting the number of queries against an API endpoint can prevent stealing the model or inferring information about the

training dataset. The queries can be limited within a timeframe or from a specific IP address. The input feeds coming from an API endpoint can also be statistically monitored for abnormalities that could indicate an attack. The problem with this approach is that sometimes an abnormality is real. The system can be designed to monitor abnormal predictions and make sanity checks to them. One method to do this is to use ensemble models where each individual model has only a small input in the result as the prediction is winning class among them all. The AI system is also possible to design in a way that it has two modes of operation where one is the normal AI system, and the other is a more failsafe the system changes into when it discovers an abnormality. The model can also be regularized during the training stage. This means that the model is rewarded for accuracy and punished for too much complexity. Regularization is often used to avoid overfitting. This usually entails leaving some data out of the process and with luck, any poisoned data is in that part that is left out. Further, it is possible to utilize differential privacy methods. These methods aim to prevent revealing the specifics of any individual data point while maintaining the statistical property of the data. Both regularization and differential privacy can worsen the predictive capabilities of the model, so a balance must be reached between the safety and usability of the model. Models can also be trained with adversarial data, which makes it more secure, but, again, worsens its performance. (Vähä-Sipilä et al., 2021; also Oseni et al. 2021) Defensive distillation is a technique designed for using an ensemble of models or large highly regularized models and transferring their knowledge to smaller distilled models while preserving the prediction accuracy. Gradient masking is a technique for deep neural networks that seeks to reduce the sensitivity of a model to small input perturbations. Finally, homomorphic encryption allows certain mathematical operations to be carried out on encrypted data without the need to decrypt the data. This can help if the data is sensitive. The problem with these methods is that many of them have already been broken or bypassed raising concerns about the robustness of the existing defense methods. (Oseni et al., 2021) Figure 7 has a taxonomy of defenses against adversarial attacks on AI systems.

Figure 7. Taxonomy of defences against AI system attacks. Source: Oseni et al. 2021

Adversarial attacks can, luckily, also be combated also with architectural choices on central training, edge training and training on users' device. If the training is done centrally, then the process can be controlled, but all the data is at the disposal and perusal of the developers creating possible data protection issues. Distributed training, where the data sources themselves handle part of the training, does not require that all data is given to the designers. But at the same time, it widens the window for poisoning attacks through data and through model updates. If predictions are done centrally, it is easiest to protect the system, but necessitates moving possibly sensitive data around and causes the most delays in the prediction. In the edge, the predictions lessen the time delay for prediction and does not necessitate gathering sensitive data in one place, but edge systems are physically more vulnerable to attacks than a central one. If predictions are done in a clients' personal device, then personal data is easier to keep safe. (Vähä-Sipilä et al., 2021) Finally, security can be improved by using only trustworthy data from trusted sources with clear provenance and governance measures as open data sources have been known to sometimes be poisoned (Oseni et al., 2021; Vähä-Sipilä et al., 2021).

3.4.3 All things data

Data has featured in previous chapters as it is one of the most fundamental aspects of AI systems. However, it is necessary to look at a few more aspects of data. The ability to process and store huge amounts of data has been, according to Stoica et al. (2017), one of the key enablers of AI's recent success. It has allowed developing personalized systems and services and significant economic benefits. However, they also require vast amounts of sensitive data, and its misuse could affect users' economic and psychological wellbeing. Also keeping up with the data generated is becoming increasingly difficult due to the amount data growing exponentially as the rapid development of hardware technology slowing down. The challenges, therefore, is to design AI systems that enable personalization while not compromising user's privacy and security as well as address the performance needs of future AI applications with custom chips for AI workloads and edge-cloud systems, and techniques for abstracting and sampling data. (Stoica et al., 2017)

Building AI systems that do not compromise user's privacy require that their information is safe and not mandating that they provide more information than necessary. The GDPR mandates that the bare minimum information is gathered to do what is required and it is stored for the shortest amount possible. For longer use, anonymization is necessary. (Smith, 2019) A thing to consider is that from the perspective of GDPR, the trained model can be personal data and should be considered as at least pseudonymized data, while some language models can retain personal data in its original form. (Vähä-Sipilä et al., 2021) Ethical AI sees privacy as a value to uphold and as a right to be protected. Privacy as a concept is often linked to protection and security, but also to freedom and trust. Solutions to achieve this are for instance differential privacy, privacy by design, data minimization and access controls, and regulatory approaches. (Jobin et al., 2019) Also federated learning, a concept originally introduced by Google, can enable training models on data from users' mobile devices without gathering the data centrally due to privacy concerns. Federated learning provides privacy advantages since only minimal updates necessary to improve a particular model is transmitted. (Oseni et al., 2021)

Apart from data privacy, also data quality is one of the most important problems in building AI solutions. Real life data is often dirty. This means that it contains inconsistent, duplicated, inaccurate, incomplete, or stale data. It continually generates misleading or biased analytical results and decisions and requires data quality management. Data quality management enables the detection

and correction of errors in the data to improve the quality and add value to business processes. Aspects to focus on are data deduplication, data accuracy, data currency, and information completeness. (Fan & Geerts, 2012) Assessing data quality is an ongoing effort that requires awareness of the fundamental principles underlying the development of subjective and objective data quality metrics that can be seen in Figure 8 (Pipino et al., 2002). Logrén (2020) identifies five meaningful practices for quality assurance, which are data and domain understanding, design, verification and validation, documentation, and engineering practices. In particular, the engineering practices appear to have a significant impact on the quality of the machine learning development work in general. (Logrén, 2020) Data provenance can aid in maintaining quality as it captures information about the data. It records data lineage and pipelines, the associated dependencies, contexts, and processing steps (Singh et al., 2019). The entire chain of custody of your data should be documented from the contents of the dataset to the way it was collected to the transformations that have been applied at every step along the way. Without this documentation, models that might be built relying on this data are unaware of the ways the data has been manipulated. Managing and documenting the chain of custody is also vital for security. It should be very clear at every stage who is able to access the data and who is able to modify it. Besides preventing leaks, it can be vital for policy compliance. (Simpson Rochwerger & Pang, 2021)

Dimensions	Definitions
Accessibility	the extent to which data is available, or easily and quickly retrievable
Appropriate Amount of Data	the extent to which the volume of data is appropriate for the task at hand
Believability	the extent to which data is regarded as true and credible
Completeness	the extent to which data is not missing and is of sufficient breadth and depth for the task at hand
Concise Representation	the extent to which data is compactly represented
Consistent Representation	the extent to which data is presented in the same format
Ease of Manipulation	the extent to which data is easy to manipulate and apply to different tasks
Free-of-Error	the extent to which data is correct and reliable
Interpretability	the extent to which data is in appropriate languages, symbols, and units, and the definitions are clear
Objectivity	the extent to which data is unbiased, unprejudiced, and impartial
Relevancy	the extent to which data is applicable and helpful for the task at hand
Reputation	the extent to which data is highly regarded in terms of its source or content
Security	the extent to which access to data is restricted appropriately to maintain its security
Timeliness	the extent to which the data is sufficiently up-to-date for the task at hand
Understandability	the extent to which data is easily comprehended
Value-Added	the extent to which data is beneficial and provides advantages from its use

Figure 8. Data quality dimensions. Source: Pipino et al. 2002

3.4.4 Societal and environmental wellbeing

As has been established, AI system come with inherent risks and potential benefits. It may disrupt established norms and methods of work and societies. Snyder Caron and Gupta (2020) present the adoption of technology as a form of social contract which evolves and fluctuates in time, scale, and impact. This social contract arises when there is sufficient consensus within society to adopt and implement the technology. If the benefits of the technology are hard to identify, technology is difficult to control and provides unforeseeable risk and unprecedented scenarios, or there is risk of

harm, the adoption and implementation of the technology may inevitably lag. (Snyder Caron & Gupta, 2020) Hence, the technology needs to be beneficent and not maleficent. In research, beneficence is often mentioned but rarely defined. However, augmentation of human senses, promotion of human well-being and flourishing, peace and happiness, creation of socio-economic opportunities, and economic prosperity are often mentioned according to Jobin et al. (2019). These benefits should be shared, but to whom is also often left undefined. In the research of Jobin et al. the beneficiaries are mentioned as the society, all humanity, everyone, as many people as possible, all sentient creatures, environment, and the planet. Maleficence is more mentioned than beneficence by a factor of 1.5 and it encompasses calls for safety, security, or mention that AI should not cause foreseeable, or unintentional, or intentional harm. (Jobin et al., 2019)

One area that deserves a special mention here is social bubbles created by filtering, ranking, and recommendation algorithms in search engines and social media. They may unintentionally or intentionally introduce bias as they attempt to deliver relevant and engaging content. It has been suggested that this limits our exposure to diverse points of view and makes us vulnerable to manipulation and dishonesty. Nikolov et al. (2018) discovered in their research that search engines expose us to diverse set of resources with varying levels of bias, while social media traffic exhibits high popularity and homogeneity bias. They state that while the capacity of AI systems to curate individual experiences and to personalize digital services holds promise to improve consumer life and service delivery, these are clear risks that should be taken into consideration. Automatically enabled hyper-personalization limits our exposure to worldviews and, through that, polarize social relationships. (Nikolov et al., 2018) China has, as previously stated, has already started to regularize recommendation algorithms.

The social contract, mentioned above, needs a socially accepted purpose, a safe and responsible method, socially aware level of risk involved, and a socially beneficial outcome in AI systems. In AI systems, the clear identification of purpose ought to happen at the time of design and it should be done both in technical language and through unambiguous and clear language. At a minimum, it should meet existing human rights and constitutional, fundamental, and ethical values of a society. (Snyder Caron & Gupta, 2020) At a minimum, it should not further hinder the life of the already disadvantaged individuals. As a fast gut check Eubanks (2018) recommends answering questions of is it targeted at poor people, and if so, does it increase their self-determination and

agency and would it be tolerated if it was targeted at the non-poor (Eubanks, 2018). We have already covered the second aspect for the acceptance of AI system, that of a safe and robust AI method. The third aspect is a socially aware level of risk involved. Now, any technology has an inherent level of risk involved. It is almost impossible to implement a zero-risk policy and that is not required. It is more a question of a socially accepted level of risk regarding the potential benefit. So safe AI does not have to guarantee a continuous zero-risk AI. If there are identified risks, an adequate warning and disclosure of such risks, as well as instructions for use and for non-use scenarios is sufficient. Also, the context appropriate for the system and the intended purpose of the system should be clearly indicated and brought to the knowledge of the user. Information about the conditions of testing and information on the accountability of the system is also needed. This means the partial accountability of the designers, programmers, and product manufacturers to develop safe AI systems, the supervision, monitoring and enforcement to designated regulators and any punitive and compensatory legal mechanisms available for damages or other remedial and punitive measures through the public justice system. The final aspect is a socially beneficial outcome. Basically, putting together a socially accepted purpose through a safe and responsible method while being socially aware of the risk involved still requires that the AI-enabled system needs to produce a socially beneficial outcome that is in line with the context and culture of the target audience. This is simpler if the target audience is a small niche and more difficult when it is a large heterogenous groups. In larger heterogenous groups decision are made on how the target audience and the social benefit is defined. User research and feedback at prototype phase is one way to aid in the definition of these to increase diversity of world views. (Snyder Caron & Gupta, 2020)

Designers and users of AI systems should remain aware that these technologies may have transformative and long-term effects on individuals and society and developers should proceed with a continuous sensitivity to the real-world impacts that deployed systems have. It is recommended that the social impact and sustainability of the AI project is evaluated. (Leslie, 2019.) One aspect that is important to focus on is environmental sustainability. It calls for development and deployment of AI to protecting the environment, improving ecosystems and biodiversity, contributing to more equal societies, and promoting peace. Further thought should be focused on energy efficiency and minimizing the ecological footprint of training algorithms. (Jobin et al., 2019) Desislavov et al. (2021) noticed in their research that if the increase of AI solutions is kept at a constant multiplicative factor, algorithmic improvements, hardware specialization and hardware consumption efficiency compensate for the growth. However, as more and more devices utilize AI, the energy

consumption can escalate (Desislavov et al., 2021). Ojika et al. (2021) mention that by 2040 it is estimated that 14% of the world's carbon emissions come from data centers and that developing nations are particularly susceptible to the adverse effects that follow. It is possible to utilize AI to help in this by modelling and simulating energy usage and predicting equipment failure. (Ojika et al., 2021)

3.5 Trustworthiness in different project stages

The machine learning pipeline involves a series of choices and practices from evaluation methodology to model definition. All of them can lead to unwanted effects. It is not straightforward to identify, what problems might be present, or once identified, how they should be solved and how these solutions might generalize over factors such as time and geography. (Suresh & Guttag, 2020) The requirements trustworthy AI need to be taken into consideration in all stages of the AI systems' lifecycle. This is a conceptually challenging task. (Euijong Whang et al., 2021.)

Let's start close to the beginning, from data collection, although most machine learning practitioners use existing datasets rather than collecting new ones (Suresh & Guttag, 2020). The quality of the data is essential. The data needs to answer the questions being asked. This is vital and causes a precarious task when data is imported from one context to another, a process that is increasingly easy and often used. If the data is exported from another context, then a challenge is to distinguish any poisoned data from the rest and check for fairness and robustness in data labeling. (de Laat, 2017; Euijong Whang et al., 2021.) In all cases, it is important to check that the data is free from bias, that the labeling is equally correct regarding all classifications, does not contain more or less errors or detail in any group or underrepresented groups. Finally, it is required to check for proxies that contain discriminatory information that affects the model. (de Laat, 2017) Once these have been checked for, the data needs to be cleaned of any discovered problems. One framework that can help with this is MLClean, that performs data cleaning, data sanitization, and unfairness mitigation together, as not much is known how different techniques from different frameworks work together especially when the data is dirty, biased, and poisoned. Data sanitization protects the data against adversarial poisoning instead of just adding noise. (Euijong Whang et al., 2021)

After the data has been acquired and cleaned, it is time for model construction. Concerns in this stage are for instance overfitting, where the data fits the training data too well but does not generalize to unseen data. This can be combatted with regularization, or for instance early stopping when error rate between the training data and test data start moving in different directions. Also using several models can help. Another problem to be faced in model cleaning and model construction is class imbalance. This means that the target variable is unevenly represented in the population. For instance, in tax evasion or monetary fraud, the problematic cases make up only a tiny fraction of all transactions and a model that fits to the majority and creates excellent accuracy scores can perform abysmally in finding the problematic cases that are searched for. Choosing the most appropriate performance metric can focus on identifying the correct cases. Another approach is undersampling, where one deletes data points from the overrepresented class, or oversample, which means adding data points from the underrepresented class. Oversampling can be accomplished by artificially creating new datapoints that are located nearby the available minority points. Also, preferential sampling deletes and or duplicates training instances. Finally, one can subtly alter the classification rules in the post-processing stage if necessary. (de Laat, 2017)

Upon completion, the model is ready to be used for making decisions. At this point there are many possibilities in deciding the degree of automation from mainly human to fully automated decision-making. The choice of the degree of automation should be documented and justified. (de Laat, 2017) After deployment, the fairness and robustness of the solution needs to be monitored (AI HLEG, 2019; Suresh & Guttag, 2020). There may also be a need to integrate real-time feedback into the system. (Euijong Whang et al., 2021; Suresh & Guttag, 2020.) Model can be debugged by model assertions, security audits, sensitivity analysis, and variants of residual analysis (Hall et al., 2019). Finally, resources should be allocated to ongoing training of the model to avoid model or concept drift. It is a good idea to refresh models at least monthly. This all seems like a lot, but even something as small as writing down the data used puts the company in a better place. (Simpson Rochwerger & Pang, 2021.) Hall et al. (2019) have drawn a diagram, in Figure 9, of a proposed holistic ML workflow in which explanations highlighted in red are used along with interpretable models, disparate impact (DI) analysis and remediation techniques, and other review and appeal mechanisms to create an understandable and trustworthy AI system.

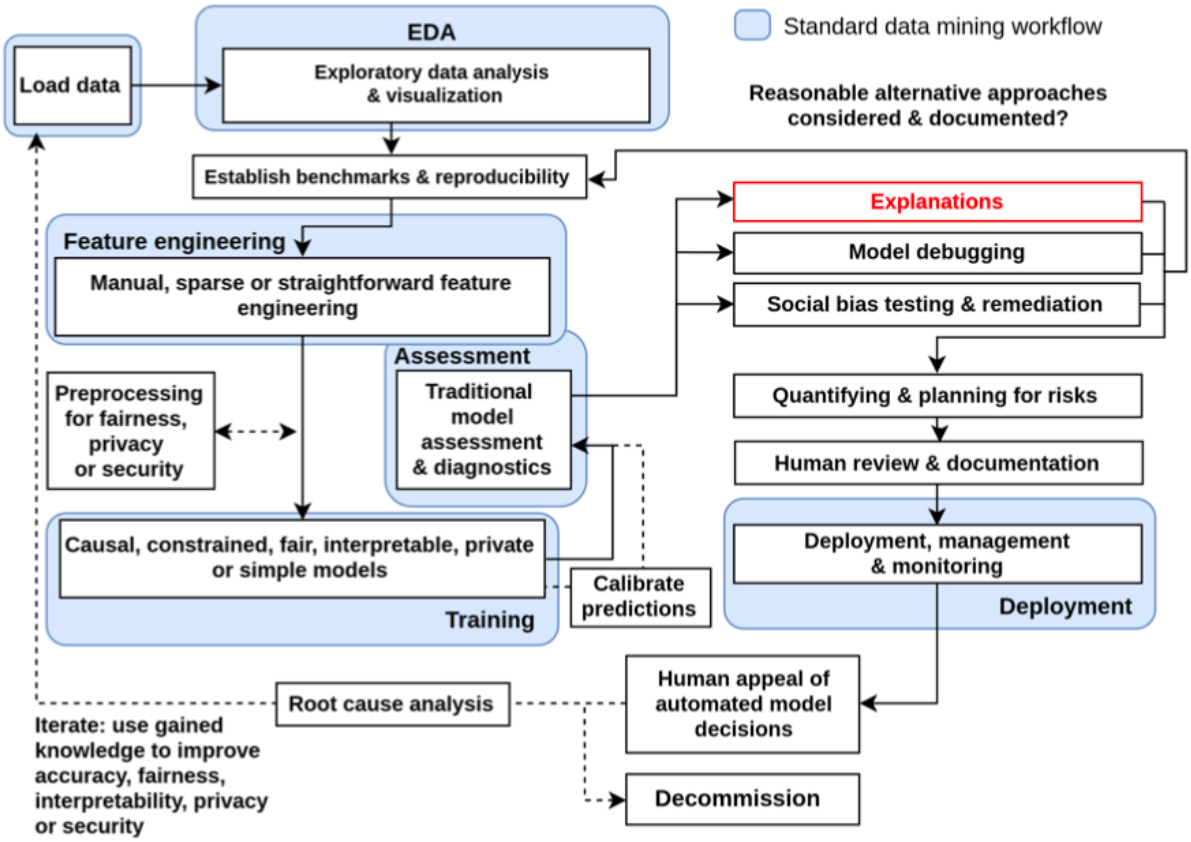


Figure 9. Proposed holistic ML workflow. Source: Hall et al. 2019

3.6 The Big Questions of Trustworthy AI for AI projects – in summary

To summarize the chapter on literature a list of questions has been created and collected to aid the development of trustworthy AI by offering a tool for the necessary discussions around AI solutions. An attempt was made for brevity and conciseness and, also, easy adaptability to different contexts. The importance of each question varies based on the context. For instance, an electric company utilizing AI to predict a suitable price of electricity has a different importance of the question than does a recommendation agent designed to recommend suitable medical treatments to individuals.

1. Does it comply with all relevant laws and regulations and human rights? Basically, is it legal?
2. Is it ethical?
    a. Interpretability, explainability, and transparency

        i. What interpretability methods are used to understand the working of the model in the usual cases as well as in anomalies?
        ii. What methods and delivery channels are used to deliver timely and understandable explanations to stakeholders and what is the content for each stakeholder group?
        iii. How transparent should we be about each aspect of the project, why and to whom?
        iv. What needs to be done to enable these requirements and how, as well as, by whom is the documentation etc. kept up to date?
    b. Diversity, non-discrimination, and fairness
        i. How was fairness defined and by whom?
        ii. What sources of bias were considered and mitigated during the design, development, and deployment of the AI system and how?
        iii. What other means to assure a fair system were used and how?
        iv. How diverse is the team responsible for the life-cycle of the AI solution?
    c. Human agency
        i. What is the role of the AI system in the human-AI interaction, what does this role need to succeed and who is responsible for enabling these requirements?
        ii. Are humans safe and in control and is the system built according to the best practices in human agency?
    d. Accountability
        i. Who is responsible and of which aspect?
        ii. To what degree is the system auditable and documented?
        iii. How can users report problems or seek redress?
3. Is it robust?
    a. Robustness:
        i. What is the best suited performance metric for the use case and what is a suitable performance level?
        ii. Are modeling and IT best practices used in the development of the system?
        iii. Is testing thorough and documented covering also unlikely, but possible, scenarios?
        iv. Are failure and recovery methods planned and implemented?
    b. Safety:
        i. Has the attack surface of the use case been evaluated?
        ii. What security measures have been implemented and how?
        iii. What is the plan for mitigation in case of an attack?
    c. Data:
        i. Are data best practices utilized and data minimized?
        ii. Is the governance of the data documented and continually maintained?
        iii. How is the quality of the data assured?
    d. Societal and environmental wellbeing:
        i. Has it beneficence and maleficence to humans, society, all living beings, and to the environment been considered and evaluated, how, by whom and with what results?
        ii. How well does it meet the requirements of socially accepted risk, safe and robust model, socially aware level of risk involved and socially beneficial outcome?
4. How have these requirements been taken into consideration in all stages of the AI system lifecycle?

4 Research Design

4.1 Case study as methodology

This chapter aims at describing the choices made in the research design, the organization in question, and the data used. The research was conducted as a case study. Tang (2021) describes a case study as a practical and result-driven way to investigate a specific phenomenon or study a problem in-depth creating often descriptive and explanatory results which may not be generalizable or applicable to other solutions (Tang, 2021). Järvinen (2012) describes a case study as an intensive analysis of an individual unit stressing developmental factors in relation to the environment (Järvinen, 2012). Khairul baharein (2008), on the other hand, describes case study as being concerned with how and why things happen, allowing the investigation of contextual realities and the difference between what is planned and what actually occurred (Khairul baharein, 2008). Simons (2009) quotes Merriam in saying that a case study can be defined as an intensive, holistic description and analysis of a single entity, phenomenon, or social unit. Case studies are particularistic, descriptive, and heuristic, and rely heavily on inductive reasoning. They can be intrinsic, where the case is studied for the intrinsic interest in the case itself, or they can be instrumental, where a case is chosen to explore an issue determined on some other ground and the case is chosen to gain insight or understanding into something else. They can be theory-led where the case is explored through a particular theoretical perspective, or they can be theory-generated, where the theory is generated from the observations. (Simons, 2009)

The purpose of a case study is to study a specific case instead of sampling from a specified population. The purpose, in software engineering, is also to not only increase knowledge, but also to bring about some change in the phenomenon being studied, to improve the software engineering process and results in some way. Research methodology close to case study is action research, where the purpose is to influence or change some aspect of whatever is the focus of the research. In a case study, however, the goal is to purely observe. (Host et al., 2012) The case study is intended to focus on a particular issue, feature, or unit of analysis (Khairul baharein, 2008).

Study objects in software engineering case study are usually private companies or units of public agencies developing software rather than entities using software and they aim to improve engineering practices. In a typical situation actors apply technologies in the performance of activities

on an existing or planned software related product or interim product. As the research is carried out in real-world setting, a researcher needs to consider not only the practical requirements and constraints from the researcher's perspective, but also the objectives and resource commitments of the stakeholders who are likely to be participating in, or supporting, the case study. (Host et al., 2012)

Data is naturally, and often automatically generated. It, as well as the research, has a high degree of realism but at the expense of the level of control. One gets what one gets. (Host et al., 2012) The design of a case study is very flexible, where the key parameters of the study may be changed during the study. The stages of the research are design of the objectives and the case study itself, preparation of the data collection, collecting the data, analysis and reporting. (Host et al., 2012)

Evaluating a case study can be done based on several indicators based on Host et al. (2012). They state that the study is to be of a significant topic. It needs to be complete in that the boundaries are made explicit, the collection of evidence is comprehensive, and there are no significant constraints on the conduct of the study. Alternative perspectives must be considered. The research must respect ethical, professional, and legal standards relevant to the study. It needs to describe the theoretical basis, offer a chain of evidence with traceable reasons and arguments that are fully documented. It needs to draw inferences from the data to answer the research question. (Host et al., 2012)

Key factor in the ethicality of a case study is informed consent of participants, suitable and agreed upon level of confidentiality. Key factor in the quality of a case study, on the other hand, is that the quality of the data is assessed in all stages of the case study. (Host et al., 2012) The validity of a case study can be approximated from the relative neutrality and reasonable freedom from unacknowledged research biases. Also, the interpretations that are made need to be traceable. (Benedichte Mayer, 2001) A case study can be considered to be reliable if it is consistent and reasonably stable over time and the same findings could be carried out by other researchers (Benedichte Mayer, 2001)

The benefits of a case study are in its ability to gain a holistic view of the issue under investigation. A case study is able to capture the emergent and immanent properties of a phenomenon. (Khairul

baharein, 2008) As it allows to reach a deeper understanding of the phenomena under study, it does not generate the same results on causal relationships as controlled experiments but allow to reach a deeper understanding of the phenomena under study. Therefore, there is an issue with generalizability. Another possible issue is the possibility of bias by the researcher as they are more involved in the process instead of just analyzing it. (Host et al., 2012) There can also be a lack of reliability (Khairul baharein, 2008) and it is possible that the personal involvement of the researcher causes issues with subjectivity (Simons, 2009)

4.2 The design of the current study

The purpose of this study is both exploratory and improving. Host et al. (2012) define the purpose of the exploratory case study as a way to find out what is happening, seeking new insights, generating ideas and hypothesis for new research. Improving, on the other hand is trying to improve a certain aspect of the studied phenomenon. (Host et al., 2012) These both are present in the current research as the goal is to seek insights into the creation of trustworthy AI and generate ideas as well as improve the way AI systems are created. The study is instrumental (see Simons 2009) in that the case is chosen to explore the issue of trustworthiness in AI solutions and to gain insight and understanding on this process of creating trustworthy AI. The study is also theory-led as the case is explored through the theoretical perspective created in the previous chapter. The hypothesis of this research is that using the set of questions created at the end of the previous chapter, it is possible to keep in mind all the various aspects of trustworthy AI during development and that the list allows for checking the work to make sure no aspect is forgotten improving the trustworthiness of the result as well as the process.

The current study is a single-case study, as there is only one instance under observation (Host et al. 2012). The company chosen for the case agreed to participate, participated in several meetings during the process and were continually consulted as to the conclusions draws and inferences made improving the ethicality of the research.

The design of the case study based on Host et al. (2012), firstly, consists of the rationale of the study. Here it is the desire to explore and experiment on the development process of an AI solution to improve its trustworthiness. Secondly the design needs an objective of the study, which is a statement of what the researcher expects to achieve as a result of undertaking that study. The objective of the current research is to test the set of questions developed at the end of the previous chapter as a tool and checklist to guide the development process to ensure trustworthiness. The design, thirdly, needs methods of data collection and analysis as well as a data selection strategy. The data collected here is data owned by Aveti Learning that is gathered in the normal process of their software solution. The data selection strategy here is to utilize all relevant data to answer the question agreed to and with the company in question. Fourthly, a case study needs theory and research questions. In this research, they are already provided in the previous chapter. Finally, a case study needs a way to check the data for quality. Here a constant dialogue with the company was utilized to make sure the data used reflected truthfully to the data they had.

4.3 The Case Organization: Aveti Learning

The focus in this case study is to develop an AI solution for Aveti Learning in a trustworthy way encompassing the important questions from literature in the design, development, and preliminary deployment of the solution, take note of any deviations and describe them as points of future development as no solution is complete and perfect. Due to the need to limit the scope of this thesis, the AI solution will be a proof of concept that is deployed for testing instead of containing the entire lifecycle of the system.

Aveti Learning began as a small initative by Biswajit Nayak called Shikhya in 2014 aiming to bridge the urban-rural gap of quality education in rural India by creating learning content in local Indian languages, so that the content would be accessible and relatable for rural students living in remote areas of India where electricity is inconsistent, and the internet is almost absent. It first began by developing learning centers in villages and orphanages that allowed students to share 5 tablets connected to a local server that contains the learning platform and quality learning material first in math and English, then in sciences and reading. Each learning center has a mentor that can help the students using the Aveti Learning's smart learning curriculum. Currently, there are 18 team

members, 120 centers with mentors, over 15 Indian languages represented, and the smart learning curriculum used also in over 400 schools in Odisha under the official name Aveti Learning. (Barua, 2020) It has been used by over 120 000 students so far.

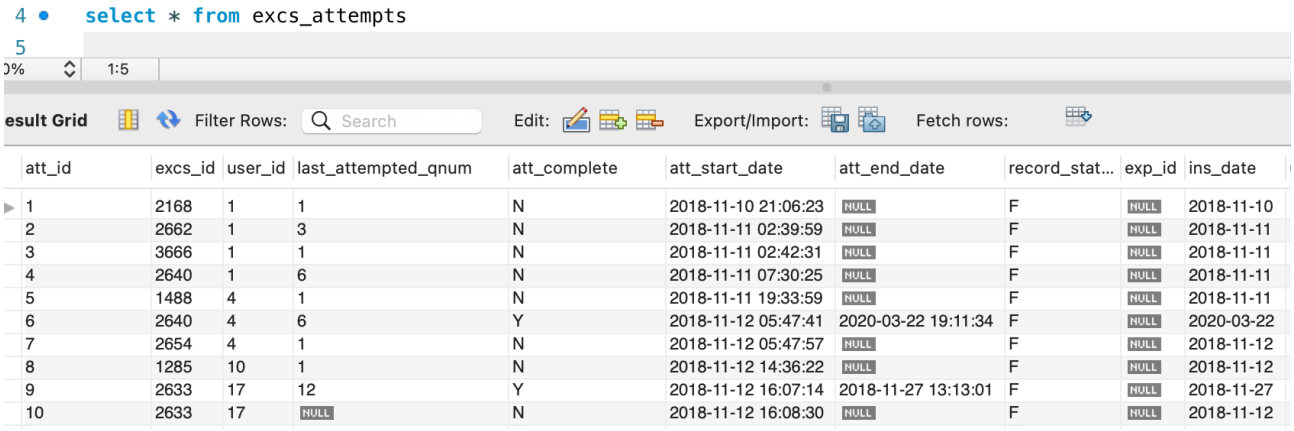
Aveti Learning is built around the concept of mastery, so that the student first goes over the learning material and tests their understanding through practice questions. After 5 questions of the topic are answered correctly, the topic is considered to be mastered and the student can move on.

4.4 Data source, primary data description and limitations created

The data for this case study consists of a copy of the production database of Aveti Learning. The database is the only source of data for this study. As a production database, most of the data contained are of no interest to this study and are therefore left untouched. The database contains information about 91 781 individual accounts that have joined since 11th of November 2018. Of these individuals, 745 have reported themselves to be male and 547 to be female. Of the rest, there is no knowledge of their gender as it is a voluntary field. Of the students, 43 051 are in the 10th grade, 24 194 are in first grade, 11 272 are in the ninth grade, and the rest with the information found on grades 2-8. There is also no information about residency, previous school success, age, or any other information on their background. From the viewpoint of data privacy, this is a good thing as the data collection is focused on only the data needed for the functioning of the learning system. At the same time, it creates a limitation on studying fairness of the learning system and the machine learning model built on this data. It is impossible to study equality based on features that are not in the data. Should this be a topic is special interest, the data on background information should be gathered, but that would have an impact on the data privacy of the users. A middle ground could be selecting some centers and schools at random for a study on the topic, gather background information on those students for the duration of the study, and check the situation and then destroy the information. This is one topic for later development.

The preliminary idea for an AI solution for this case was to identify students at risk of failing so that mentors can contact them and help them in their learning journey. For this use case, a few

tables in the database are especially important. These are *excs_attempts*, which holds the information about the attempt id, exam id, user id, attempt start date and whether the attempt is complete or not meaning if the student demonstrated mastery or not. This table holds information about 210 694 attempts by 37 732 distinct users. An example of this data can be seen in Figure 10.



```
4 •  select * from excs_attempts
5
```

1:5

esult Grid Filter Rows: Search Edit: Export/Import: Fetch rows:

att_id	excs_id	user_id	last_attempted_qnum	att_complete	att_start_date	att_end_date	record_stat...	exp_id	ins_date
1	2168	1	1	N	2018-11-10 21:06:23	NULL	F	NULL	2018-11-10
2	2662	1	3	N	2018-11-11 02:39:59	NULL	F	NULL	2018-11-11
3	3666	1	1	N	2018-11-11 02:42:31	NULL	F	NULL	2018-11-11
4	2640	1	6	N	2018-11-11 07:30:25	NULL	F	NULL	2018-11-11
5	1488	4	1	N	2018-11-11 19:33:59	NULL	F	NULL	2018-11-11
6	2640	4	6	Y	2018-11-12 05:47:41	2020-03-22 19:11:34	F	NULL	2020-03-22
7	2654	4	1	N	2018-11-12 05:47:57	NULL	F	NULL	2018-11-12
8	1285	10	1	N	2018-11-12 14:36:22	NULL	F	NULL	2018-11-12
9	2633	17	12	Y	2018-11-12 16:07:14	2018-11-27 13:13:01	F	NULL	2018-11-27
10	2633	17	NULL	N	2018-11-12 16:08:30	NULL	F	NULL	2018-11-12

Figure 10. Select query and first 10 lines from table *excs_attempts*

Another important table is the *excs_detail*, that holds information about the individual questions in each attempt. It contains the attempt id, user id, question id, submit time, status of correct or not and information about how many the question was within the attempt. If the status of a question is 0 the answer has been incorrect and if it is 1, the answer is correct. This table has 1 370 594 lines of data and information concerning 110 720 distinct exam attempts by 8460 distinct students. This means that nearly half of the exam attempts do not have any line data attached to them in *excs_detail*. An example of the data for *excs_detail* can be seen in Figure 11

```
4 •  select * from excs_detail
5
```

0% 26:4

esult Grid | Filter Rows: Search | Export: | Fetch rows:

att_id	question_num	status	submit_time	record_stat...	exp_id	question_id	user_id
1	1	0	2018-11-10 21:06:23	F	NULL	18401	1
2	1	1	2018-11-11 02:39:59	F	NULL	23767	1
3	1	1	2018-11-11 02:42:31	F	NULL	33117	1
4	1	1	2018-11-11 07:30:26	F	NULL	23531	1
5	1	1	2018-11-11 19:33:59	F	NULL	10929	4
6	1	1	2018-11-12 05:47:41	F	NULL	23531	4
7	1	1	2018-11-12 05:47:57	F	NULL	23681	4
8	1	0	2018-11-12 14:36:22	F	NULL	8967	10
9	1	0	2018-11-12 16:07:14	F	NULL	23478	17
9	2	1	2018-11-12 16:08:30	F	NULL	23479	17

Figure 11. Select query and first 10 lines from table *excs_detail*

The preliminary idea based on the column descriptions and preliminary information of what should be in the database was to utilize exam start times and exam end times divided by the number of questions to serve as a proxy for the difficulty of the exam for the student. However, about 45 000 exam attempts that are marked as completed have the end time missing. This means that information about the duration of an exam is too unreliable to be used. Also, the idea was to utilize information on the use of hints to serve as a proxy for the difficulty felt by the student and hence indicate on their mastery of the concept and problems faced. However, information on the hints was found to be missing.

Finally, there is a table called *video_view_log* that contains information on the teaching videos watched, user id, and the length of view time and the date of the viewing. It contains information about 5788 distinct users and 4091 videos. Due to exams attempted by 8460 students and videos watched by only 5788 students, also this data is not used at this time. There is no mapping done at this time how the videos relate to the exams but building a mapping between videos, the material, and the questions measuring mastery of the topics in the videos is one direction for further development. It would allow relating the exam results with the length and number of times the videos are watched. This could allow for personalized feedback based on this information. Also, it would allow suggesting videos that have helped others with similar problems by a recommendation algorithm. Further, saving the students' answers to the database is not done at this time. Should they be gathered in the future, it could be possible to create an algorithm that could look for similar mistakes and again, offer content that has helped others who make similar mistakes.

4.5 Trustworthy AI questions within AI lifecycle and the focus of the case study

This case study focuses on the proof-of-concept phase of an AI system's lifecycle. This encompasses the ideation and design phase, data processing, model construction, and testing. The model is deployed to be used for user testing and documentation created to enable and facilitate the testing is created. Should the model itself be robust and trustworthy enough to consider deployment, user testing an evaluation will be made whether to deploy the solution to end users or develop it further. The focus of this thesis ends at that evaluation, so solution will not be deployed to be used by end-users within the scope of this thesis. Throughout the stages within this thesis a series of questions are asked, and answers sought to facilitate in the creation of a trustworthy AI solution. These questions are created from the previous listing and focused on specific areas in the development lifecycle and form the research questions for this case study. These questions are as follows:

1. Design-stage:
    a. What is the purpose of the AI solution?
        i. Is it a socially accepted purpose?
        ii. Does it increase social fairness?
        iii. Does it increase human autonomy?
        iv. Is the outcome socially beneficial?
        v. Is the benefit distributed equitably and how is this defined and measured?
        vi. What are the risks involved and are they acceptable in relation to the purpose? Is there a possibility of harm and how is that mitigated?
    b. Design of solution
        i. How is the human kept in the loop?
        ii. How is the data protected?
        iii. Are there risks to safety and how are they mitigated?
        iv. How is failure of the AI component guarded against? What are possible reasons for failure and how are they mitigated?
        v. What is anticipatory accountability in this project and are there possible remediation and redress?
        vi. How is decision provenance built into the system?
        vii. What is the acceptable accuracy level of the chosen performance metric?
        viii. How diverse is the team in charge?
        ix. How can users report problems?
        x. What type of models can be used?
        xi. What models are chosen to be tested?
        xii. What is the degree of autonomy in the decision making?
    c. Communication with stakeholders
        i. How is the purpose and design communicated with different stakeholders?
        ii. How transparent is the design of the system?

2. Data processing stage:
    a. What unfairness-mitigation techniques is possible to use with this data?
    b. Is the data labeled? Is it feasible to label if it isn't?
    c. How is the data processed?
    d. What is the quality of the data?
    e. How is the data governed?
3. Model construction:
    a. What is the baseline performance on a simple and explainable model?
    b. What performance metric is best fitting and why?
    c. Is model overfitting or underfitting?
    d. Is the model robust and explainable?
    e. Interpretation: What methods can be used for interpretability?
4. Testing and evaluation:
    a. What to test against?
    b. Performance metric on test data, if applicable
5. Explanation:
    a. How to explain results?
    b. What to teach about AI to different stakeholders so they may understand the system being used?
6. Deployment:
    a. What is the possible attack surface?
    b. How are the risks for safety mitigated against?
    c. How often and how to retrain the model due to possible model drift?
    d. How is failure noticed and what happens next?
    e. How can users report problems?
7. User testing and feedback:
    a. Have the users understood the material?
    b. Can they use the system or are there issues?
    c. What feedback do they provide and how will it shape the further development?
8. Future development: What can be learned from the proof of concept and what can be improved upon?

5 Trustworthy AI solution to identifying students at risk in rural India

5.1 The Design of the AI solution

The purpose of the AI solution is to further enhance the capability of students in rural India with limited resources to learn and advance in their lives. This is a socially acceptable purpose, and it increases social fairness by helping those advance who would otherwise struggle. By helping students understand math and science concepts, learn to read better, and communicate in English, it increases the autonomy of the students. For them, if the AI solution delivers on the goal, the outcome is socially beneficial. However, should the solution fail and misidentify students needing

help, some of them could be left more alone with their studies than they currently are by directing the attention and help of the mentors away from them. So, there is a clear risk in failing the students relying on this system. A risk here is to misidentify students needing help. It is less harmful to offer a student who is doing okay help than it is to not offer help to a student needing it. A way to mitigate this would be to add into the learning platform a way for a student to request help from their mentor.

The students are not the only stakeholder group to consider. Another clear stakeholder group is the mentors. It can be difficult to see, who is struggling and why and offer the suitable help at the right time. This AI solution can help them by offering advice on the students with whom they should check in with. It does not, however, offer advice on how to best do this and where the problems are. At its worst the solution can offer them a list of students that is too long with information that is too hard to understand making their job harder than it is. Mitigating this problem is possible through user testing and feedback as well as educating them on AI and how it functions as well as the limitations of the system.

The administrators of the learning platform are also a stakeholder group. With basic data analytics on the material provided it is possible to identify those exams where failure is common and improve the learning material on them. Also, it is straightforward to produce failure rates for individual questions helping the administrators and content creators to review at the questions again. However, the high amount of missing line items makes this a process rife with uncertainties. Hence, it is advisable to first make modifications in the system to better improve the data and only then analyze the line data better.

For the benefit to distribute equally amongst those in need of help, there needs to be a logging procedure put into place that needs to capture the prediction, the action by the mentor or lack thereof and follow-up on the success of the student. If these actions are not logged in some way, there is little to no chance to see how the system helps or, if indeed, it does. Also, a feedback route provided to the mentors as well as the students could allow for the development of the AI system after the proof-of-concept phase. This type of logging also enables decision provenance as it is possible to demonstrate the reason for action, the action, and the result of the action.

To measure the benefit of the AI system or lack thereof, one possible way is to look at the overall improvement of the students. This, however, would not show cause and effect, but rather a simple correlation. An AB testing would be required with similar groups of students, half of which get help from their mentors without the help of the AI system and another half where the mentors are augmented with the AI system.

In this solution, the human is always in the loop, as it is a human who acts based on the results or does not act. It is the mentor that decides how to react; therefore, it is not an autonomous system at all, but rather augments the skills of the mentor by analyzing the data created by the students and offering further information based on the results.

The data used in the modeling procedure does not contain any sensitive information on the students. The only information available is their id, how many questions they answered correctly and how many incorrectly and if they passed the exam. This data is kept behind access controls to keep it safe from poisoning or tampering. It is the data used in the learning platform allowing the possibility for mentors, teachers, and students to see their data and flag it for correction if there are inaccuracies in it.

The biggest risk to the AI solution comes from the elements. With electricity shortages and limitations to access the internet, a second system needs to be built to act as a way to offer the mentors the same service even when it is impossible to update models due to lack of connection. This secondary system utilizing local data can also work as a way to test the predictions of the AI solution. This secondary system can also kick in in the case of for example DDoS attack. Further, the proof of concept will be deployed as an API endpoint. Limitations can be built to guard against unauthorized use and brute force solutions by limiting the number of queries from the same IP within a limited timeframe.

Anticipatory accountability is shown in the upmost interest of the entire and very diverse team to build something that helps those less fortunate than themselves. The team has people from very different cultures, backgrounds, genders, and educational backgrounds. Great care will be shown in the design, creation, deployment, and testing of the system. Further, a way for the mentors and the students to report problems needs to be built into the learning platform so that problems can

be discovered as early as possible, and actions taken promptly. This can act as a possible remediation as students in need of help can request it.

The data does not contain information on the students that are failing or thriving. This must be deduced from the little data available. The sparsity of data available does not allow for a lot of feature processing. There is data on the number of correct responses, indicating mastery, and the number of tries of a particular exam, indicating perseverance. Both mastery and perseverance are needed for learning. The group most in need of help would be those that have high perseverance and low mastery as they are probably the group with the most to gain and a good group to start with.

There are different ways to approach this type of data. One approach is to use unsupervised approaches, for instance utilize a K-Means algorithm to find groups of students that differ from another. Then the group with high perseverance and low mastery can be contacted by the mentors to offer them further aid. This is a fairly straightforward modeling with high interpretability and is therefore selected as one approach. It can start the development process which can be further developed with more data collected to allow for a more sophisticated approach. Another is to label the data in some way based on the students' previous success and utilize a supervised learning approach. This can be attempted and can also be built on top of the unsupervised approach. Third approach is to predict the number of correct replies a student is likely to produce in their next exam. From this it would be possible to contact those students whose prediction of correct replies is very low. This approach requires a lot of sequential data points, which would further dramatically decrease the amount of data used while also limiting the generalizability of the predictions. This idea is then better suited for further development options than utilization at this time. The K-Means will be created, with possible classification system as well, and the results are compared to a system based on simple analysis techniques of previous success. The overall preliminary design of the system is shown in Figure 12. In the design the data is extracted from the database with SQL-queries, preprocessed and then modelled. The model is evaluated and interpreted after which a choice is made about going forward with the model or solution. The development code is refactored for production and the machine learning model is deployed as a REST API to allow for incorporation into the learning platform. Also a more secure system that is even easier to incorporate into the edge devices is created.

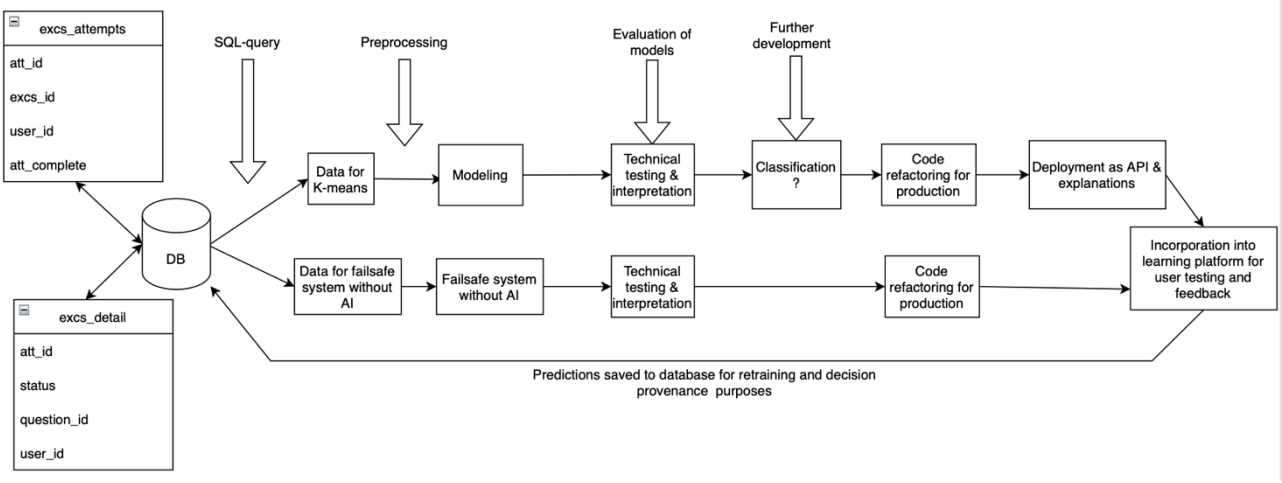


Figure 12. Overall preliminary design of the AI system

The design, development, and deployment of this AI system will be very transparent as this thesis of its' creation will be public and open to anyone to view. This, in and of itself, is not enough as the purpose and design of the system will need to be communicated with stakeholders. They will also need to be taught to understand the technology enough for them to understand the limitations of this system. This process, however, takes place after the proof-of-concept phase.

5.2 Data exploration and processing

The data for models in this thesis was retrieved from the database with an SQL-query in code block 1 below. As previously stated, the *excs_detail* contains the status, which informs whether the answer was correct or not. This information was further processed in a common table expression (CTE) called *excs* to calculate the number of correct answers, incorrect answers, and answers in total to each exam. This information was then combined with the *excs_attempts* data with a left join meaning that it takes all the lines in *excs_attempts* here as that table is mentioned first in the main select and combines the line data in *excs_detail* to these attempts. Should there be lines with *att_id* that is not in the *excs_attempts*, that would not appear in the data with left join, but it was checked that this is not the case, and all line data is present. The percentage of correct answers and percentage of wrong answers are calculated in the joined data, though it could also be done in the CTE. The used SQL-query can be seen in Code Block 1.

```
with excs (att_id, user_id, correct, total, wrong)
as
(select att_id,
user_id,
sum(status) as ed_correct,
count(status) as ed_total,
count(status) - sum(status) as ed_wrong
from excs_detail
group by att_id, user_id)

select ea.att_id,
ea.user_id,
att_complete,
ed.correct,
ed.wrong,
ed.correct/ed.total as percentage_correct,
ed.wrong/ed.total as percentage_wrong,
ed.total
from excs_attempts ea
left join excs ed on ea.att_id = ed.att_id
group by ea.att_id,
ea.user_id
```

Code block 1. SQL-query to abstract data for preprocessing and modeling from the database

This extraction and processing give 210 694 rows of data with many missing values as only 110 720 exam attempts have line data. As previously described, the data lacks classifying information on any sensitive information and hence it is quite impossible to utilize unfairness-mitigation techniques in this data. For the development team to be able to check for discrimination, sensitive information needs to be gathered. However, this encroaches on privacy. For this data the line was drawn in favor of privacy so, the data was not collected. The data is also not labelled, so many supervised learning algorithms would require several arduous decisions and processes to proceed to modelling. Access to this data is governed by credentials and all forms of processing are documented, which is sufficient governance at this time. An example of the data can be seen in Figure 13. *Att_complete* has “Y” if the attempt is successful and “N” if it has not been successful.

att_id	user_id	att_complete	correct	wrong	percentage_corr...	percentage_wro...	total
1	1	N	0	1	0.0000	1.0000	1
2	1	N	2	1	0.6667	0.3333	3
3	1	N	1	0	1.0000	0.0000	1
4	1	N	4	2	0.6667	0.3333	6
5	4	N	1	0	1.0000	0.0000	1
6	4	Y	6	0	1.0000	0.0000	6
7	4	N	1	0	1.0000	0.0000	1
8	10	N	0	1	0.0000	1.0000	1
9	17	Y	50	21	0.7042	0.2958	71
10	17	N	NULL	NULL	NULL	NULL	NULL

Figure 13. The first 10 rows of the data as an example

Looking at the descriptives of the data in Figure 14, it seems, there are more correct (mean 7,06) than incorrect answers (mean 5,30) and the deviation of both variables is quite large (correct st.dev 12,32, incorrect st.dev. 10,41). The maximum value of correct responses is 2013, which seems unlikely to occur within the normal use of the learning platform. Also, the maximum of 838 incorrect answers, seems likely to be an outlier rather than an actual value. Descriptives, count, mean, standard deviation, minimum value, maximum value and values at quartiles 25%, 50% and 75% are given are given also for *att_id* and *user_id,* but are unnecessary for analysis. An outlier was determined to be 3 standard deviations away from quartile 3, which is the line indicating that 75% of data is below this quartile point. 3 standard deviations from quartile 3 is referred to as an extreme outlier in statistics. This extreme outlier for correct responses is all points above 78 per exam, and for incorrect responses above 54 per exam. The SQL-statement was altered to drop the outliers and can be found in Code Block 2.

	count	mean	std	min	25%	50%	75%	max
att_id	210694.0	112550.289861	61930.381101	1.0	60189.2500	113274.50	165984.7500	218680.0
user_id	210694.0	38241.590363	28761.675348	-1.0	11040.0000	30187.00	65531.0000	94027.0
correct	110714.0	7.075772	12.320631	0.0	2.0000	5.00	9.0000	2013.0
wrong	110714.0	5.303584	10.410782	0.0	1.0000	3.00	6.0000	838.0
percentage_correct	110714.0	0.565917	0.335669	0.0	0.3333	0.64	0.8333	1.0
percentage_wrong	110714.0	0.434083	0.335669	0.0	0.1667	0.36	0.6667	1.0
total	110714.0	12.379356	18.794916	1.0	5.0000	10.00	14.0000	2640.0

Figure 14. Descriptives of the original dataset

Also, here, too, it is clear to see the large amount of missing lines as the count of *att_id* and *user_id* is 210 694, but of the rest only 110 714. This means that a large percentage of the line

data is missing for some reason or another. This brings with it two different ways of interpreting the situation. Either, for some exam attempts, all the lines are missing, but for exams that have lines, all lines are present. Or then lines can also be missing from attempts that have line data. It is impossible to say, which is the case or if, as is likely, both are.

```
with excs (att_id, user_id, correct, total, wrong)
as
(select att_id,
user_id,
sum(status) as ed_correct,
count(status) as ed_total,
count(status) - sum(status) as ed_wrong
from excs_detail
group by att_id, user_id
having sum(status) < 78 and count(status) - sum(status) < 54
)

select ea.att_id,
ea.user_id,
att_complete,
ed.correct,
ed.wrong,
ed.correct/ed.total as percentage_correct,
ed.wrong/ed.total as percentage_wrong,
ed.total
from excs_attempts ea
left join excs ed on ea.att_id = ed.att_id
group by ea.att_id, ea.user_id
```

Code block 2. Modified SQL to drop outliers

Dropping the outliers dropped 607 lines of line data as can be seen in Figure 15 leaving 110 107 exams with line data attached. The tendency to have more correct than incorrect answers in an exam is still visible in the data. The average of the number of questions students answer is 11, but with almost as large a variation. Average of correct answers is 6,78 (st.dev. 7,18) and of incorrect answers is 4,86 (st.dev. 6,48). Figure 16 shows the distribution of the data for the incorrect answers and the correct answers.

	count	mean	std	min	25%	50%	75%	max
att_id	210694.0	112550.289861	61930.381101	1.0	60189.2500	113274.5000	165984.7500	218680.0
user_id	210694.0	38241.590363	28761.675348	-1.0	11040.0000	30187.0000	65531.0000	94027.0
correct	110107.0	6.781095	7.176388	0.0	2.0000	5.0000	9.0000	79.0
wrong	110107.0	4.857293	6.484535	0.0	1.0000	3.0000	6.0000	53.0
percentage_correct	110107.0	0.567337	0.335089	0.0	0.3333	0.6429	0.8333	1.0
percentage_wrong	110107.0	0.432663	0.335089	0.0	0.1667	0.3571	0.6667	1.0
total	110107.0	11.638388	10.170311	1.0	5.0000	10.0000	14.0000	130.0

Figure 15. Descriptives without the outliers

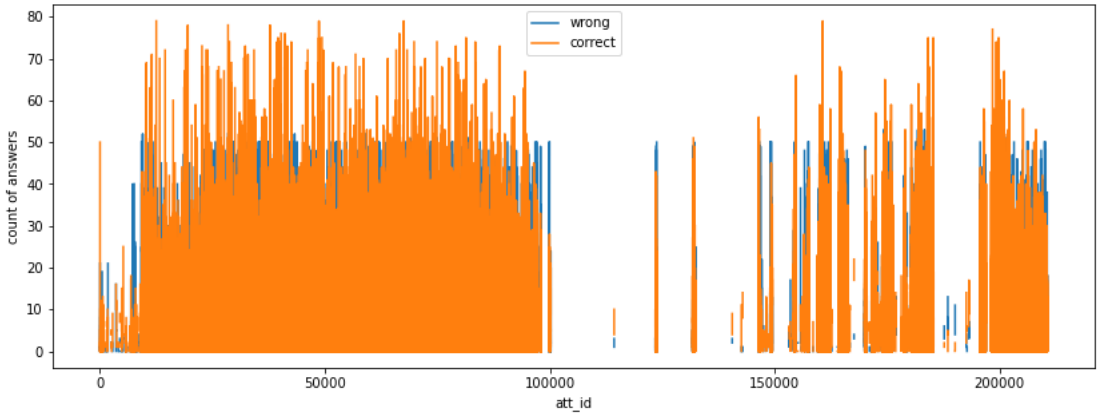


Figure 16. Preliminary imaging of data. Amount of wrong and correct answers based on attempt id

Looking back to data quality dimensions, this data is accessible as it is available and easily retrievable. The volume of data is also sufficient for the task at hand, even if only the attempts that have line data are used. The data is not completely believable and free of errors, as there seems to be outliers, but mostly the data that is present does seem to be accurate. The biggest issue to data quality is completeness, which is the extent to which data is not missing. Here nearly half of line data is missing. This creates a clear problem from the viewpoint of data quality and causes any results of the modeling to be viewed with some measure of unreliability. The data is secure, timely, interpretable, and understandable. It is also fairly relevant, although more data would bring more possibilities to focus on learning. The code for this exploratory data analysis (EDA1) is in Appendix 3, Code Block 8. Due to the sizable problem with missing data, the developed model is unlikely to be deployed, although it is developed here to offer a starting point for further development once the issue with the missing data is fixed and a more complete dataset is gathered. This reduces the need to thoroughly test the built model here as the need for testing is transferred to the point in future with a higher quality dataset.

5.3 K-Means cluster and Random Forest Classifier

5.3.1 Machine learning problem formulation

The main idea is to predict which students need help. This is not a yes or no question but a matter of who needs more and who needs less. Firstly, the problem requires that the data is viewed not from the viewpoint of exam attempt as was shown above but from the viewpoint of the student. Secondly, as this is not measured directly, suitable proxies for need for help are required.

Two distinct points of information can act as proxies of learning in this data. One is the relationship between wrong answers and right answers. If a student frequently answers incorrectly, they have not mastered the material and may need help to do so. For this, intuitively, a high rate of errors is a suitable proxy for needing help. Another is the average number of attempts that is required to take exams before mastery is shown. If the number of attempts is very high, this shows problems in the mastery but also perseverance. It serves as a proxy for someone who is really trying but is just not succeeding. Here a high number indicates both problems in mastery and perseverance that is beneficial in accepting help.

A final consideration before data extraction is the choice of the scope of exams. A student may struggle with an individual exam or with more than one. In this proof-of-concept stage the focus is on offering help to those students that struggle in general, so the need is to extract the average difficulty for the student and use that for modelling. In later development, as more line data is captured, it may be beneficial to focus on only the latest datapoints, such as for instance the last 5 or 10 exams.

5.3.2 Data extraction and preprocessing

So, the problem formulation is to extract the average number of attempts per exams and the average percentage of wrong answers. Necessary query to fetch this data is shown in Code Block 3. The data produces the number of attempts, the number of exams, average attempts per exam, correct responses, incorrect responses, percentage of incorrect and correct responses and the total responses for 4264 students with more than 1 exam try in the database. This should give enough information for modelling for this proof-of-concept phase and testing the concept, but a

reliable model would need a dataset with less missing data. This produces the data that can be viewed from Figure 17.

```
with excs (user_id, correct, total, wrong) as
(select user_id,
sum(status) as ed_correct,
count(status) as ed_total,
count(status) - sum(status) as ed_wrong
from excs_detail
group by user_id
having sum(status) < 78 and count(status) - sum(status) < 54)
,
attempts (user_id, n_attempts, n_exams, avg_attempts) as
(select user_id,
count(att_id) as n_attempts,
count(distinct(excs_id)) as n_exams,
count(att_id)/count(distinct(excs_id)) as avg_attempts
from excs_attempts)

select
ea.user_id,
ea.n_attempts,
ea.n_exams,
ea.avg_attempts,
ed.correct,
ed.wrong,
ed.correct/ed.total as percentage_correct,
ed.wrong/ed.total as percentage_wrong,
ed.total
from attempts ea
left join excs ed on ea.user_id = ed.user_id
where correct is not null and n_attempts > 1
group by ea.user_id
```

Code block 3. SQL-query to fetch the necessary data for K-means cluster analysis

user_id	n_attempts	n_exams	avg_attempts	correct	wrong	percentage_corr...	percentage_wro...	total
6	6	5	1.2000	7	5	0.5833	0.4167	12
10	4	4	1.0000	0	4	0.0000	1.0000	4
35	3	3	1.0000	1	2	0.3333	0.6667	3
41	5	3	1.6667	2	3	0.4000	0.6000	5
44	3	1	3.0000	2	1	0.6667	0.3333	3
55	73	21	3.4762	28	45	0.3836	0.6164	73
82	6	1	6.0000	1	5	0.1667	0.8333	6
84	4	2	2.0000	2	2	0.5000	0.5000	4
110	2	1	2.0000	0	2	0.0000	1.0000	2
122	3	3	1.0000	8	42	0.1600	0.8400	50

Figure 17. Example of dataset for K-means

Looking at the data more closely, the average value of exams attempted is approximately 5 and the number of attempts is the same. The average number of correct answers is about 14 and of wrong answers is about 20. The average total amount of answers is about 34. Figure 18 shows the unmodified columns in the data and their average values.

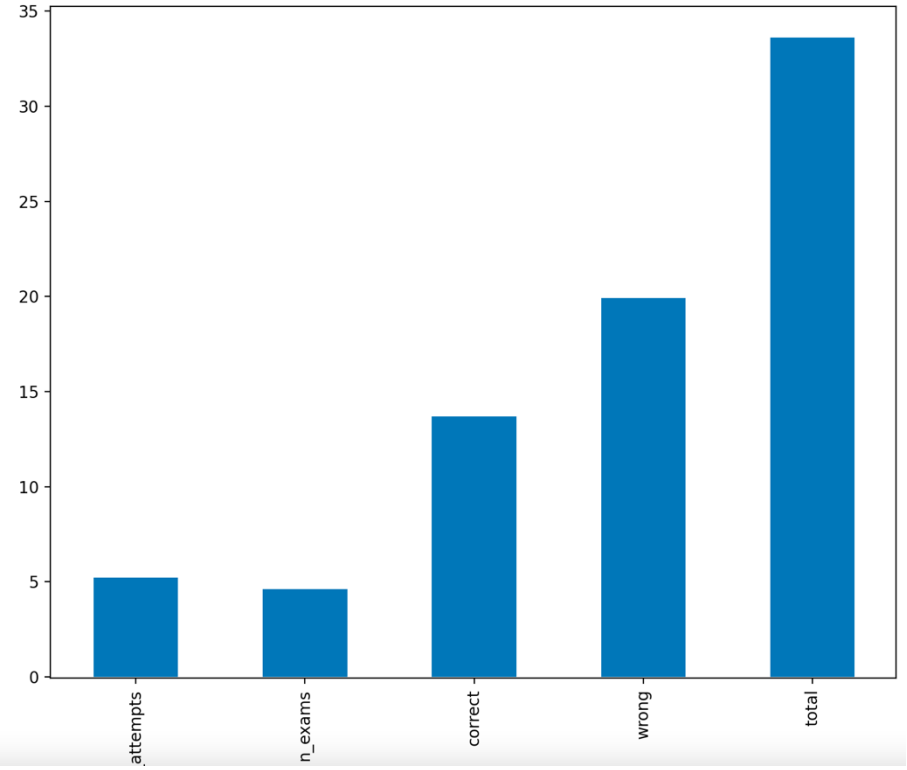


Figure 18. Averages of unmodified columns in data

Averages, however, only tell so much as distribution is also very important. From the histograms in Figure 19 it is clear to see that the distributions of all the unmodified variables in the data are very skewed and not even remotely resembling a normal distribution. The horizontal axis shows values for the variable shown in the legend. The code for the EDA for K-Means algorithm can be found in Appendix 3 Code Block 9.

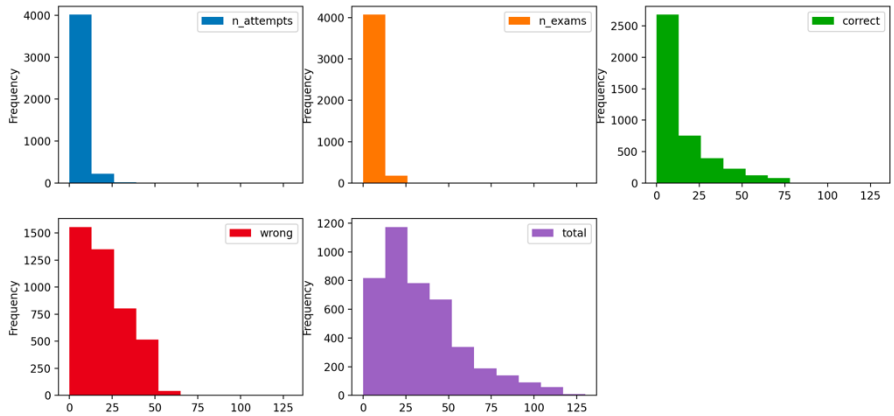


Figure 19. Histograms of 5 unmodified columns in data

5.3.3 Modelling

K-Means machine learning algorithm is a distance-based algorithm, which means that the difference of magnitude between the variables can create a problem, so the data needs to be standardized to get all variables to the same magnitude. This is done by subtracting the mean of each variable and then scaling it to unit variance. Further data processing is unnecessary for the K-Means algorithm. This is a simple, but powerful algorithm that can be used in recommendation engines, customer segmentation, and getting to know the subjects, as it is used here. The K-Means algorithm is both robust and explainable and hence a good starting point for this data.

The appropriate number of clusters created can be checked by comparing the Dunn index or inertia for different options. Inertia was used here, and it calculates the sum of distances of all the points within a cluster from the centroid of that cluster and sums them. The inertia should be as small as possible. From Figure 20 it is observable than cluster number between 3 and 7 would work. The number was chosen to be 5 after some experimentation. This resulted in the following groups: group 0 had 622 students, group 1 had 1356, group 2 had 907, group 3 had 1245, and finally group 4 had 134 students. This number of clusters was used to implement the K-Means algorithm. The code of the execution can be found in Code Block 4.

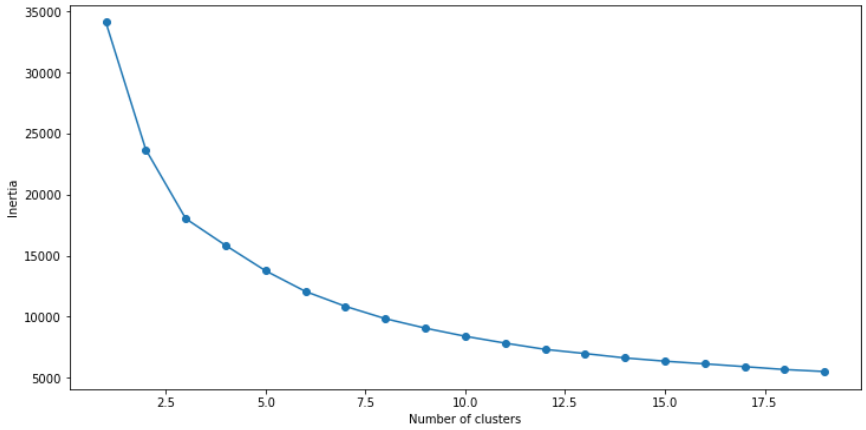


Figure 20. Inertia values for different number of clusters

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

data = pd.read_csv('data_user_nonull_over1.csv')
df = data.drop('user_id', axis=1) # drop user_id
print(df.head())
print(df.columns)

# scaling the data
scaler = StandardScaler()
df_scaled = scaler.fit_transform(df)

# fitting multiple k-means algorithms and storing the values in an empty list
SSE = []

for cluster in range(1, 20):
    kmeans = KMeans(n_jobs = -1, n_clusters= cluster, init='k-means++')
    kmeans.fit(df_scaled)
    SSE.append(kmeans.inertia_)

# plotting the results
frame = pd.DataFrame({'Cluster':range(1,20), 'SSE':SSE})
plt.figure(figsize=(12,6))
plt.plot(frame['Cluster'], frame['SSE'], marker='o')
plt.xlabel('Number of clusters')
plt.ylabel('Inertia')

# KMeans using 5 clusters and k-means++ initialization
kmeans = KMeans(n_jobs=-1, n_clusters=5, init='k-means++')
kmeans.fit(df_scaled)
pred = kmeans.predict(df_scaled)

# examining value count of points in each cluster
frame = pd.DataFrame(df_scaled)
frame['cluster'] = pred
print(frame['cluster'].value_counts())

# examining the clusters
data['cluster'] = pred
print(data.groupby('cluster')['n_attempts','n_exams','avg_attempts', \
                              'correct', 'wrong','percentage_wrong', \
                                  'percentage_correct', 'total'].mean().T)
```

Code block 4. K-Means cluster algorithm execution with commentation

5.3.4 Interpretability

Next task is to understand the results and an easy first look is through the means of each variable in each group. It enables one to create a description of each cluster (Figure 21). In figure 21 the correct, wrong and total refer each to the amount of questions answered correctly, incorrectly and in total. Cluster 0 with 622 students has individuals that have the second largest number of attempts and of wrong and correct answers. Their total number of questions is largest of all clusters. They are clearly very engaged in the learning platform and get more right than wrong. Cluster 1 with 1356 students has the lowest number of total questions answered, very few of them correct and many more wrong. They have the lowest count of attempts. While this group seems to need help getting started, it may not be the group that most benefits from it if resources are very scarce. However, should help be given, most benefit could be achieved by helping them get started. Cluster 2 with 907 students has more attempts in total and they have the highest average attempt per exam. Their percentage of wrong answers is the highest and the total amount of questions answered is the second highest. It seems this group is quite engaged and persevere but are not succeeding very well. This group could benefit from extra help from mentors.

cluster	0	1	2	3	4
n_attempts	9.294212	3.087758	4.746417	4.085141	21.723881
n_exams	8.065916	2.749263	3.977949	3.797590	18.947761
avg_attempts	1.209147	1.232353	1.286507	1.128451	1.229430
correct	45.528939	1.609882	6.608600	15.683534	17.805970
wrong	30.893891	13.603982	36.465270	9.108434	21.671642
percentage_wrong	0.404659	0.899255	0.860434	0.369905	0.605658
percentage_correct	0.595341	0.100745	0.139567	0.630096	0.394343
total	76.422830	15.213864	43.073870	24.791968	39.477612

Figure 21. Means of each variable in each cluster

Cluster 3 with 1245 students answer more questions right than wrong but have the second lowest total number of answered questions and the second lowest number of exams. This is a group of students that have not used the platform very much so far and are doing okay. Cluster 4 with 134 students have the highest number of attempts with a slight majority of questions answered incorrectly. This group would benefit from help as they are very engaged, but they get more answers wrong and right. The differences are easier to spot with an image, so Figure 22 shows the different columns and the mean value of each for the main columns of the dataset. One thing that is clearly

noticeable in this image is that in groups 1 and 2 the percentage of incorrect answers is quite clearly bigger than the percentage of correct answers. The same is true for cluster 4 but less clearly.

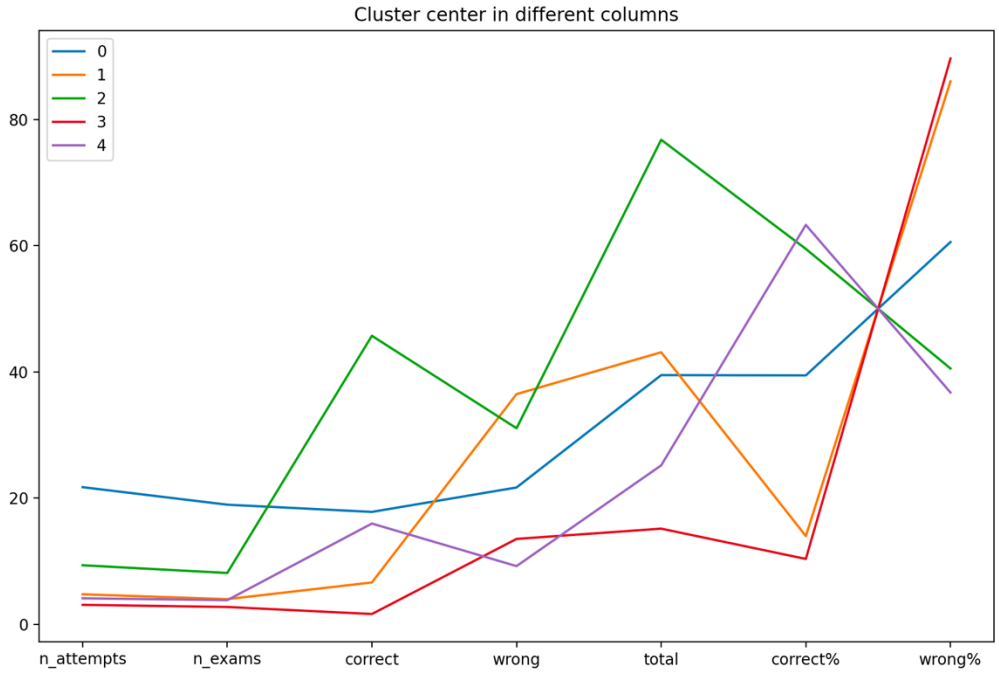


Figure 22. Visualisation of the different clusters based on average values for each cluster

The K-Means cluster algorithm is an explainable algorithm where some idea of the nature of the clusters can be gleaned from the cluster centers. Interpreting the meaning of the clusters boils down to characterizing the clusters as done above. It is possible to get a clearer view by utilizing principal component analysis (PCA) for imaging. The code for this can be found in Appendix 3 in Code Block 11. It is quite difficult to create an image with as many dimensions as is used here as each variable is one dimension. So, the variables were mapped onto a 2D vector space (Figure 23). The code for this is in appendix 3 Code Block 11 and 12.

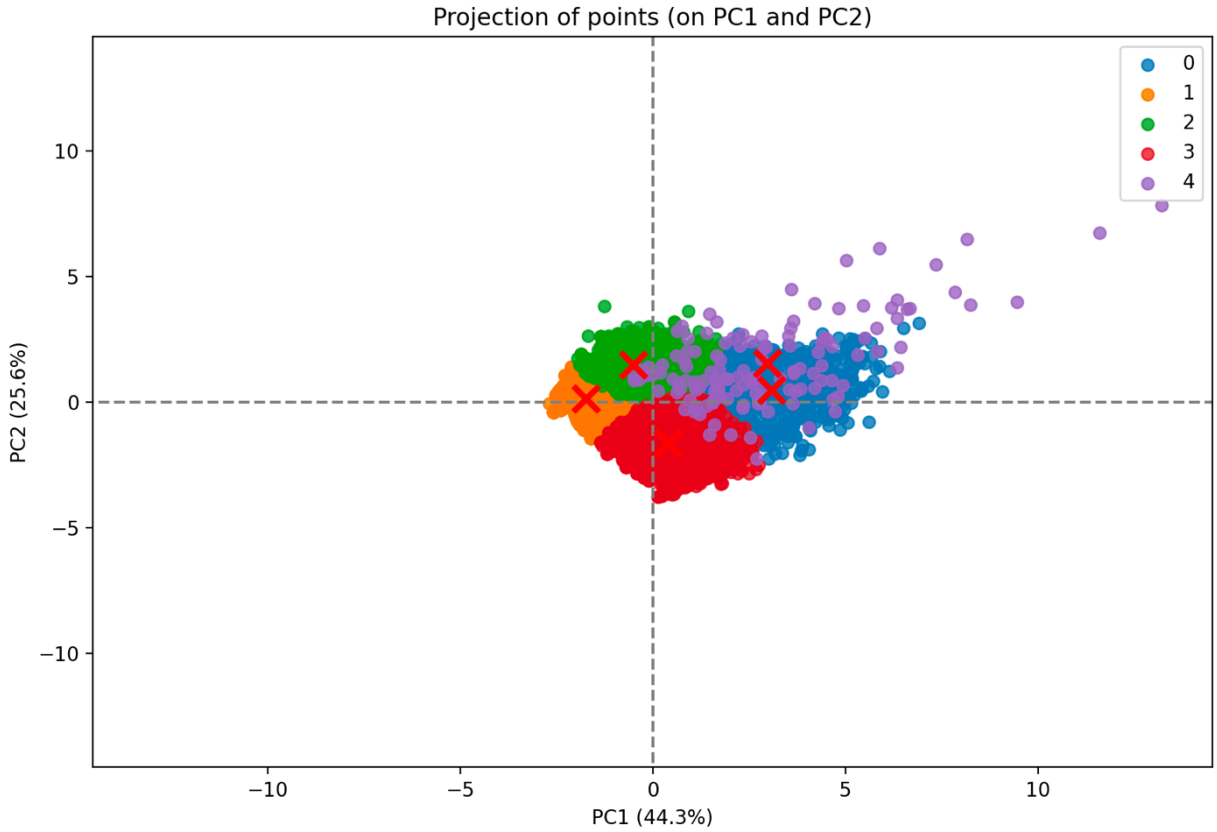


Figure 23. Clusters and their centroids in 2D

The Parallel Coordinates Plot can help further to characterize the clusters (Figure 24). The code for this can be found in Appendix 3 in Code Block 13 and 14. It shows how individual data points sit across all the variables. A plot can be drawn for each cluster and that can help get a feel for what the clusters actually represent. From the figure it is fairly clear that clusters 1 and 2 have more answers wrong than right and that in group 3 this setting is clearly reversed, while in clusters 0 and 4 the difference between right and wrong answers is more even. However, from both Figure 23 and 24 it is clear to see that the data points overlap without a clear divide between them. It is then not so clear why a specific datapoint is for instance in cluster 0 instead of cluster 4.

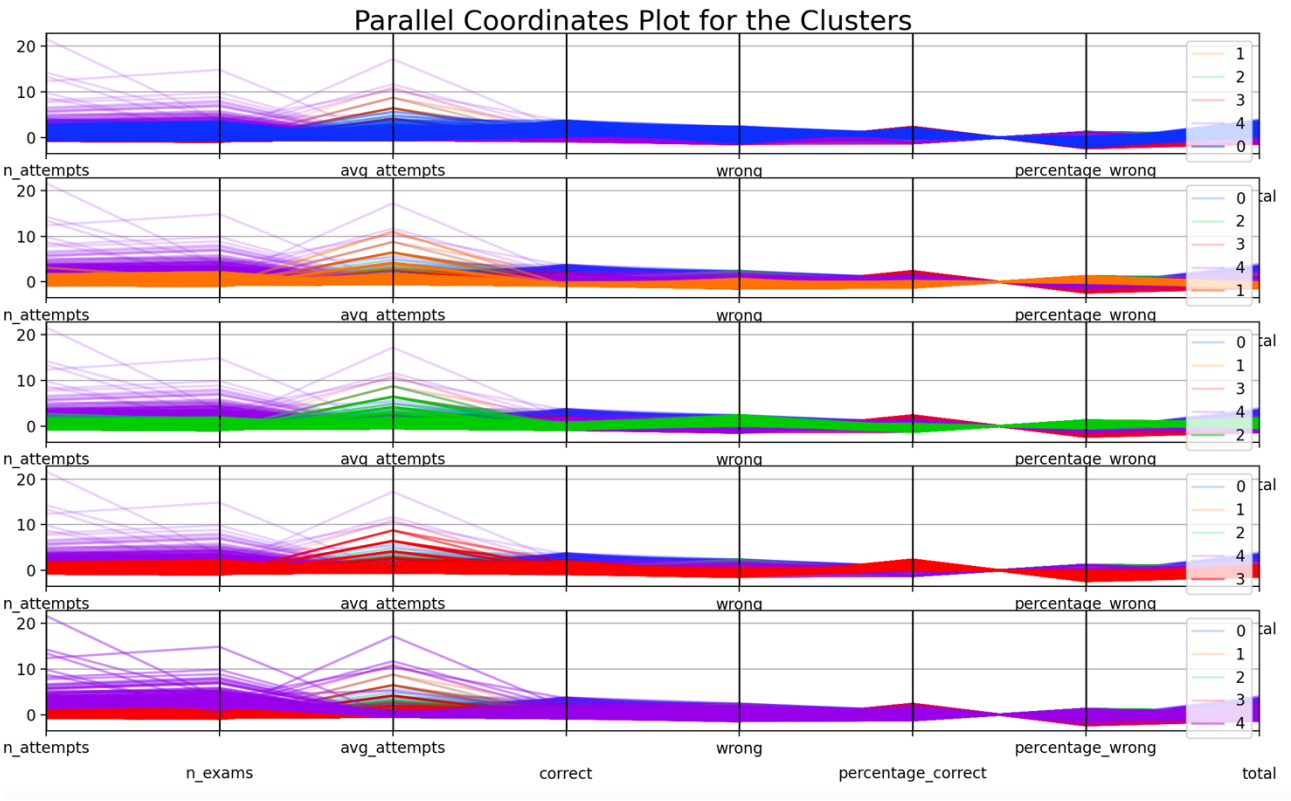


Figure 24. Individual lines in the dataset from the viewpoint of clusters in all variables

To further clarify the differences between the clusters, it is possible to use the clusters as labels and utilize a supervised learning algorithm, like random forest, with SHAP for instance to give further information on the importance of each feature, or variable, to each class. The code is in Code Block 5. This code first splits the dataset into test and train datasets and then trains the random forest classifier on the train dataset. Then it creates a crosstabulation to see how well each cluster is predicted in the test dataset (Figure 25). It shows that the classifier is very accurate, as there are only a very small number of instances that are misclassified. This classifier can also be used to make predictions on new datapoints.

Predicted cluster	0	1	2	3	4
Actual cluster					
0	121	0	1	1	0
1	0	281	2	0	0
2	1	1	174	1	0
3	1	2	1	241	0
4	2	0	1	2	20

Figure 25. Crosstabs of predicted and actual cluster labels in test data of the Random Forest Classifier

Then the performance of the classifier is assessed through several metrics. Figure 26 shows them for each cluster. Precision attempts to answer the question of what proportion of positive identifications are actually correct. Recall attempts to answer the question of what proportion of actual positives was identified correctly. These two are often in tension meaning that improving one may reduce the other. The F1 score is the weighted average of precision and recall and therefore takes both false positives and false negatives into account. Accuracy attempts to answer the question of what is the fraction of all predictions our model got right.

	precision	recall	f1-score
0	0.96800	0.98374	0.97581
1	0.98944	0.99293	0.99118
2	0.97207	0.98305	0.97753
3	0.98367	0.98367	0.98367
4	1.00000	0.80000	0.88889
accuracy			0.98124
macro avg	0.98264	0.94868	0.96342
weighted avg	0.98140	0.98124	0.98098

Figure 26. Performance metrics of the Random Forest Classifier

Shapley values for this dataset are as follows: [('n_attempts', 0.083), ('n_exams', 0.063), ('avg_attempts', 0.009), ('correct', 0.100), ('wrong', 0.153), ('percentage_correct', 0.191), ('percentage_wrong', 0.202), ('total', 0.198)]. A positive shap-value means a positive impact on prediction, as all of these have, and the size represent how much this feature contributes to the output. In general, it seems that *percentage_wrong* has the strongest contribution followed closely by total

number of questions and *percentage_correct*. Let's look at the shap values for each cluster separately. The SHAP summary plot (Figures 27-31, one for each cluster) is interpreted as follows. The vertical axis indicates the variable name in order of importance from top to bottom. On the horizontal axis is the SHAP-value. Each dot is a single row in the data. For the cluster 0, the number of exams and the number of attempts have a high and positive effect for explaining this cluster. Also, the correct responses have a higher explanatory value than the incorrect ones. Figure 27 contains the Shapley values for cluster 0.

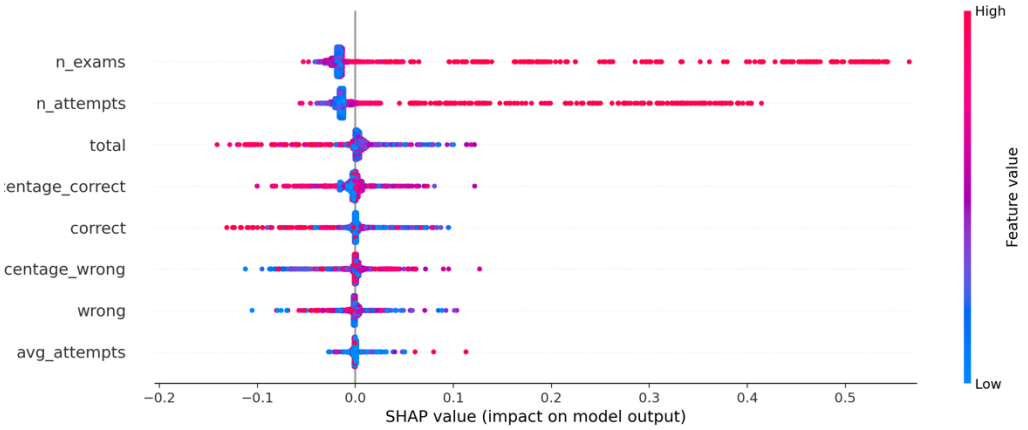


Figure 27. Shapley values for cluster 0

For cluster 1 the strongest explanatory variable is the total number of questions answered with the wrong answers having a high positive impact and the correct answers a strong negative impact. Figure 28 contains the Shapley values for cluster 1.

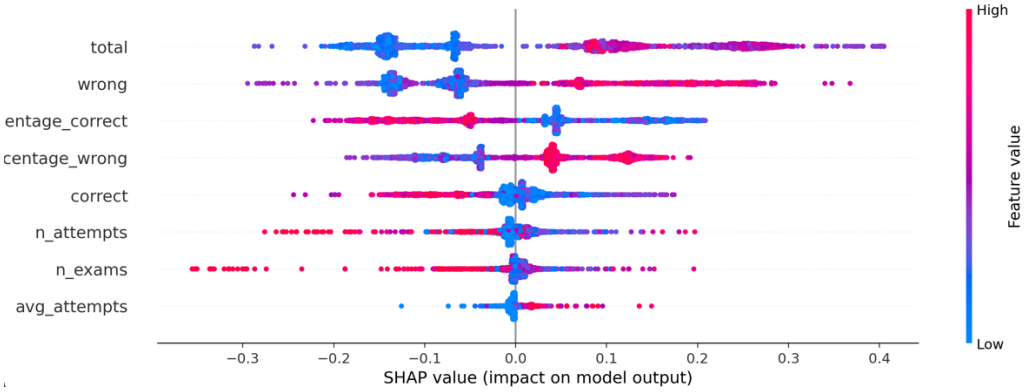


Figure 28. Shapley values for cluster 1

For cluster 2 the total number of answered questions has the highest explanatory impact and it is positive. The number of correct responses strongly explains this cluster as well as the number of attempts. Figure 29 contains the Shapley values for cluster 2.

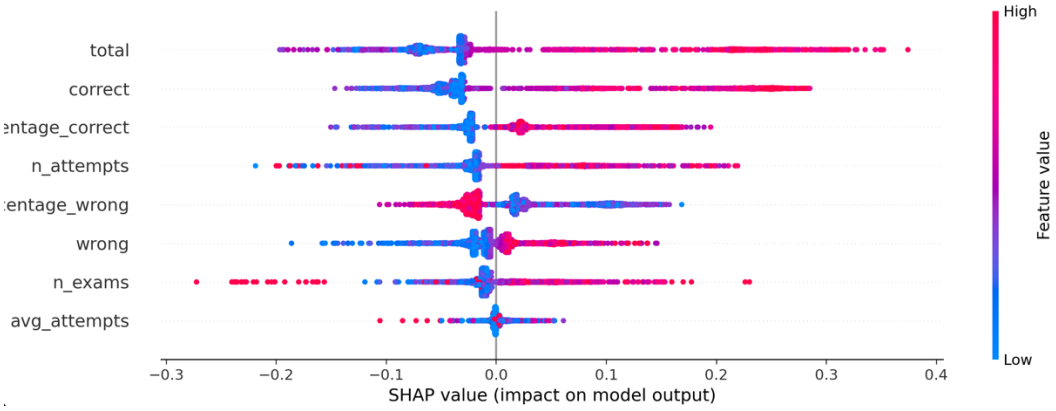


Figure 29. Shapley values for cluster 2

For cluster 3 the total number of answers has a strong negative impact, the percentage of correct responses a strong positive one and the percentage of wrong answers a strong positive impact. Figure 30 contains the Shapley values for cluster 3.

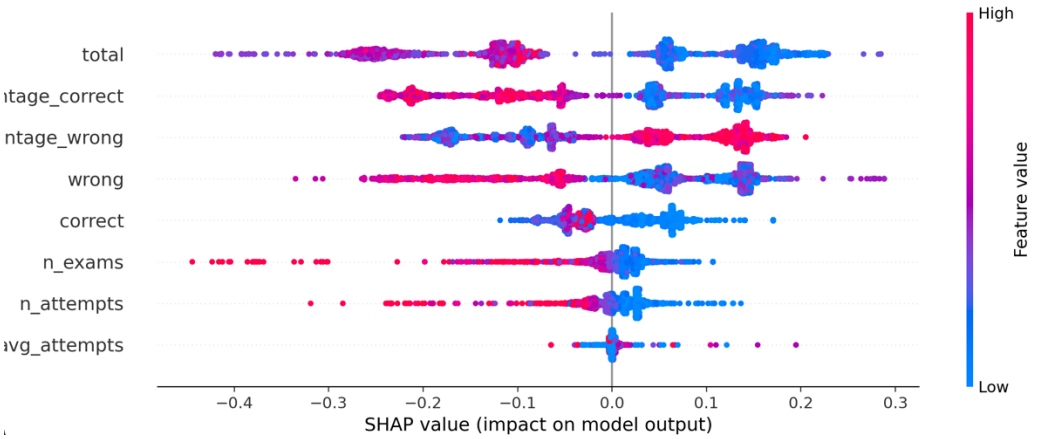


Figure 30. Shapley values for cluster 3

For cluster 4, the highest impact is on the percentage of correct and wrong answers. The percentage of correct responses has a high positive impact, and the percentage of wrong answers has a

high negative impact. The total number of answers has a strong negative impact. Figure 31 contains the Shapley values for cluster 4.

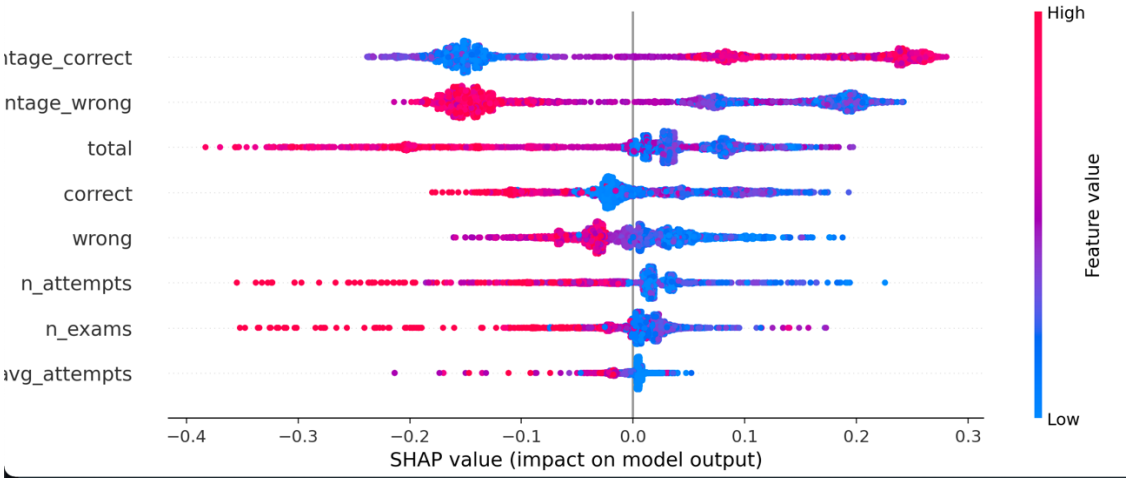


Figure 31. Shapley values for cluster 4

Bringing these bits of information together, for the cluster 0, the number of exams and attempts and total answers, so basically activity on the learning platform are clearest indicators of belonging to this group. They are very engaged, and this engagement is more important than success. They also get more right than wrong. For cluster 1, the total number of questions and especially the ones answered incorrectly explain belonging to this cluster. They have the lowest engagement and many wrong answers. For cluster 2, the number of questions answered has the highest explanatory impact as well as the low number of correct answers. They are engaged but not succeeding very well. For cluster 3, they get more answers right than wrong, but have not answered that many questions and have not participated in that many exams. For cluster 4, the percentage of correct and incorrect answers explains their belonging to this group. They have answered a lot of questions and participated in a lot of exams but are not succeeding all that well.

```
#%% Random forest with shap

#Importing Libraries
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
import sklearn.metrics as metrics
import shap

# create X (independent variables) and y (dependent variable)
X = df_scaled.copy()
y = pred

# divide into train dataset to train algorithm and test to check for performance
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)

#Fitting Random Forest Classifier to the training set
classifier = RandomForestClassifier(n_estimators=10, criterion='entropy', random_state=42)
classifier.fit(X_train, y_train)

# predicting the test set results
y_pred = classifier.predict(X_test)
print(pd.crosstab(y_test, y_pred, rownames=['Actual cluster'], colnames=['Predicted cluster']))
print(metrics.classification_report(y_test, y_pred, digits=5))

print(list(zip(df.columns[2:10], classifier.feature_importances_)))

data_shap = df.copy().drop(['Unnamed: 0', 'user_id', 'cluster'], axis=1)

explainer = shap.Explainer(classifier)
shap_values = np.array(explainer.shap_values(X_train))
shap_values_ = shap_values.transpose((1,0,2))

np.allclose(
    classifier.predict_proba(X_train),
    shap_values_.sum(2) + explainer.expected_value
)

shap.summary_plot(shap_values[4],X_train, feature_names=['n_attempts','n_exams','avg_attempts',
                                          'correct','wrong','percentage_correct',
                                          'percentage_wrong','total'])

dump(classifier, 'randomforestmodel.pkl')
pickle.dump(explainer, open('Kshap_explainer2.pickle', 'wb'))
```

Code block 5. Random Forest Classifier and SHAP code

5.3.5 Business result

Directing help to any of the clusters is a matter of prioritizing business objectives. Of these groups, the students in cluster 1 are struggling the most to get started. Their percentage of wrong answers is a lot higher than that of correct answers. The same is true for cluster 3. For these two groups in both percentages as in amounts they get more wrong than right and could use more help. Clusters 2 and 4 are more engaged. Cluster 2 has students that have a lot more answers wrong than correct. For cluster 4, the difference is only slight. The models, the scaler and SHAP explainer was saved for deployment purposes later. The information of the prediction for each datapoint was also saved. This code can be found in Appendix 3 in Code Block 10.

5.4 Failsafe method without machine learning

5.4.1 Problem formulation

In the production phase AI system, for this type of environment, an edge deployment would be best fitting as then the model could make predictions on the servers close to the users and only periodically reload the updated and retrained model. This way the lack of connection to the internet does not interrupt functionality of the system. This system is not feasible to build in the proof-of-concept phase. Therefore, it is necessary to build another solution that can be utilized on the edge to perform a similar function of shortlisting the students that most need help and would most benefit from it without the help of machine learning.

To build a failsafe system it is necessary to define by hand what is meant by needing help and benefitting from it. To be a safer option, it needs to be very simple. The same proxies could be used as for the K-Means cluster algorithm with the number of attempts and the number of questions answered acting as proxies for perseverance and the percentage of correct answers as a proxy for mastery. As these are hard coded instead of adapting to the system, a classificatory approach is difficult as that leaves those just on the wrong side of the hard coded classification in trouble.

The percentage of correct answers is comparable between students. A similarly comparable proxy for perseverance is also needed. For this seeing the percent rank of the total number of questions answered could act as a starting point. Then the problem formulation is organizing all the students with line data in the database according to their percentage of correct answers and their percent rank of total number of questions. This list can then be organized either based on mastery, or lack thereof, or based on perseverance.

5.4.2 Data extraction and preprocessing

The data is extracted from the database by an SQL-query in Code Block 6.

```
select user_id,
(sum(status) / count(status) )*100 as percentage_correct,
percent_rank() over (ORDER BY count(status) desc)*100 as perseverance
from excs_detail
group by user_id
```

Code block 6. SQL-query to extract the desired data for failsafe system

The data needs no further processing, and the query can be run within the learning platform itself without the need to download any libraries that would take up space in the servers that is needed for content and operating system. This option utilizes every single line of data in the *excs_detail* table so no datapoints are lost. An example of the data is in Figure 32.

user_id	percentage_corr...	perseverance
1098	0.0000	95.6850691600
1148	74.6988	17.3779406500
1268	0.0000	95.6850691600
1391	15.4839	6.5019505900
1413	65.5319	13.0866532700
1414	46.0238	1.6313985100
1415	80.8534	3.3810143000
1418	59.1592	9.6347085900
1419	68.5185	24.0335737100

Figure 32. Example of data for failsafe system of identifying students needing help

5.4.3 Interpretability and explanation of results

Here the percentage of correct answers is calculated by combining all attempts into one and calculating their total number of correct answers and dividing that sum with the total number of answered questions. The higher the percentage is the more mastery the student has shown with less problems with failing to succeed.

For the perseverance indicator, it looks at the total number of answered questions, organizes all students from the one with the most answers to the one with the least and gives everyone a number that relates to their place on the organized list. The higher the value is, the more questions the student has answered compared to other users. This is only an indicator of perseverance and should be reviewed after thorough user testing. Both values can be found for 8460 users whose information can be found in *excs_detail* table of the database.

5.5 Testing

Testing was done for the machine learning models by utilizing performance metrics and test groups. However, further testing is required when a more complete dataset has been gathered. To get more information on the performance of the model and SQL-query, a figure was created with the clusters including the two new variables from the failsafe query. This image also shows the disparity for clusters 1 and 3 between wrong and correct answers and suggest offering them help first. Between them the group 3 shows stronger perseverance. Further testing needs to be done with the students by people who know the students and can give invaluable feedback on their actual needs. Figure 33 shows a line chart depicting the average performance of the 5 clusters in the test data.

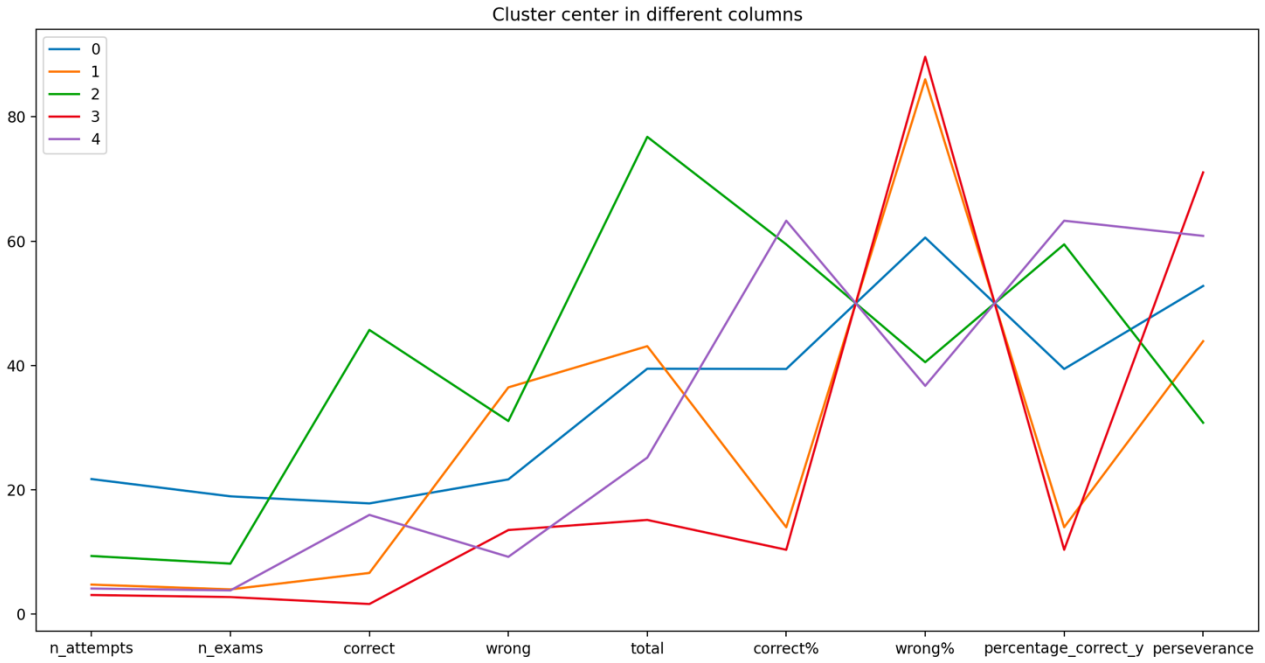


Figure 33. Mean performance of the 5 clusters

5.6 Design choices after modelling and before deployment

This process proceeded by building a K-Means clustering algorithm that could produce a classificatory label, that can then be predicted on new data with a very high degree of the performance

metric, being F1 in this case. The classification model will be deployed as a REST API endpoint to allow for testing. Due to the requirements of this business case, another method to identify students at risk was also developed to act as needed but to be as simple as possible to allow for ease of integration into the learning platform. This failsafe method is possible to produce with a simple SQL-query that is passed on to the developers of the learning platform. The updated design of the system backend is in Figure 34. The unsupervised learning algorithm K-Means is not published as an endpoint but is used in the retraining of the system and the Random Forest classifier.

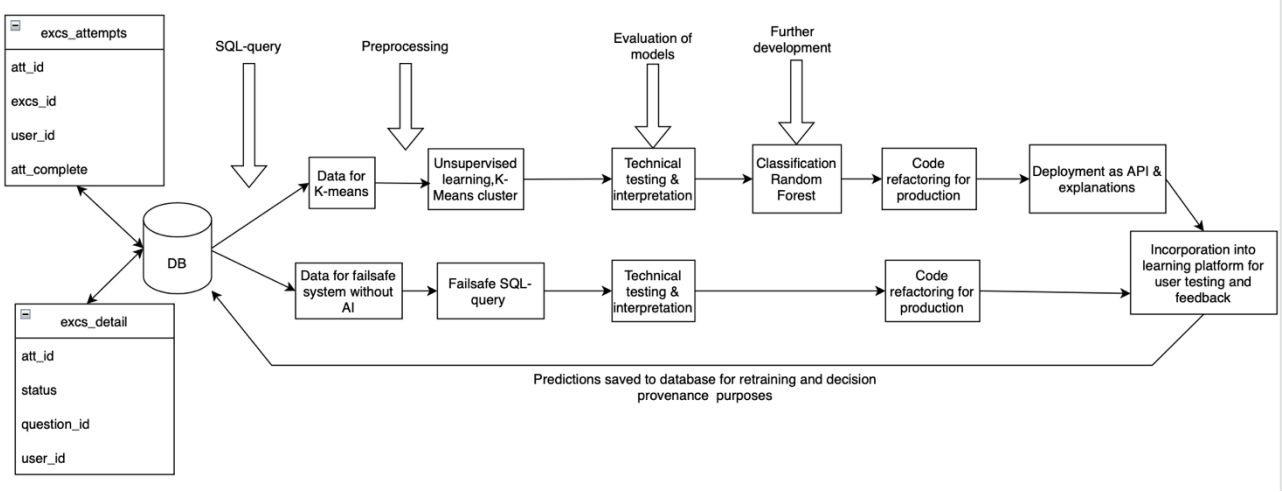


Figure 34. Final design of the AI system for the proof-of-concept phase

These models are robust and can be explained as can be seen in chapter 5.3. Further explanation options will be created in the deployment phase where it is possible to attain an explanation of an individual prediction. This thesis acts as a documentation of each phase and choice made in the development and design of the system as well as in the deployment.

This system will not be autonomous but instead its' role is to augment the human skillset, so the responsibility of the actions of the mentor rests with the mentor. However, they need to be trained and it is also necessary to ascertain that they understand this system and its' limitations. As for risks to the safety of this system, there are always some. The database can be hacked even though there is access control. If it is hacked the data can be leaked, poisoned, or both. The individual servers in India can stop working or updating themselves leading to a student missing out on essential help. However, there is contact between the mentors and the administrators that al-

low for events such as this to surface and be repaired. Should the REST API fail to work as intended, it is very easy to take offline and switch to the failsafe system until it can be repaired and retested.

It would be beneficial if the users could report problems in the learning platform, and, also ask for an explanation should they need more information than is provided to them. This way problems would be noticed early on. Finally, the K-Means cluster and Random Forest algorithm should be retrained at least monthly to allow for data or model drift. Furthermore, it is vital to find the cause why line data goes missing at such a large percentage and fix this to enable putting this AI system to production. As it is, trained on data that is missing nearly half of what is supposed to be there, the results are too unreliable to utilize in production, but sufficient to use as a proof-of-concept and a starting point for further development.

5.7 Deployment

The proof-of-concept model needs to be deployed for it to be tested by Aveti Learning administration and mentors. The deployment opens a way for different people and groups from various geographical locations to test the model simultaneously and gather feedback that helps further development. Many of the cloud providers have simple solutions to deploy ML models as endpoints, but for this purpose a choice is made to utilize open-source libraries, namely Flask.

The REST API that is built here will allow data to be sent to it in json format. The data will be processed so that it can be used to get a prediction from the model which is then sent back as response. This API will also have a rate limiter, that helps to reduce the attack-surface by limiting the number of queries that is possible to send to the API endpoint from a specific IP address. It does not completely prevent, for instance, a DDoS attack, but it does make it harder. It also prevents web scraping and bots and prevents server resource exhaustion and controls the flow of system processes and data. It works by controlling the frequency of the repetition of an operation and ensures that the set constraints guiding this are not exceeded within a specific timeframe. This is done by first defining the rate limit rules for specific operations. Then the system counts every operation and request made by users. If the frequency reaches the rate limit, further requests are not processed until the limitation is lifted or modified. (Esenyi, 2021.)

The REST API was built to allow for ease of use and deployment. The code of the API is in Appendix 5. The modelling code was also refactored for production and can be found in Appendix 6 and requirements.txt in Appendix 7. The requirements-file details the libraries and their versions needed to replicate the project and run the code.

The API was called from localhost for this proof-of-concept with the following example. http://127.0.0.1:2582/predict/?n_attempts=12.0&n_exams=5.0&avg_attempts=2.4&correct=17.0&wrong=24.0&percentage_wrong=0.4146&percentage_correct=0.5854&total=41.0. This produces the following result in Figure 35. The parameters in the address field contain the information of one possible example student and should be replaced with the desired values of the student utilizing the endpoint. The result contains the SHAP-values and the prediction in *prediction_category*.


```
{
  "avg_attempts": "0.01469366",
  "correct": "-0.05203972",
  "n_attempts": "-0.04860914",
  "n_exams": "-0.00970322",
  "percentage_correct": "-0.05318007",
  "percentage_wrong": "-0.06189424",
  "prediction_category": "[2]",
  "total": "-0.12206436",
  "wrong": "0.01679006"
}
```


Figure 35. Example of output from API

The number of connections to the /predict/-endpoint was limited to 5 per minute from a single endpoint for testing. The endpoint can also be contacted from python code as can be seen in Code Block 7. This endpoint can be connected to the learning platform and care should be taken to save the prediction to the database for traceability and auditability.

```
import requests

url = 'http://127.0.0.1:2581/predict/'  # localhost and the defined port + endpoint
body = {
    "n_attempts": 12.0,
    "n_exams": 5.0,
    "avg_attempts": 2.4,
    "correct": 17.0,
    "wrong": 24.0,
    "percentage_wrong": 0.4146,
    "percentage_correct": 0.5854,
    "total": 41.0}

response = requests.post(url, params=body)
print(response)
```

Code block 7. Using the endpoint from python code

5.8 Explanation of results for end users for their own prediction

SHAP-values are extracted from the endpoint with the prediction. Those, however, are not very explainable and certainly not understandable for a child, or a mentor, without training in machine learning or statistics. Explanations need to be understandable and in plain language. An explanation for the student example from the precious subchapter could go something like this: “You’ve started to use the learning app and from your results it seems you’ve gotten off to a good start. However, it seems you get more answers wrong then right. Do you feel you understand the material? If you get questions wrong, please look through the material again and, if that does not help, talk with your mentor. S/he can help.”

If another example student would be classified into cluster 1 and considered to need help, an explanation of the help could be something like this: “We’ve looked at the results you have been getting and we think you could benefit from a little more help. What do you think? Would sometime next week be ok?”

While it might be completely unnecessary to go into more detail with a child, a parent could be another story. Also, the mentor should understand the results better. For this a figure could prove beneficial for explanation. Figure 36 shows the SHAP-values for each variable. This, however, most likely will not help the mentor explain to the parents, why

the student needs help. They can, however, help the administrators of the learning platform in developing more detailed explanations and learning material to mentors and other interested parties.

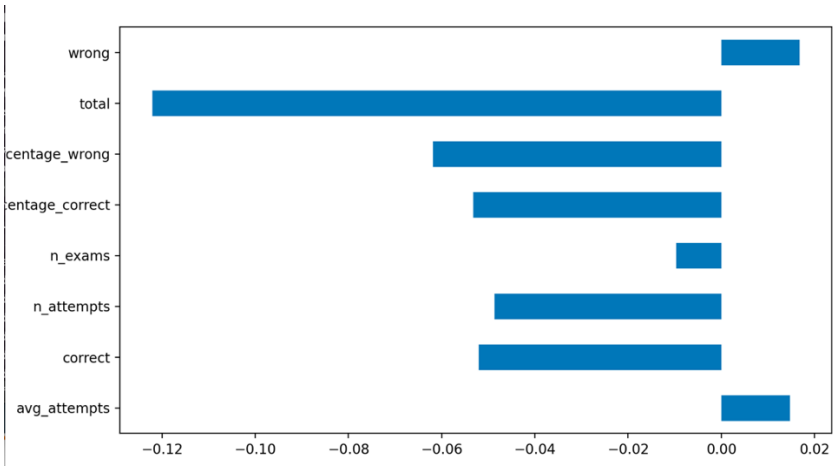


Figure 36. Shap-values for example student

A better approach to use in the preliminary testing would be to look at the information from the database concerning the performance of the student (Figure 37). The mentor could show this to the parents and tell them that the student is doing quite well but might benefit from further help based on the fact that they have more incorrect answers than correct answers. Here informing the student or parent about utilizing AI in the data analysis is not absolutely necessary as it is not an autonomous process. However, for full disclosure, is should be described that the student was selected for further help with statistical and machine learning methods and if there is a disagreement with the student receiving further help, or not receiving it, based on the recommendation, such discussions can be taken up with the mentors.

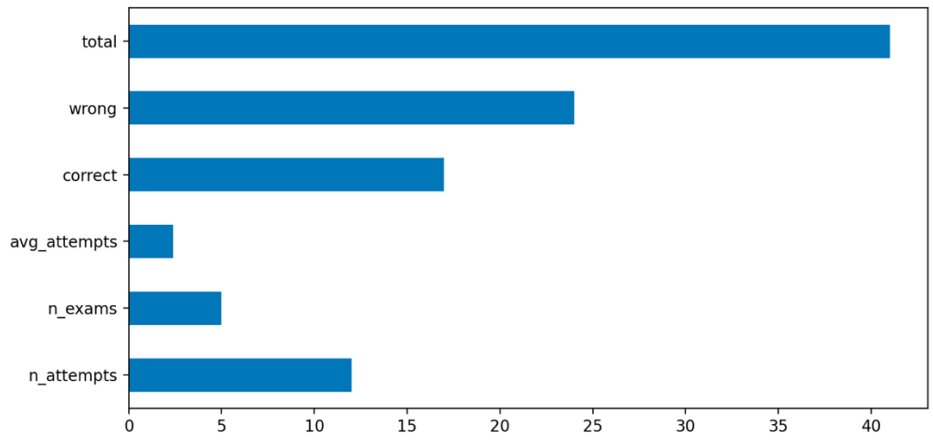


Figure 37. Student's information from the database

5.9 Answering the big questions based on this proof of concept

A set of questions was created based on the literature. These questions were then modified, and a set of questions was created that allowed looking at the bigger picture from the viewpoint of each stage in the development of the AI system. Now, it is time to come back to the original set of questions and to evaluate the development from the viewpoint of the complete project presented in this thesis.

1. Does it comply with all relevant laws and regulations and human rights? Basically, is it legal?
   Yes, it does.

2. Is it ethical?
   a. Interpretability, explainability, and transparency
      i. What interpretability methods are used to understand the working of the model in the usual cases as well as in anomalies?
         Data exploration and shap values

      ii. What methods and delivery channels are used to deliver timely and understandable explanations to stakeholders and what is the content for each stakeholder group?
         An few examples were created to serve as a starting point should the company wish to pursue this development further. It is still required that these be developed more in depth and in various instances should this development be continued.

      iii. How transparent should we be about each aspect of the project, why and to whom?

For this project, full transparency except from the viewpoint of the data, was chosen. The data is not shared due to privacy concerns.

iv. What needs to be done to enable these requirements and how, as well as, by whom is the documentation etc. kept up to date?
A choice needs to be made whether to continue with the development of AI augmentation of the mentors. If it is chosen to continue, then these questions need to be answered.

b. Diversity, non-discrimination, and fairness
   i. How was fairness defined and by whom?
   Fairness was defined by the development team as offering each student the help they need instead of equally offering aid to everyone despite the need.

   ii. What sources of bias were considered and mitigated during the design, development, and deployment of the AI system and how?
   The data did not allow for a lot of bias mitigation. Should this proof of concept be continued, a more in-depth study on possible other sources of bias needs to be conducted.

   iii. What other means to assure a fair system were used and how?
   Ideas for further development were suggested that allow for greater fairness and suitable redress.

   iv. How diverse is the team responsible for the life-cycle of the AI solution?
   The team building the entire solution is diverse, but the person developing the AI aspect is one individual with a very different background from the others responsible for the development. Further collaboration is needed if development is continued.

c. Human agency
   i. What is the role of the AI system in the human-AI interaction, what does this role need to succeed and who is responsible for enabling these requirements?
   The role of the AI is to augment the human making decisions. This role requires good explanations and, also information on the limitations of the system. Decisions on the who remain with the company.

   ii. Are humans safe and in control and is the system built according to the best practices in human agency?
   Yes, but only so far as the humans in charge are safe and fair.

d. Accountability
   i. Who is responsible and of which aspect?
   The developer of the AI is responsible for building a robust and safe system and detailing all limitations to the company so they may make informed decisions. The company is in charge of making those decisions and holding themselves accountable for them. Each mentor is in charge of the decisions they make, but their responsibility is limited by their understanding of the system. Each student is responsible for their own learning and reaching out if they need further help.

   ii. To what degree is the system auditable and documented?
   This system currently is highly auditable and documented.

iii. How can users report problems or seek redress?
Such methods need to be built into the learning platform.

3. Is it robust?
    a. Robustness:
        i. What is the best suited performance metric for the use case and what is a suitable performance level?
        Several of performance metrics were checked for the random forest algorithm. All of which passed the 0,9 level of very acceptable behavior.

        ii. Are modeling and IT best practices used in the development of the system?
        Yes, for this proof of concept.

        iii. Is testing thorough and documented covering also unlikely, but possible, scenarios?
        The testing done is not thorough as it was seen from the quality of the dataset that deployment of the AI model is not advisable due to the percentage of missing data. Should those problems be fixed, a more thorough testing regime is needed.

        iv. Are failure and recovery methods planned and implemented?
        Yes, a failsafe method that can be built into the system was designed.

    b. Safety:
        i. Has the attack surface of the use case been evaluated?
        Yes, it has.

        ii. What security measures have been implemented and how?
        The data training was kept centralized. A rate limiter was developed for the API endpoint to prevent several attack types.

        iii. What is the plan for mitigation in case of an attack?
        Should an attack occur, then the mitigation follows based on the protocols of the learning platform that are outside the scope of this proof-of-concept

    c. Data:
        i. Are data best practices utilized and data minimized?
        Yes, they were. The data was minimized and great care shown in access control and all data processing was documented.

        ii. Is the governance of the data documented and continually maintained?
        Yes, during the proof-of-concept phase.

        iii. How is the quality of the data assured?
        Through testing the dataset. However, there were problems especially in data completeness.

    d. Societal and environmental wellbeing:
        i. Has it beneficence and maleficence to humans, society, all living beings, and to the environment been considered and evaluated, how, by whom and with what results?
        They were assessed by the author of the proof-of-concept. The purpose is very beneficial. Also retraining monthly on a small dataset will not have a significant impact with regards to the use of electricity needed in its performance

ii. How well does it meet the requirements of socially accepted risk, safe and robust model, socially aware level of risk involved and socially beneficial outcome?
Very well.

4. How have these requirements been taken into consideration in all stages of the AI system lifecycle?
This thesis holds a clear documentation as to how these requirements are taken into consideration. All future development needs to be documented.

5.10 Considerations for further development

Building this proof of concept has enabled a deeper look at the data from Aveti Learning database and focus on the problem behind the desire to develop more analytics possibilities. The data is not without problems as quite a large percentage of the line data is missing. Therefore, that is suggested as the main issue needing development. While it may be impossible to find the missing data, finding the reason behind it not arriving to the database can allow for more data collection, so that in the future, the data quality will be improved. Also, gathering the timestamps from the servers when the exam attempt ends, would enable study in how long each exam takes, which would give further indication to problems the student faces. Saving information on the hints used would serve a similar purpose and offer yet another datapoint to use in building a more comprehensive model.

If the students' answers would be saved to the database, it would allow for deeper personalization with machine learning as it would enable the search of similar errors made by other students and recommending the same content that helped them. This would also need a mapping between each question and exam and the learning material provided.

In conclusion, while this proof of concept has offered a view into what could be possible with this data and offered many points for further development, based on the issues with the data, it is recommended to move forward with the solution that does not utilize machine learning. That solution isn't reliable with the amount of missing data. Once the quality of data has improved, it is possible to build upon this proof of concept and develop more advanced analytics to augment the skills of the mentors in deciding to whom to offer aid.

6 Conclusions

The purpose of this thesis was, firstly, to look at the literature on trustworthy AI and develop a set of questions that are usable in various projects utilizing AI. This was done and the questions developed that were then first adopted to serve the building of the proof-of-concept phase and then to overview the entire building part of it. The questions served both purposes well. The need to adopt them to each building phase emphasizes their general nature and, also, the need to adopt them according to the specifications of each project. They also seemed as a good tool to overview that all aspects are considered and as a way to pinpoint and document possible problem areas in each application as well as steps to mitigate each possible problem.

Secondly, the purpose was to implement the ideas and guidelines in the literature in building a beneficial and trustworthy AI proof of concept for a company serving a very important and socially beneficial role in rural India. This was also done as well as directions for future development were given. However, as the data lacked sufficient completeness, it is stated that it is advisable to deploy the failsafe method utilizing basic data analytics without the AI while continuing to develop the system to enhance the quality of the data.

Looking back at the criteria of a good case study in chapter 4.1 and based on Host et al. (2012), they state that the study is to be of a significant topic. This current thesis is just that as it is vital that our AI solutions are trustworthy, and they bring beneficial outcomes into our societies instead of maleficent. Second criterion is that the boundaries are made explicit, evidence is comprehensive and there are no significant constraints on the conduct of the study. These all hold true for this research. Alternative perspectives and solutions were also considered and implemented, as is required of a good case study. A good case study also respects ethical, professional, and legal standards relevant to this study, describes the theoretical basis and offers a fully documented chain of evidence with traceable reasons and arguments. This has been done to the best of abilities of the author. This then, can be seen as a good case study based on these criteria. As to the ethicality, all parties consented and had thorough information at hand. A suitable level of confidentiality was agreed upon and the quality of data was assessed in all stages of the case study. The traceability of interpretations, as was done here, also increases the validity of this research. Further, it is believed that the same findings could be reached by other researchers giving credence to the reliability of this research.

Building trustworthy AI is possible and research is conducted, and methods developed constantly to aid in this process and goal. However, creating it requires more time and resources than creating a solution without going over all these possible problem areas. It is therefore understandable that many AI solutions can be developed without these safeguards. Such an undertaking is shortsighted as the risks involved are substantial for the company taking the shorter route. Even if nothing untoward happen and the risks don't materialize, the coming regulation can signal great difficulties and use of resources and money in recreating the already created solutions to assess their trustworthiness. It is therefore advisable to take the cautious route from the start. As Simpson Rochwerger and Pang (2021) state:

> *No matter how you deploy machine learning, you are deploying bias at scale. By definition, you are encoding bias and decision-making into a very big, fancy engine that is going to make decisions on behalf of a human. When you participate in the creation of this engine, you have a basic moral responsibly to do so responsibly* (Simpson Rochwerger & Pang, 2021)

References

AI HLEG. (2019). *Ethics Guidelines for Trustworthy AI.* European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Ammanath, B. (2020, March 26). *Forbes Insights: Trust At The Center: Building An Ethical AI Framework*. Forbes. https://www.forbes.com/sites/insights-ibmai/2020/03/26/trust-at-the-center-building-an-ethical-ai-framework/

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete Problems in AI Safety*. arXiv:1606.06565v2 [cs.AI]. https://arxiv.org/abs/1606.06565v2

Banerjee, D. N., & Chanda, S. S. (2020). *AI Failures: A Review of Underlying Issues*. arXiv:2008.04073 [cs.CY]. https://arxiv.org/abs/2008.04073

Barclay, I., Preece, A., Taylor, I., & Verma, D. (2019). *Quantifying Transparency of Machine Learning Systems through Analysis of Contributions*. arXiv:1907.03483v1 [cs.LG].

Barua, A. (2020, June 18). NRI Gives Back to His Roots, Educates 2 Lakh Kids While Living in USA! *Better India*. https://www.thebetterindia.com/230406/odisha-nri-california-ngo-classes-online-rural-students-donation-aveti-learning-ana79/

Benedichte Mayer, C. (2001). A Case in Case Study Methodology. *Field Methods*, *13*(4), 329–352.

Bhattacharya, P. (2020). Safeguarding Intelligent Decision-Making for Business: Towards A Model. *19th International Symposium INFOTEH-JAHORINA*. IEEE.

Bigham, T., Gallo, V., Nair, S., Lee, M., Soral, S., Mews, T., Tua, A., & Fouché, M. (2018). *Deloitte-uk-ai-and-risk-management.pdf* (p. 29). Deloitte Centre for Regulatory Strategy. https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/financial-services/deloitte-uk-ai-and-risk-management.pdf

Criado Perez, C. (2019). *Invisible women. Exposing data bias in a world designed for men*. Abrams Press.

de Laat, P. (2017). Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability? *Philosophy & Technology*, *31*(4), 525–541.

Desislavov, R., Martínez-Plumed, F., & Hernández-Orallo, J. (2021). *Compute and Energy Consumption Trends in Deep Learning Inference*. arXiv:2109.05472v1 [cs.LG].

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M., & Weisz, J. (2021). *Expanding Explainability: Towards Social Transparency in AI systems*. arXiv:2101.04719v1 [cs.HC].

Esenyi, S. (2021, March 31). *Implementing Rate Limiting in Flask APIs*. https://www.section.io/engineering-education/implementing-rate-limiting-in-flask/

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

Euijong Whang, S., Huyn Tae, K., Roh, Y., & Heo, G. (2021). *Responsible AI Challenges in End-to-end Machine Learning*. arXiv:2101.05967v1 [cs.LG]. https://arxiv.org/pdf/2101.05967.pdf

Neuvoston direktiivi 2000/43/EY, annettu 29 päivänä kesäkuuta 2000, rodusta tai etnisestä alkuperästä riippumattoman yhdenvertaisen kohtelun periaatteen täytäntöönpanosta, 2000/43/EY (2000). https://eur-lex.europa.eu/legal-content/FI/TXT/?uri=celex%3A32000L0043

European Commission. (2020a). *Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics.* (COM(2020) 64 Final). European Commission.

European Commission. (2020b). *Public consultation on the AI White Paper. Final report.* European Commission. https://www.standict.eu/sites/default/files/2021-02/PublicConsultation-AIWhitePaper_Finalreportpdf.pdf

Fan, W., & Geerts, F. (2012). Foundations of Data Quality Management. *Synthesis Lectures on Data Management*, *4*(5), 1–217. https://doi.org/10.2200/S00439ED1V01Y201207DTM030

Hall, P., Gill, N., & Schmidt, N. (2019). *Proposed Guidelines for the Responsible Use of Explainable Machine Learning*. arXiv:1906.03533v3 [stat.ML].

Hellström, T., Dignum, V., & Bensch, S. (2020). *Bias in Machine Learning—What is it Good for?* arXiv:2004.00686 [cs.AI]. https://arxiv.org/abs/2004.00686v2

Hildebrandt, M. (2020). 10. "Legal by Design" or "Legal Protection by Design"? In *Law for Computer Scientists and other folk*. Oxford University Press. https://lawforcomputerscientists.pubpub.org/pub/doreuiyy/release/7

Host, M., Rainer, A., Runeson, P., Regnell, B., & Regnell, B. (2012). *Case study research in software engineering: Guidelines and examples*. John Wiley & Sons, Incorporated.

Jansen Ferreira, J., & Monteiro, M. (2021). *The human-AI relationship in decision-making: AI explanation to support people on justifying their decisions*. arXiv:2102.05460v2 [cs.HC]. https://arxiv.org/abs/2102.05460

Järvinen, P. (2012). *On Research Methods*. Opinpajan Kirja.

Jobin, A., Ienca, M., & Vayena, E. (2019). Artificial Intelligence: The global landscape of ethics guidelines. *Health Ethics & Policy Lab*. arXiv:1906.11668. https://arxiv.org/abs/1906.11668

Khairul baharein, N. (2008). Case Study: A Strategic Research Methodology. *American Journal of Applied Sciences*, *5*(11), 162–164.

Krishna, D., Albinson, N., & Chu, Y. (2017). *Managing algorithmic risks. Safeguarding the use of complex algorithms and machine learning.* Deloitte Centre for Regulatory Strategy. https://www2.deloitte.com/us/en/pages/risk/articles/algorithmic-machine-learning-risk-management.html

Kumar, M., Roy, R., & Oden, K. (2020). Identifying Bias in Machine Learning Algorithms: Classification without Discrimination. *The RMA Journal*, *103*(1), 42–48.

Leslie, D. (2019). *Understanding artificial intelligence ethics and safety. A guide for the responsible design and implementation of AI systems in the public sector* (Public Policy Programme). The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529

Logrén, C. (2020). *TOWARDS QUALITY ASSURANCE OF MACHINE LEARNING SYSTEMS. Master of Science Thesis.* Tampere University. Faculty of Information Technology and Communication Sciences.

Molnar, C. (2021). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable.* https://christophm.github.io/interpretable-ml-book/index.html

Mueller, S., Veinott, E., Hoffman, R., Klein, G., Alam, L., Mamun, T., & Clancey, W. (2021). *Principles of Explanation in Human-AI Systems*. arXiv:2102.04972 [cs.AI].

Nair, S., Gallo, V., Fouché, M., & Thornhill, B. (2020). *Deloitte-ch-en-audit-building-trustworthy-ai.pdf* (p. 30). Deloitte Centre for Regulatory Strategy. https://www2.deloitte.com/content/dam/Deloitte/ch/Documents/audit/deloitte-ch-en-audit-building-trustworthy-ai.pdf

Nemitz, P. (2018). Constitutional Democracy and Technology in the age of Artificial Intelligence. *Royal Society Philosophical Transactions*. https://doi.org/10.1098

Nikolov, D., Lalmas, M., Flammini, A., & Menczer, F. (2018). *Quantifying Biases in Online Information Exposure*. arXiv:1807.06958v1 [cs.SI].

Yhdenvertaisuuslaki, Pub. L. No. 1347/2014 (2015). https://www.finlex.fi/fi/laki/alkup/2014/20141325

Ojika, D., Strayer, J., & Kaul, G. (2021). Towards Sustainable Energy-Efficient Data Centers in Africa. *OCP Future Technologies Symposium, San Jose, California*. arXiv:2109.04067 [cs.CY].

Oseni, A., Moustafa, N., Janicke, H., Liu, P., Tari, Z., & Vasilakos, A. (2021). *Security and Privacy for Artificial Intelligence: Opportunities and Challenges*. arXiv:2102.04661v1 [cs.CR]. https://arxiv.org/abs/2102.04661

Pasquale, F. (2015). Introduction—The Need to Know. In *The Black Box Society. The Secret Algorithms That Control Money and Information* (pp. 1–18). Harward University Press.

Pery, A., Rafiei, M., Simon, M., & van der AAlst, W. (2021). *Trustworthy Artificial Intelligence and Process Mining: Challenges and Opportunities*. arXiv:2110.02707v1 [cs.SE].

Pipino, L., Lee, Y., & Wang, R. (2002). Data Quality Assessment. *Communications of the ACM*, *45*(4), 211–218.

Rabiul Islam, S., Eberle, W., Khaled Ghafoor, S., & Ahmed, M. (2021). *Explainable Artificial Intelligence Approaches: A Survey*. arXiv:2101.09429v1 [cs.AI]. https://arxiv.org/abs/2101.09429

Rességuier, A., & Rodrigues, R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data & Society*, *July-December*, 1–5. https://doi.org/10.1177/2053951720942541

Rosenfeld, A., & Richardson, A. (2019). Explainability in human-agent systems. *Autonomous Agents and Multi-Agent Systems*, *33*, 673–705.

Saif, I., & Ammanath, B. (2020, March 25). *'Trustworthy AI' is a framework to help manage unique risk*. MIT Technology Review. https://www.technologyreview.com/2020/03/25/950291/trustworthy-ai-is-a-framework-to-help-manage-unique-risk/

Sapni, G. K., & Mihir, M. (2021). Understanding China's Draft Algorithm Regulations. *The Diplomat*. https://thediplomat.com/2021/09/understanding-chinas-draft-algorithm-regulations/

Scantamburlo, T., Cortés, A., & Schacht, M. (2020). *Progressing Towards Responsible AI*. arXiv:2008.07326v1 [cs.CY].

Simons, H. (2009). *Case study research in practice*. Sage Publications.

Simpson Rochwerger, A., & Pang, W. (2021). *Real World AI: A Practical Guide for Responsible Machine Learning*. Lioncrest Publishing.

Singh, J., Cobbe, J., & Norval, C. (2019). Decision Provenance: Harnessing Data Flow for Accountable Systems. *IEEE Access*, *7*, 6562–6574. https://doi.org/10.1109/ACCESS.2018.2887201

Smith, C. (2019, October 8). Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development. *AAAI FSS-19: Artificial Intelligence in Government and Public Sector*. arXiv:1910.03515 [cs.AI].

Snyder Caron, M., & Gupta, A. (2020). The Social Contract for AI. *IJCAI 2019 AI for Social Good Workshop*. arXiv:2006.08140 [cs.CY].

Stevenson, A. (2018). Facebook Admits It Was Used to Incite Violence in Myanmar. *The New York Times*. https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html

Stoica, I., Song, D., Ada Popa, R., Patterson, D., Mahonet, M., Kantz, R., Joseph, A., Jordan, M., Hellerstein, J., Gonzalez, J., Goldberg, K., Ghodsi, A., Culler, D., & Abbeel, P. (2017). *A Berkeley View of Systems Challenges for AI*. arXiv:1712.05855v1 [cs.AI].

Suresh, H., & Guttag, J. (2020). *A Framework for Understanding Unintended Consequences of Machine Learning*. arXiv:1901.10002v3 [cs.LG].

Tang, H. (2021). *Engineering Research. Design, Methods and Publication.* Wiley.

The Committee of experts on internet intermediaries (MSI-NET). (2018). *Algorithms and Human Rights. Study on the human rights dimensions of automated data processing techniques and possible regulatory implications*. the Council of Europe.

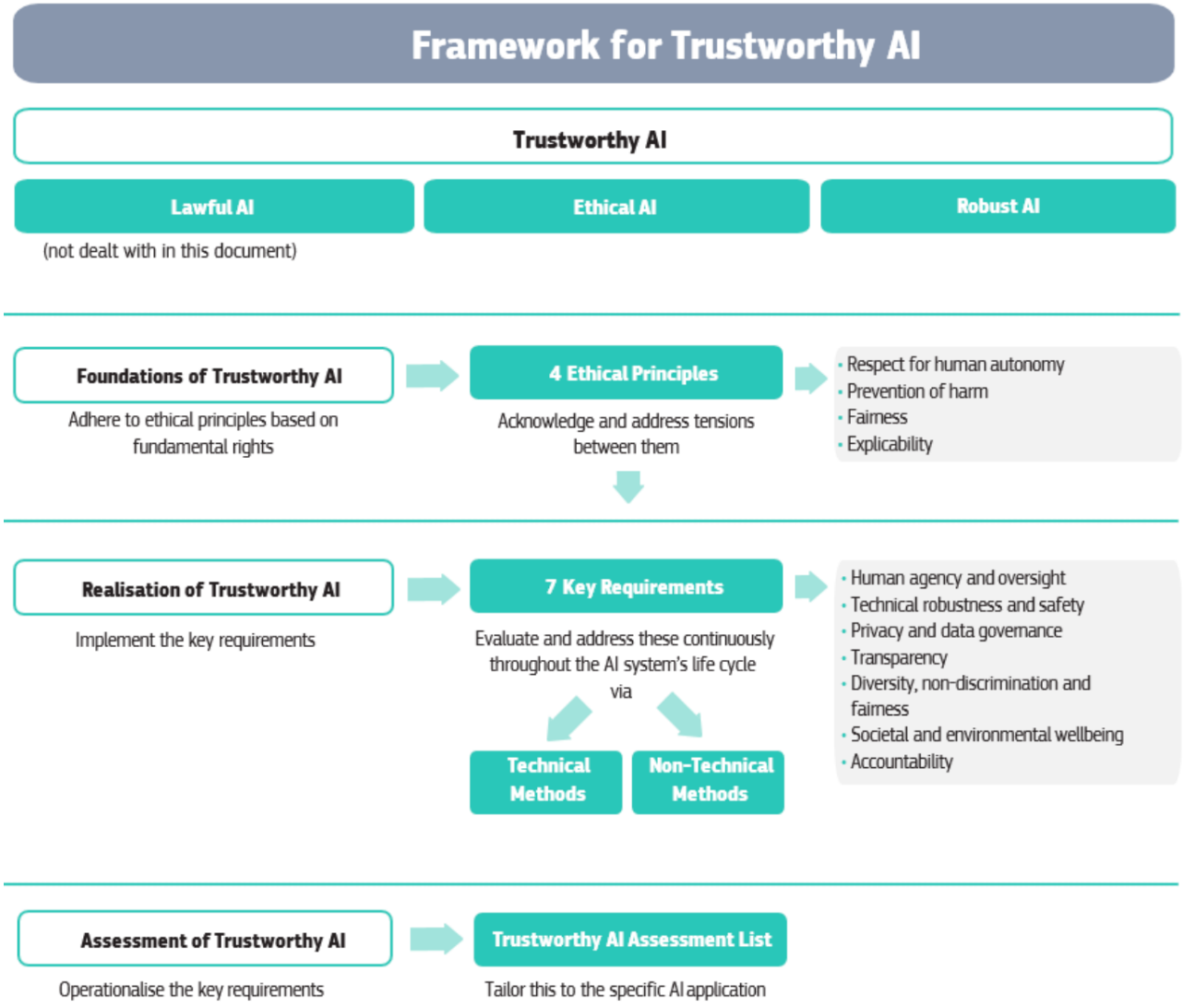
Vähä-Sipilä, A., Marchal, S., & Aksela, M. (2021). *Tekoälyn soveltamisen kyberturvallisuus ja riskienhallinta* (978-952-311-771-6; No. 9/2021). Traficom & Kyberturvallisuuskeskus; ISBN. https://www.traficom.fi/sites/default/files/media/publication/Tekoälyn%20soveltamisen%20kyberturvallisuus%20ja%20riskienhallinta.pdf

Vought, R. (2020). *MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES. Subject: Guidance for Regulation of Artificial Intelligence Applications*. White House. https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf

Zhang, J., Liu, K., Khalid, F., Hanif, M. A., Rehman, S., Tehocharides, T., Artussi, A., Shafique, M., & Garh, S. (2019). Invited: Building Robust Machine Learning Systems: Current Progress, Research Challenges, and Opportunities. *2019 56th ACM/IEEE Design Automation Conference (DAC),* 1–4. IEEE Xplore.

Appendices

Appendix 1. EU Framework for Trustworthy AI



Source: AI HLEG, 2019

Appendix 2. The Human-Machine Teaming Framework

Use the HMT Framework to guide development of accountable, de-risked, respectful, secure, honest and usable AI systems, with a diverse team aligned on shared ethics.

We are confident that we have designed our AI system so that:
We ensured humans are always in control, able to monitor and control risk.

We designated responsibility to humans for all decisions and outcomes.
We explicitly defined responsibility and who shares respon- sibility.

We preserved human responsibility for final decisions that affect a person's life, quality of life, health, or reputation. Significant decisions made by the AI system are: Appealable, able to be overridden, and reversable.

We identified the full range of risks and benefits:

- Harmful, malicious use
- Good, beneficial use
- Blind spots and unintended consequences

We have created plans:

- Communication plan(s) for misuse/abuse of AI system
- Mitigation plans for misuse/abuse of AI system

We value transparency with the goal of engendering trust:

The purpose and limitations of the AI system are explained in plain language.
Data sources and training methods have unambiguous sources and are verifiable.

Confidence and context are presented for humans to base decisions on.
We provided transparent justification for outcomes.
The AI system includes straightforward, interpretable, monitoring systems.

The AI system explicitly states its identity, is honest and usable:

Humans can easily discern when they are interacting with the AI system vs. a human.
Humans can easily discern when and why the AI system is taking action and/or making decisions.

Improvements will be made regularly to meet human needs and technical standards.

A form for your team to use as a checklist and to sign in agreement with the HMT Framework principles, is available online as a PDF: https://drive.google.com/open?id=1aI- oJb2hen-bufT5eZ2MxTrQfdWrplYFc2

Source: Smith, 2019

Appendix 3. Code bits used in development

```
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv('data.csv')

# df['correct'] = df.correct.where(df.correct < 250)
# df['wrong'] = df.correct.where(df.correct < 250)

print(df.describe().T.to_string())

plt.figure(figsize=(14, 5))
plt.plot(df.wrong, label='wrong')
plt.plot(df.correct, label='correct')
plt.ylabel('count of answers')
plt.xlabel('att_id')
plt.legend()
```

Code block 8. Code to create a figure of the right and wrong answer columns in the dataset

```
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv('Kmeans.csv')

#%%

means = df[['n_attempts', 'n_exams', 'correct', 'wrong', 'total']]
means.mean().plot(kind='bar')


#%%

means.plot(subplots=True, layout=(2,3), kind='hist')

#%% eda of clusters

cluster_centers = df.groupby('cluster')['n_attempts','n_exams', \
                            'correct', 'wrong','total'].mean().T

cluster_centers.columns = ['0','1','2','3','4']

cluster_centers.plot(kind='line')
plt.title('Cluster center in different columns')
plt.legend(loc='upper left')
```

Code block 9. Code for the EDA of the KMeans data

```
#%% saving the model for deployment with pickle, and saving the prediction to the data
from joblib import dump, load
import pickle

pickle.dump(kmeans, open('Kmeans-model.pickle', 'wb')) #Saving the model

dump(scaler, 'std_scaler.bin', compress=True) # save the scaler for deployment

data.to_csv('Kmeans.csv')
```

Code block 10. Saving the model and the prediction

```
def display_factorial_planes(X_projected, n_comp, pca, axis_ranks, labels=None, alpha=1, illustrative_var=None):
    '''Display a scatter plot on a factorial plane, one for each factorial plane, from
    https://github.com/OpenClassrooms-Student-Center/Multivariate-Exploratory-Analysis.git functions.py
    '''

    # For each factorial plane
    for d1,d2 in axis_ranks:
        if d2 < n_comp:

            # Initialise the matplotlib figure
            fig = plt.figure(figsize=(7,6))

            # Display the points
            if illustrative_var is None:
                plt.scatter(X_projected[:, d1], X_projected[:, d2], alpha=alpha)
            else:
                illustrative_var = np.array(illustrative_var)
                for value in np.unique(illustrative_var):
                    selected = np.where(illustrative_var == value)
                    plt.scatter(X_projected[selected, d1], X_projected[selected, d2], alpha=alpha, label=value)
                plt.legend()

            # Display the labels on the points
            if labels is not None:
                for i,(x,y) in enumerate(X_projected[:,[d1,d2]]):
                    plt.text(x, y, labels[i],
                             fontsize='14', ha='center',va='center')

            # Define the limits of the chart
            boundary = np.max(np.abs(X_projected[:, [d1,d2]])) * 1.1
            plt.xlim([-boundary,boundary])
            plt.ylim([-boundary,boundary])

            # Display grid lines
            plt.plot([-100, 100], [0, 0], color='grey', ls='--')
            plt.plot([0, 0], [-100, 100], color='grey', ls='--')

            # Label the axes, with the percentage of variance explained
            plt.xlabel('PC{} ({}%)'.format(d1+1, round(100*pca.explained_variance_ratio_[d1],1)))
            plt.ylabel('PC{} ({}%)'.format(d2+1, round(100*pca.explained_variance_ratio_[d2],1)))

            plt.title("Projection of points (on PC{} and PC{})".format(d1+1, d2+1))
            #plt.show(block=False)
```

Code block 11. Function to aid in the visualisation of PCA results

```
from sklearn.decomposition import PCA

# Create a PCA model to reduce our data to 2 dimensions for visualisation
pca = PCA(n_components=2)
pca.fit(df_scaled)

# Transfor the scaled data to the new PCA space
X_reduced = pca.transform(df_scaled)

# Convert to a data frame
X_reduceddf = pd.DataFrame(X_reduced, index=df.index, columns=['PC1','PC2'])
X_reduceddf['cluster'] = pred
X_reduceddf.head()

# to be able to show the clusters, they also need to be transformed with the PCA
centres_reduced = pca.transform(kmeans.cluster_centers_)

display_factorial_planes(X_reduced, 2, pca, [(0,1)], illustrative_var = pred, alpha = 0.8)
plt.scatter(centres_reduced[:, 0], centres_reduced[:, 1],
            marker='x', s=169, linewidths=3,
            color='r', zorder=10)
```

Code block 12. Picturing the clusters in 2D space with PCA

```
""" from https://github.com/OpenClassrooms-Student-Center/Multivariate-Exploratory-Analysis.git functions.py"""
from pandas.plotting import parallel_coordinates
import seaborn as sns

palette = sns.color_palette("bright", 10)

def addAlpha(colour, alpha):
    '''Add an alpha to the RGB colour'''

    return (colour[0],colour[1],colour[2],alpha)

def display_parallel_coordinates(df, num_clusters):
    '''Display a parallel coordinates plot for the clusters in df'''

    # Select data points for individual clusters
    cluster_points = []
    for i in range(num_clusters):
        cluster_points.append(df[df.cluster==i])

    # Create the plot
    fig = plt.figure(figsize=(12, 15))
    title = fig.suptitle("Parallel Coordinates Plot for the Clusters", fontsize=18)
    fig.subplots_adjust(top=0.95, wspace=0)

    # Display one plot for each cluster, with the lines for the main cluster appearing over the lines for the other clusters
    for i in range(num_clusters):
        plt.subplot(num_clusters, 1, i+1)
        for j,c in enumerate(cluster_points):
            if i!= j:
                pc = parallel_coordinates(c, 'cluster', color=[addAlpha(palette[j],0.2)])
        pc = parallel_coordinates(cluster_points[i], 'cluster', color=[addAlpha(palette[i],0.5)])
        # Stagger the axes
        ax=plt.gca()
        for tick in ax.xaxis.get_major_ticks()[1::2]:
            tick.set_pad(20)

def display_parallel_coordinates_centroids(df, num_clusters):
    '''Display a parallel coordinates plot for the centroids in df'''

    # Create the plot
    fig = plt.figure(figsize=(12, 5))
    title = fig.suptitle("Parallel Coordinates plot for the Centroids", fontsize=18)
    fig.subplots_adjust(top=0.9, wspace=0)

    # Draw the chart
    parallel_coordinates(df, 'cluster', color=palette)
    # Stagger the axes
    ax=plt.gca()
    for tick in ax.xaxis.get_major_ticks()[1::2]:
        tick.set_pad(20)
```

Code block 13. Functions for the Parallel Coordinates Plot

```
X_clustered = pd.DataFrame(df_scaled,index=df.index, columns=['n_attempts','n_exams','avg_attempts',
                                        'correct','wrong','percentage_correct',
                                        'percentage_wrong','total'])
X_clustered['cluster'] = pred

# display parallel coordinates plot, one for each cluster
display_parallel_coordinates(X_clustered, 5)
```

Code block 14. Code for the parallel coordinates plot

Appendix 4. Code for app / api endpoint

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Fri Jan 21 14:36:53 2022

@author: satumkorhonen
"""

from flask import Flask
from flask_restful import reqparse
from joblib import load
import numpy as np
from flask_limiter import Limiter
from flask_limiter.util import get_remote_address

app = Flask(__name__)
limiter = Limiter(app, key_func=get_remote_address) # limit traffic to endpoint


@app.route('/')
def welcome():
    return "Hello World and welcome!"

@app.route("/ping/")
def predict():
    return "PONG"

@app.route("/predict/")
@limiter.limit("5/minute")
def post() :
    try: # load retrained versions
        model = load('randomforest-model.pkl')
        scaler = load('std_scaler.pkl')
        explainer = load('explainer.pkl')
    except:     # load proof-of-concept versions
        model = load('randomforestmodel.pkl')
        scaler = load('std_scaler.bin')
        explainer = load('Kshap_explainer2.pickle')
    else:
        print("an error occurred in loading model or scaler")

    parser = reqparse.RequestParser()
    parser.add_argument('n_attempts')
    parser.add_argument('n_exams')
    parser.add_argument('avg_attempts')
    parser.add_argument('correct')
    parser.add_argument('wrong')
    parser.add_argument('percentage_wrong')
    parser.add_argument('percentage_correct')
    parser.add_argument('total')
```

```
    args = parser.parse_args() # creates dict
    # modifies and scales the data to fit the model
    X_new = np.fromiter(args.values(), dtype=float).reshape(1, -1)
    x_scaled = scaler.transform(X_new)
    # get prediction
    rfpred = model.predict(x_scaled)

    shap_values = explainer.shap_values(x_scaled)[1]

    results = {'prediction_category': np.array2string(rfpred),
               'n_attempts': np.array2string(shap_values[0][0]),
               'n_exams': np.array2string(shap_values[0][1]),
               'avg_attempts': np.array2string(shap_values[0][2]),
               'correct': np.array2string(shap_values[0][3]),
               'wrong': np.array2string(shap_values[0][4]),
               'percentage_wrong': np.array2string(shap_values[0][5]),
               'percentage_correct': np.array2string(shap_values[0][6]),
               'total': np.array2string(shap_values[0][7])}

    return results

@app.errorhandler(429)
def ratelimit_handler(e):
    return "You have exceeded your rate-limit"

if __name__ == '__main__':
    app.run(debug=True, port='2582')
```

Appendix 5. Refactored code for retraining the model

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Fri Jan 21 13:15:27 2022

@author: satumkorhonen
"""

import pandas as pd
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
import pickle
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
import sklearn.metrics as metrics
import shap

def kmeans(scaled_data, clusterN) :
    """ KMeans using N clusters and k-means++ initialization,
    Input scaled data and cluster number, output predictions
    saves model"""
    kmeans = KMeans(n_jobs=-1, n_clusters=clusterN, init='k-means++')
    print(scaled_data.shape)
    kmeans.fit(scaled_data) # learning phase for algorithm
    pred = kmeans.predict(scaled_data) # predicting classes for each datapoint

    pickle.dump(kmeans, open('Kmeans-model.pkl', 'wb')) #Saving the model

    return pred

def randomforest(X, y) :
    """Random Forest Classifier with train test split 0,8,
    input X-data and y_data, X_train and classifier
    saves model, prints metrics"""

    # create test and train datasets and a classifier
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
    classifier = RandomForestClassifier(n_estimators=10, criterion='entropy',
                                        random_state=42)

    # fit the data and make the prediction for test set
    classifier.fit(X_train, y_train)
    y_pred = classifier.predict(X_test)
    print(metrics.classification_report(y_test, y_pred, digits=5))

    pickle.dump(classifier, open('randomforest-model.pkl', 'wb'))

    return X_train, classifier
```

```
def shap_info(classifier, X_train) :
    """SHAP explanations for model. Input: classifier and train-data,
    output shap_values and feature_importance,
    saves model, prints feature_importances"""
    explainer = shap.Explainer(classifier)

    shap_values = explainer.shap_values(X_train)[1]
    print(shap_values.mean())
    pickle.dump(explainer, open('explainer.pkl', 'wb'))

    return shap_values

def retrain(clusterN) :
    """retrains models and saves the retrained models. Requires dataname
    and path and the desired number of clusters, calls functions K-means,
    random forest and shap"""

    try :
        data = pd.read_csv('data.csv')
    except:
        data = pd.read_csv('data_user_nonull_over1.csv')

    # scaling the data
    scaler = StandardScaler()
    df_scaled = scaler.fit_transform(data) # scale data
    pickle.dump(scaler, open('std_scaler.pkl', 'wb')) # save scaler

    data['cluster'] = kmeans(df_scaled, clusterN)
    X_train, classifier = randomforest(df_scaled, data['cluster'])
    shap_vals = shap_info(classifier, X_train)
    print(shap_vals)

retrain(3)
```

Appendix 6. Contents of requirements.txt

```
# This file may be used to create an environment using:
# $ conda create --name <env> --file <this file>
# platform: osx-64
aniso8601=9.0.1=pypi_0
blas=1.0=mkl
bzip2=1.0.8=h1de35cc_0
ca-certificates=2021.10.26=hecd8cb5_2
certifi=2021.10.8=py39hecd8cb5_2
click=8.0.3=pyhd3eb1b0_0
cloudpickle=2.0.0=pyhd3eb1b0_0
colorama=0.4.4=pyhd3eb1b0_0
dataclasses=0.8=pyh6d0b6a4_7
flask=2.0.2=pyhd3eb1b0_0
flask-limiter=2.1=pypi_0
flask-restful=0.3.9=pypi_0
intel-openmp=2021.4.0=hecd8cb5_3538
itsdangerous=2.0.1=pyhd3eb1b0_0
jinja2=3.0.2=pyhd3eb1b0_0
joblib=1.1.0=pyhd3eb1b0_0
libcxx=12.0.0=h2f01273_0
libffi=3.3=hb1e8313_2
libgfortran=3.0.1=h93005f0_2
libllvm11=11.1.0=h9b2ccf5_0
limits=2.3.0=pypi_0
llvm-openmp=12.0.0=h0dcd299_1
llvmlite=0.37.0=py39he4411ff_1
markupsafe=2.0.1=py39h9ed2024_0
mkl=2021.4.0=hecd8cb5_637
mkl-service=2.4.0=py39h9ed2024_0
mkl_fft=1.3.1=py39h4ab4a9b_0
mkl_random=1.2.2=py39hb2f4e1b_0
ncurses=6.3=hca72f7f_2
numba=0.54.1=py39hae1ba45_0
numpy=1.20.3=py39h4b4dc7a_0
numpy-base=1.20.3=py39he0bd621_0
openssl=1.1.1m=hca72f7f_0
pandas=1.1.3=py39hb2f4e1b_0
pip=21.2.4=py39hecd8cb5_0
python=3.9.7=h88f2d9e_1
python-dateutil=2.8.2=pyhd3eb1b0_0
pytz=2021.3=pyhd3eb1b0_0
readline=8.1.2=hca72f7f_1
scikit-learn=0.23.2=py39hb2f4e1b_0
scipy=1.7.3=py39h8c7af03_0
setuptools=58.0.4=py39hecd8cb5_0
shap=0.39.0=py39hb2f4e1b_0
six=1.16.0=pyhd3eb1b0_0
slicer=0.0.7=pyhd3eb1b0_0
sqlite=3.37.0=h707629a_0
tbb=2021.5.0=haf03e11_0
threadpoolctl=2.2.0=pyh0d69192_0
tk=8.6.11=h7bc2e8c_0
tqdm=4.62.3=pyhd3eb1b0_1
tzdata=2021e=hda174b7_0
werkzeug=2.0.2=pyhd3eb1b0_0
wheel=0.37.1=pyhd3eb1b0_0
xz=5.2.5=h1de35cc_0
zlib=1.2.11=h4dc903c_4
```