Master of Science in Software Engineer
January 2023

From Software Engineer to being a Machine Learning Engineer - A Study on Skills and Responsibilities

Pannala Sai Pranay Reddy
Sai Manvith Markonda

Faculty of Computing, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden

This thesis is submitted to the Faculty of Computing at Blekinge Institute of Technology in partial fulfilment of the requirements for the degree of Master of Science in Software Engineer. The thesis is equivalent to 20 weeks of full-time studies.

The authors declare that they are the sole authors of this thesis and that they have not used any sources other than those listed in the bibliography and identified as references. They further declare that they have not submitted this thesis at any other institution to obtain a degree.

Contact Information:
Author(s):
Pannala Sai Pranay Reddy
E-mail: sapn21@student.bth.se

Sai Manvith Markonda
E-mail: samk20@student.bth.se

University advisor:
Kai Petersen
Department of Department of Software Engineer

Faculty of Computing	Internet	:	www.bth.se
Blekinge Institute of Technology	Phone	:	+46 455 38 50 00
SE–371 79 Karlskrona, Sweden	Fax	:	+46 455 38 50 57

Abstract

Background: With significant advancements in Machine Learning, the most advancing technology in multiple domains regardless of the field. It has primarily been the critical component of artificial intelligence and is used in a wide range of Artificial Intelligence applications. In the industry, there is growing importance on machine learning, as organisations are actively looking for Machine Learning Experts who develop and create self-running AI systems that will automate processes; With the explosion of data in recent years and the increase in computing power, machine learning systems have become much more capable. Thus the demand for machine learning skills is high in the job market due to the need for more advanced technology.

Objectives: Our thesis aims to provide a comprehensive understanding of primary machine learning skills that are required by organizations to engineer a machine learning system. The main objective of this research is to identify the important skills required to engineer a ML system; the important responsibilities looked for in a Machine Learning engineer.

Methods: There are two parts of the methodology involved in this research. The first part of the research employs the method of Archival Research. In this method, we extracted the 3.4 million job ads dataset. We have got 6497 job ads related to machine learning. The Spacy Natural Language processing automated processing tool will be used to process the job advertisements and identify the required skills and responsibilities. After the skills and responsibilities are obtained, a qualitative survey is conducted on industrial expertise in machine learning to identify the most important skills that are needed and the most looked-for responsibilities for a ML engineer, which is the second part.

Conclusions: The findings of this research have the potential to open up numerous opportunities for individuals who aspire to learn about machine learning systems. This includes software engineers and anyone else who wishes to gain a comprehensive understanding of the skills and responsibilities required to work on machine learning systems. The insights gained from this study may serve as a roadmap to transition into the machine-learning field and can also contribute to advancements in the field.

Keywords: Machine Learning, Skills, Responsibilities, ML Design and Engineer

Acknowledgments

We sincerely express our gratitude to our supervisor, Kai Petersen for his expert advice and immense support throughout the thesis. We are really grateful to him for being patient and providing all the inputs required to complete the thesis. We are deeply indebted to his constant encouragement, motivation, and guidance for providing this opportunity to pursue our master's in software Engineer.

We would also like to express our special thanks to Micheal Unterkalmsteiner for his constant communication with us at all times.

We finally like thank our family and friends who boosted us morally, corrected us educationally, and stood with us through thick and thin. Thank You :)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The goal of ML is to duplicate or emulate human behaviour [1]. The objective is to increase one's knowledge base and organize it so that it may be used to improve performance steadily. It serves as the cornerstone of Artificial Intelligence(AI) and is a key component of the mechanism that endows computers with intelligence. It is used in a wide range of AI applications and mostly depends on induction and synthesis approaches rather than deduction. The ML research develops each type of study theory and method, researches the general automated system and conducts the theoretical study, and sets up a study system that has the specific application facing the duty. It does this by aligning the understanding model or computation model with the study mechanism of civilization through physiology and cognitive science. [1]

ML engineer is an Information Technology(IT) specialist with expertise in the study, creation, and design of self-running AI systems that automate predictive models. ML developers developed and created the AI algorithms that makeup ML; these algorithms are capable of learning and making predictions. As a member of a larger DS team, a ML Engineer frequently works with data scientists, administrators, data analysts, data engineers (DE), and data architects. They may also communicate with groups outside of their teams, such as the IT, application development, sales, or website development teams, depending on the size of the company. [2]

ML is at the crossroads of Data Science(DS) and Software Engineer(SE). DS who concentrate on statistical and model-building work and the development of ML and AI systems are connected by ML engineers that comprise complex ML components that use that data and automate prediction models. Large amounts of data need to be evaluated, analyzed, and organized. Tests need to be run, and ML models and algorithms need to be optimized that require particular skills for execution. [2]

When ML systems become more prevalent or crucial in particular industries, three skills are necessary. In the beginning, a fundamental understanding of how data and these systems are used will become an essential ability for individuals of all ages and experiences as most people's daily interactions with ML become the norm. This could be achieved by teaching young children key ML principles. In order to guarantee that a variety of businesses and professions have the absorptive capacity to apply ML in ways that are beneficial to them, new techniques to create a pool of qualified users or professionals are necessary. Third, greater support is needed to acquire advanced ML abilities. The ML Engineer process Zimmerman describes shows that a ML engineer has a number of duties and obligations to carry out. [3]

It is becoming increasingly vital for software engineers to possess machine learning skills due to their growing importance in the software development industry and the numerous benefits it offers to both careers and projects [4]. Machine learning can enhance the performance and functionality of software systems, as well as provide new opportunities for creating applications and services [5]. The demand for machine learning skills is high in the job market due to the growth of data and the need for more advanced technology, making it a valuable asset for software engineers. Learning machine learning can bring many benefits to software engineers, including the ability to develop software systems that can continuously improve, create new applications and services, stay current with industry trends and advancements, increase earning potential and job opportunities, and solve real-world problems in areas such as healthcare, energy, and security. [6]

A software engineer with a background in programming and development may be interested in exploring the transition to a role as a machine learning engineer, where they can apply their skills in a new and expanding field. The study aims to highlight the differences in skills and responsibilities between the two roles, providing insights for software engineers who are considering making the transition.

Now that the basic aspects of ML and benefits of learning Machine learning for software engineers have been outlined, ML is on the rise can be as evident from the discussion in the following sections, and the current research has its primary focus on determining the skills and responsibilities of a ML engineer. The following section delves deeper into expanding the various aspects of this research by discussing the background, research questions, objectives, and so on.

1.1 Background

The phrase "artificial intelligence (AI)" was first used in the 1950s to describe the straightforward hypothesis that machines may demonstrate human intellect. [7] AI would usher in the "post-physician era" in the 21st century, predicted Jerrold S. Maxmen in 1976. [8] Today's era of quick technical development and exponential growth in extraordinarily huge data sets (also known as "big data") has allowed AI to proceed from theoretical study to practical application on a previously unheard-of scale [9].

As a subset of artificial intelligence (AI), machine learning (ML) exhibits the experiential learning associated with human intelligence while also having the capacity to learn from and improve its analysis by utilizing computer methods. By employing enormous amounts of data as inputs and outputs to recognize patterns and, in essence, learn these algorithms teach the machine to make independent suggestions or conclusions. After sufficient rounds and algorithm tweaks, the computer can receive input and anticipate an output. [7, 10]

In the Fourth Industrial Revolution era, the digital world has access to a wide range of data, such as Internet of Things (IoT) data, cyber data, mobile data, business information, social media data, health data, etc. For the purpose of intelligently analyzing these data and developing related smart and automated applications, knowledge of AI, and more particularly, ML, is required. The area of ML employs a wide range of algorithms, including supervised, unsupervised, semi-supervised, and reinforcement learning. Deep learning (DL), a member of a larger family of ML approaches, can also efficiently analyze a large amount of data. A full overview of these ML techniques that could be applied to increase the functionality and intelligence of an application is provided in the study by Sarker and colleagues [11]. The main contribution of this work is thus to identify the fundamentals of various ML approaches and how they may be applied in a number of real-world application domains, such as cybersecurity systems, intelligent buildings, health, e-commerce, agribusiness, and many more. [11]

ML has made significant advances in recent years, expanding its capabilities across a wide range of applications. ML systems can now be trained on a vast pool of instances thanks to increased data availability, and their analytical capabilities have been bolstered by increased computer processing power. There have also been significant advancements in the use of the algorithms used within the sector, like activities involving massive sets of very high-dimensional input data, such as classification, regression, clustering, or dimensionality reduction giving ML more strength. [12] Due to these developments, machines that performed significantly worse than people only just a few years ago are now capable of outperforming humans at specific tasks.

With a rising number of companies researching and utilizing Data Science (DS) methods such as AI and ML, the demand for competent experts in these fields is growing. a ML Engineer (Engineer) is one of the most sought-after DS positions in the market. [13] As technology focus changes SE require to expand their skill in order to rise up to the level of Engineer a ML system.

ML approaches are used in my areas, including Natural Language Processing (NLP), Image classification, and so on. Massive data's ubiquity has made these algorithmic solutions viable and accurate across a wide range of areas. Supervised learning algorithms used to classify images are built using millions of labelled images making the Engineer of a ML system a great challenge. As ML applications mature, they are becoming an essential component of industries for making faster and more accurate judgments for crucial and high-value transactions.

1.2 Scope

The scope of this research extends to finding the skills required for becoming a ML Engineer, and the focus will be on research pertaining to the prime skills and responsibilities that a Software Engineer aspiring to be a ML Engineer will have to take up.

The following aim and objectives help to work towards answering the research questions.

Aim: This research aims to provide software engineers a clear understanding of the skills and responsibilities required to become a Machine Learning Engineer, and assist them to excel in the field.

Objectives: The various objectives framed towards achieving this aim are discussed below:

- To determine the skills required by software engineers to become machine learning engineers and work with a Machine Learning system.
- To determine the responsibilities that software engineers aspire to become machine learning engineers and are required to work with a Machine Learning system.
- To determine the most important skills and responsibilities that are required to become Machine Learning Engineers.

1.3 Methodology

Research methodology refers to the methods or tactics used to locate, pick, process, and analyze data regarding a topic. This section discusses the details regarding the method to be employed in this research, more of which can be seen in the upcoming chapters. There are two parts of the methodology involved in this research. The first part of the research employs the method of Archival Research based on different job advertisements that have been published for the position of a ML Engineer.

Once the skills needed to become a ML Engineer are obtained, the survey will be conducted to identify the most important skills that are needed and the most looked-for responsibilities for a ML Engineer, which is the second part.

1.4 Chapter Outline

In addition to the current chapter, which is the Introduction chapter where the background for the research and the various objectives, research questions, and basic method to be employed are all discussed, the chapters to follow are mentioned below:
CHAPTER 2: Related works – This chapter gives the different works that are related to the research that is the skills and responsibilities for a ML engineer.
CHAPTER 3: Methodology – This chapter will provide the detailed process involved in the research and the various steps to be followed in order to achieve the objectives.
CHAPTER 4: Result and Analysis – This chapter gives the various results obtained from the method followed, the data obtained and the analysis of the results
CHAPTER 5: Discussion – This chapter gives the interpretation of the result obtained.
CHAPTER 6: Conclusion and Future Work – This chapter concludes the research and also shows if the objectives are achieved and what the future lies ahead pertaining to the research done currently.

Chapter 2

Related Work

This chapter presents the previous works done by several researchers in the field of ML and the skills and responsibilities that were there, related to the literature, have been arranged based on different aspects of ML.

2.1 Literature Review

2.1.1 Literature related to ML, Modeling, and Training

According to Murphy (2013) [14], a model structure and training data sets are used to develop a machine learning (ML) model, and test datasets are used to evaluate learned models. This is a data-driven inductive approach as opposed to the deductive creation of conventional systems. In this study, a model structure is a ML model with unknowable parameters. Employing ML models in a deductively created system necessitates breaking down the data-driven training of model parameters into requirements, designs, and verifications, especially for models employed in safety-sensitive systems.

2.1.2 Literature related to Skill Development using ML and AI

The studies determine how software Engineer and machine learning collaboration benefit from each other in the industry. The Du Zhang [15], explore the intersection of machine learning and software Engineer and highlight how the two fields can benefit from each other. The authors provide an overview of machine learning and its various techniques and explain how it can be applied to software Engineer tasks. The authors explain in detail how software Engineer can benefit from machine learning in several tasks, such as software testing by generating automatic test cases and detecting defects in the software prototypes, project cost prediction, and validation to verify whether the software system is developed as the user intended, reverse Engineer, requirement Engineer etc., The article concludes that there is scope for improvement to develop ML applications, as well application of ML models into software Engineer would improve and transform the way software is developed and maintained, which leads to building effective and efficient software systems. Correspondingly, Ben and Andrew [16], conducted a study on the accountability of machine learning datasets;

they defined a framework which would facilitate transparency of datasets development to normalize modern AI outcomes. The proposed framework leverages the best software Engineer practices, facilitating machine learning datasets to fair and unbiased data. The authors conclude the importance of implementing these critical practices from software Engineer to ensure transparency in machine learning datasets.

Software engineers adapt to AI capabilities by integrating Machine Learning workflow into the organisation's development workflow. Saleema and Christian [17], have performed a case study to analyse the Microsoft team's experiences in building AI applications. The authors have conducted a survey; they designed a questionnaire related to the challenges, best practices and current work design. The stakeholders surveyed were related to different positions in the Mircosoft Organization and had prior experience in Artificial Intelligence(AI) and Machine Learning(ML). The authors discuss their main findings after analysing the survey that describes software engineers, at early stages of adaptation to ML integration, face challenges related to data discovery and management, Customization and reusability of ML models in different domains and ML modularity. The authors propose a set of best practices for integrating ML into existing software Engineer processes to address these challenges at various organisations. The authors highlight the collaboration between data scientists and software engineers in developing Machine Learning.

Artificial intelligence and Machine Learning are becoming increasingly important in software Engineer, particularly in Tech Startups. In the study by Yehia and Jon [18], the authors discuss the impact of machine learning on software Engineer practises and describes how machine learning techniques are transitioning software development and maintenance. The authors also describe the challenges of the need for more expertise of software engineers with machine learning traits and how it impacts the software's data quality, availability, and performance. The authors conclude that software engineers should know the challenges and techniques to develop the software system effectively. In a recent study by Olimar and Valentina [19], the authors conducted a survey to gain insights into how software Engineer practitioners utilise AI/ML techniques in their software development tasks and processes. The authors have investigated the software practitioners(respondents) who have prior experience in improving the software development process in software startups. The authors designed the survey to identify the challenges that software engineers faced in the software development process before the application of AI/ML, as well as how software engineers are currently tackling these challenges using AI/ML techniques. The survey results showed that software practitioners commonly use AI/ML tools and techniques to automate software development processes, including Neural Networks, Supervised Learning Algorithms (such as SVM, DT, NB), and Unsupervised Learning Algorithms (such as clustering and K-means).

As software development processes evolve, learning an essential skill for software engineers is vital in the rapidly changing field of software Engineer. Harman [20] examines that software engineers are required to be equipped with new skills, which is growing in the industry; the study determines the growing importance of machine learning and how software engineers need to adapt to the change. Learning machine learning by software engineers also meets the demand in the job market and can be more competitive. The authors determine the benefits and challenges of implementing ML into software Engineer tasks such as testing, refactoring, maintenance, optimization, etc., and predict that ML will be an essential component of the software Engineer field. The author concludes by suggesting that organizations develop and train their software engineers by equipping them with resources and training them with new skills to be competitive in the market. Correspondingly, Will balard [21] also examines many advancements to the latest applications, such as image recognition systems or secure car systems etc., highlighting that these advancements are due to the understanding the large datasets and potentially developing more advanced applications through machine learning. The author suggests that learning machine learning would improve problem-solving ability, which is crucial for solving complex problems in large datasets. Finally, Zimmermann [22] has performed an empirical study on software engineers to identify the thoughts on ML in an organization. The survey was conducted on 643 software engineers to recognize the importance of ML. The results show that 84% of the participants expressed that ML skills are essential for market competitiveness. The studies define that ML has a significant skill gap in software Engineer and suggests that software engineers should invest themselves in developing machine learning skills.

2.1.3 Literature related to Job advertisements and talent acquisition

Romeo acknowledges the effect of Industry 4.0 on the labour market (2020). [23] The author stated that the hiring environment had altered significantly as a result of rapid advances in AI. Due to the lack of trained AI talent, businesses have been compelled to look for creative solutions, such as upskilling. In this procedure, the businesses retrain their current staff members to get knowledge in AI. Text mining was also employed in Fareri et al. (2020) [24] to look at how Industry 4.0 will affect job profiles and skills. They looked into job descriptions from the global company Whirlpool, which had positions and skills classified using a lexicon of technology and procedures. The authors provided metrics to assess workers' Industry 4.0 readiness.

Pejic Bach et al. (2020) [25] created a job profile highlighting the requisite skills using text mining to analyze Industry 4.0 job descriptions. The initial generation of these job descriptions focused primarily on the Internet of Things and intelligent product design and control. Application domains like enterprise software, supply chains, and customer support were included in the second category. In addition to Industry 4.0, other closely connected areas of interest include business analytics, data analytics, and big data. Numerous studies have utilised job advertising to investigate the skill requirements for various occupations. For instance, Debortoli et al. (2014) [26] examined the variations in job posts from a big data and business intelligence online employment platform. This technique produced a competency framework based on a latent semantic analysis of the text descriptions. They discovered that business skills were just as important for both job positions as technical ones. Gardiner et al [27] research which was built on a system of skill areas revealed that, in addition to conventional hard skills, analytical abilities and soft skills are in high demand in their analysis of 1,216 job listings with big data in the headline.

Between June and September 2005, Lovaglio et al (2018) [28] web's scraping of information and communication technology (ICT) and statistical Italian employment adverts. Their essay sought to outline the qualifications for these positions. Additionally, they sought to distinguish between ICT positions and statistical positions. The results show that computational and analytical skills are more important for statistical experts than soft skills. Verma et al. (2019) [29] provided a skill breakdown for comparable jobs like Analyst, business intelligence analyst, data analyst, and DS using content analysis of job adverts. The authors came to the conclusion that these vocations require decision-making, organization, communications, and database management skills.

The competencies needed for three career fields—DS and Engineer, SE and development, and in the development of business and sales—have recently been studied by Anton et al. (2020). [30] Their mixed-methods technique combines text mining of employment adverts with content analysis of scientific publications. The crucial administrative and technical skills necessary for these positions were given.

According to recent research by Kortum et. al. (2022) [31], Information system research often discusses the topic of human capital. Companies need skilled employees in order to develop and utilize IT artefacts. This is particularly true when sophisticated technologies like artificial intelligence are employed. NLP and computer vision (CV) are two of the main subfields of AI. This essay makes employment market-based analyses and comparisons of the abilities needed for CV and NLP professionals. To accomplish this, we analyze job postings for artificial intelligence using a text mining-based analysis pipeline. Job postings for both sub-disciplines were specifically scraped from a sizable international online job board and examined using named entity recognition and word vectors. It might be demonstrated that the two job descriptions demand different skill sets. A professional in AI does not have a standard need profile; thus, it is important to consider diverse factors.

In another research, Mrisc et al. (2020) [32] research has been done to provide a method for extracting data from job listings on job marketplaces. It mostly deals with mining work criteria from specific job adverts. Our suggested system comprises a data mining module, a ML module, and a post-processing module. The SDCA logistic regression is the foundation of the ML module. The post-processing module uses a number of strategies to improve the job requirement identification process' success rate. The 20 most popular IT employment positions from the chosen job portal were used to verify the suggested system. A total of 9971 job postings were examined. According to our system's verification, 80% of the evaluated advertisements contain all of the job requirements. Additionally, the discovered job specifications were compared to the Open Skills database. We compiled a list of the most prevalent job skills in a small group of IT job roles using this database and the expansion of IT work positions with additional common job skills. The creation of a standardized system to identify job criteria in job postings is the key contribution. The suggested method can be applied to a variety of jobs in addition to IT employment. The data mining module that is being described can also be used for different employment portals.

The three studies by Cruzes et al. (2019) [33], López-Mena et al. (2020) [34], and Soares et al. (2021) [35] all focus on the competencies required by software engineers to work with machine learning (ML) technologies. Cruzes et al. (2019) discuss the challenges and opportunities for software Engineer in ML and identify six key competencies: domain knowledge, data management, modelling, evaluation, deployment, and monitoring. The authors argue that software engineers must be trained in these competencies to develop ML-based software effectively. López-Mena et al. (2020) focus on the competencies required by software engineers in the context of Industry 4.0, which involves the integration of advanced technologies, including ML. The authors identify six key competencies: software Engineer, ML, data management, human-computer interaction, cybersecurity, and ethics. They argue that software engineers must have a multidisciplinary skill set to develop software in this context. Soares et al. (2021) conducted a systematic mapping study to identify the competencies required by software engineers to work with ML technologies. The study identified 14 competencies in three categories: technical, professional, and personal. Technical competencies included knowledge of ML algorithms, programming languages, and tools used for ML development. Professional competencies included project management skills, software development practices, and team collaboration abilities. Personal competencies included communication, creativity, and continuous learning. All three papers highlight the importance of a multidisciplinary skill set for software engineers to work effectively with ML technologies. The competencies required include technical skills in ML algorithms and programming languages and professional skills in project management, software development practices, and team collaboration. Additionally, personal competencies such as communication, creativity, and continuous learning are also essential. These findings provide guidance for the development of training programs and curricula for software engineers working with ML technologies.

2.2 Research Gap

The following limitations observed in the literature: "Why Every Software Engineer Should Learn Machine Learning" by Will Ballard [21] is an article that argues for the importance of software engineers learning machine learning. The article discusses about advantages and disadvantages a software practitioner has if they would like to transition into a machine engineer. But, the article does not provide any guidance or resources for software engineers who want to learn machine learning, such as recommended courses or tools. We used this as a potential research aspect and conducted our research in the real world. We intended to identify a set of skills and responsibilities that would help software engineers adapt to machine learning and its applications in the respective field.

"Software Engineer Competencies for the Future of Machine Learning" by Daniela Cruzes [33]. The authors acknowledge that their study is based on their opinions and experiences and that further research is needed to validate their findings. We used this study as the foundation for our research. We focused on potentially improving its validity by conducting a survey to obtain the important set of skills required for ML engineers.

Chapter 3

Methodology

3.1 Research questions and Motivation

RQ1: What are the skills and competencies that software engineers require to work with a Machine Learning system?
Justification: There is a barrier created between Machine Learning(ML) engineers and Software Engineer(SE) engineers. This division may be due to differences in focus. The ML community is concerned with algorithms and their performance, whereas the SE community is concerned with implementing and deploying software-intensive systems. However, in today's fast-paced software development, SE engineers alone may not be able to automate the software development process, deliver and deploy software quickly, and improve software quality. By leveraging the techniques and tools of AI/ML, SE engineers can facilitate software development processes and enhance the quality of their software products. However, software engineers must be equipped with the necessary skills to apply them effectively. By bridging the gap between these two fields, SE engineers to apply ML techniques to improve software quality. Thus, archival research would help identify the skills needed to be a ML engineer.
Research Method: Research Method used to answer the research question is Archival Research(AR)

RQ2:What are the responsibilities of a Machine Learning Engineer that are required by companies for working with a Machine Learning system?
Justification: Using archival research, we can examine the job postings defined by companies and we can gain insights of the current job market and understand the responsibilities that are in demand for machine learning engineers.
Research Method: Research Method used to answer the research question is Archival Research(AR)

RQ3: What are the most important skills and responsibilities for a Machine Learning Engineer?
Justification: By conducting the survey and gathering the data from industry experts, we can identify the most important skills and responsibilities for a ML engineer, and gain a comprehensive understanding of the industry's requirements and expectations for Machine Learning engineers.
Research Method: The research method used to answer the research question is a survey.

3.2 Method

Having a solid understanding of research methodology is essential. Research methodology refers to the methods or tactics used to locate, pick, process, and analyze data regarding a topic. This chapter discusses the details regarding the method to be employed in this research.

This research aims to bring about the following aspects:

- The educators will know the skills companies look for, and they know what responsibilities the students will have after graduating before choosing their career in ML. By understanding the expectation of the companies, the educators will be able to train the students accordingly, helping them equip themselves to take up a career in ML.
- Software engineers and other personnel aspiring to be ML engineers will have an understanding of the skills and responsibilities that surround the job role.
- The companies may use the results of this research to determine skills that may be relevant for their future hires.

3.3 Research Design

The research design is the theoretical foundation upon which the study will be conducted. It serves as a guide for collecting, evaluating, and analyzing research.
There are two parts of the methodology involved in this research:

- The first part of the research employs the method of Archival Research based on different job advertisements published for the ML Engineer position [36].
- Once the skills needed to become a ML Engineer are obtained, a survey will be conducted to identify the most important skills that are needed and the most looked-for responsibilities for a ML Engineer, which is the second part.

Archival research can be beneficial for this study because the data required to answer the research questions are available in the form of job descriptions taken from the database. This makes the data collection process more valid, efficient, and less resource-intensive. It provides a broader perspective on the topic, as it gives access to a wide range of materials and sources. Archival research provides access to a large set of historical information that is related to the research topic and can provide valuable insights into the development of the field over time. This can help identify the skills and competencies required by software engineers to design a ML system and the responsibilities of ML engineers in companies working with ML systems. However, there are other alternatives such as field experiments and focus groups. Field experiments and focus groups can also bring a more current and deep understanding of the topic. Field experiments are experiments conducted in real-world use cases. However, field experiments are time-consuming and resource-intensive and can also bring ethical concerns. Focus groups involve small groups of people who give valuable insights into the perspectives and opinions on a specific topic. However, they may not be suitable for studying a large and diverse population and concluding the result.

A qualitative expert survey can be used to identify the most important skills and responsibilities of a ML engineer. By gathering opinions and perspectives from experts in the field, the research can provide valuable insights into the expectations and requirements for ML engineers. Surveys provide standardized data, and this data can provide valuable insight into the current state of the industry and what skills and responsibilities are deemed most important by industry professionals. The results can be compared and analyzed to draw conclusions about the research questions. However, there is another alternative research method i.e., interviews. Interviews can provide in-depth data and have the opportunity to ask follow-up questions and probe for further information. However, interviews can be time-consuming and resource-intensive because industry-expertise individuals often avoid answering, which may lead to confusion about whether they are industry experts, and they may have time constraints which lead to invalid responses.

Therefore, research methods archival research and survey can provide data and information that can help answer the research questions and together, they can provide a comprehensive understanding and valid information to support the research findings of the skills and responsibilities required for ML Engineer.

3.4 Archival Research (AR)

AR techniques examine the output and by-products of the SE field, such as source code, documentation, and reports (Lethbridge et. al., 2005) [37]. Locating and gathering the data systematically and evaluating and interpreting data are some of the tasks included (Wohlin and Aurum, 2015) [38]. It does not need a lot of commitment from the subjects and is non-intrusive. Real data, however, may occasionally have something amiss or may be hard to interpret. Additionally, triangulating data from archive research can be very useful for both case studies and research.

Archival research can be classed as either qualitative or quantitative, depending on the data source. Meeting minutes and software documentation are examples of sources of qualitative data, whereas project databases, software repositories, and fault reports are sources of quantitative data. Although archival research is frequently done in business settings, it can also be used for academic work (Molleri et. al., 2019) [36]. The decision-making framework described in (Wohlin and Aurum, 2015) [38] states that systematic literature reviews are academic literature-based archival research. However, secondary studies' criteria entail particular stages relating to study identification, selection, etc., so Systematic Literature review(SLR) are treated individually. Postmortem evaluations look into past data from completed projects to learn lessons for new situations. These light studies make use of retroactive reflection, frequently in conjunction with assistance from project participants. Some examples of sources that could be used in Archival research include newspaper stories, Job advertisements, censuses, sports data, prominent leaders' speeches, and even tweets.

In the case of the current research, the job advertisements are the primary data collection source to find the skills and responsibilities needed to be a ML Engineer. This is because skills and responsibilities are generally included in job advertisements and hence selected as the only data source.

3.4.1 Database

On Kaggle, Data Science(DS) and ML users can communicate online. In order to solve DS problems, Kaggle users can collaborate, access, and share datasets, use notebooks with Graphics Processing Unit(GPU) integration and compete with other DS. The data set used in this research is from Kaggle – International Job postings September 2021.
The dataset link is given below:
https://www.kaggle.com/datasets/techmap/international-job-postings-september-2021

This dataset, which is a sample of 3.4 million global job postings from September 2021, is the result of web scraping work at Techmap.io. More than Job searchers can use Techmap, a workplace search engine, from the global posting list, to locate local businesses that use particular technologies. Data has been gathered and screened for job listings from around the world, identifying pertinent technology and workplace traits in order to identify the technologies employed in firms. On conducting a pre-check, it was found that sufficient ads were related to ML, DS and Artificial Intelligence(AI). Thus, around 6497 job postings were extracted for the purpose of this research based on the job postings of ML and DS as the skills and responsibilities overlapped for the two domains based on the pre-check.

3.4.2 Data Processing Tool

Python(Py) is the program used to process the data. Py is a well-known Programming language(PL) for computers that are used to build websites and applications, automate procedures, and analyze data. As a common language, Py can be used to create a wide range of applications and isn't specifically suited for solving any particular problems. Its flexibility and accessibility for beginners have propelled it to the top of the list of PLs currently in use. According to a survey conducted by the market research company RedMonk in 2021, it was the second-most well-liked PL among developers [39].

Data analysts(DA) and other experts can use Py to execute complex statistical calculations, create ML algorithms, and handle and analyze data, among other tasks now that it has become a standard in the field of data science. A wide range of data visualizations, including line and bar graphs, pie charts, histograms, and three-dimensional plots, may be made with Py. Additionally, Py offers a number of libraries like TensorFlow and Keras that make it easier and faster for programmers to construct data analysis and machine learning applications. [40]

In this research, the concept of Natural Language Processing (NLP) plays a prime role, and the Py provides the perfect platform for this. One can speak, read, and write with the aid of language, which is a form of communication. The study of AI, which enables computers to comprehend and process human language, is known as NLP. Projects involving NLP in the past were only open to professionals since they required a high level of expertise in linguistics, ML, and mathematics. Developers can now focus on creating ML models by using pre-made tools that make text pre-processing easier. To address NLP issues, numerous tools and libraries have been developed. Py is a great option for PL for an NLP project due to a number of factors. This language is a great option for applications requiring NLP because of its straightforward syntax and transparent semantics. Additionally, excellent integration support for tools and other languages that are helpful for techniques like ML is advantageous for developers. This adaptive language's other characteristic, though, is what makes it such a great tool for helping robots understand real languages. It provides developers with a rich collection of NLP tools and libraries to handle a variety of NLP-related tasks, including document classification, topic modelling, part-of-speech (POS) tagging, word vectors, and sentiment analysis. [41]

The library considered for this research is spaCy. SpaCy is a relatively new library that was created for use in actual production. It is so much more approachable than other Py, NLP libraries like Natural Language Tool Kit(NLTK) because of this. The quickest syntactic parser currently on the market is provided by spaCy. Additionally, the toolkit is incredibly quick and effective because it is written in Cython. SpaCy is an open-source, cost-free Py library for high-level NLP. Because it is designed primarily for use in production contexts, SpaCy enables you to develop programs that manage and comprehend enormous quantities of text. It can be used to develop a system for text pre-processing before DL, Natural Language Understanding(NLU), or retrieval of information. [42]

Python is used and imported a few libraries such as Spacy, pandas, numPy etc., to process the data. The data is extracted from the Kaggle website where we found the dataset link which consists of job ads.

3.4.3 Data Preparation or Pre-processing pipeline

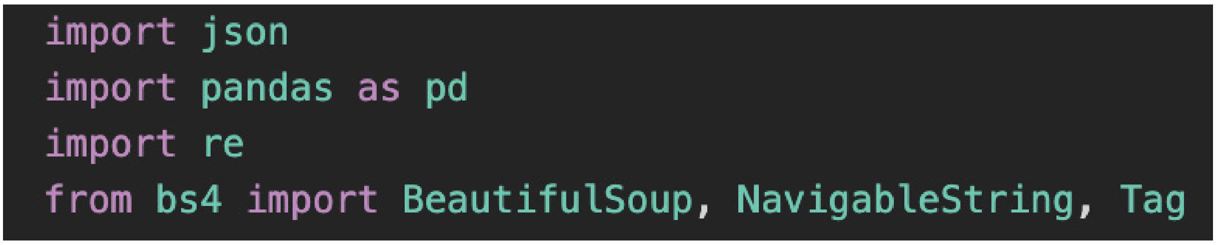
Data preparation for use in research and data visualization applications involves several phases, including data gathering, combining, structure, and organization.

The entire dataset is processed in two stages:

3.4.3.1 Stage 1: Data Loading and Profiling

This stage can be further categorized into the following steps:

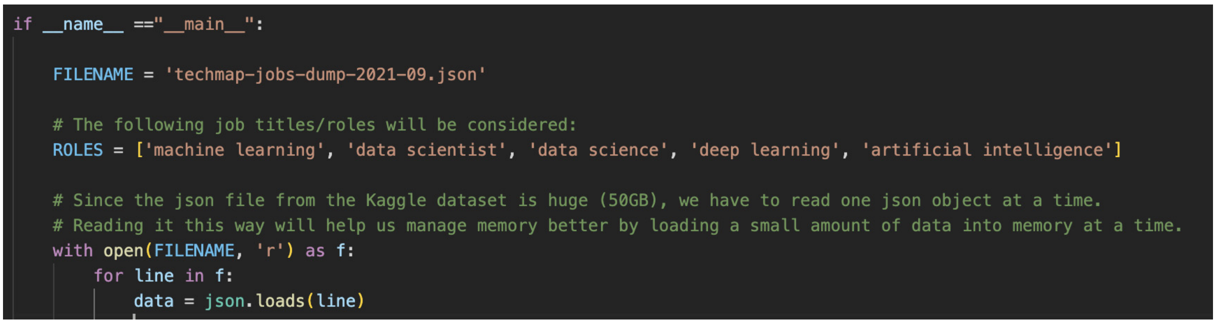
1. **Importing libraries:**
   We have imported json, pandas, re, and beautiful soup. As kaggle dataset is in json format, we have imported json library. A regular expression(re) is used for cleaning the text strings. Beautiful Soup is used for navigating, searching, and modifying a parse tree in HTML.



```
import json
import pandas as pd
import re
from bs4 import BeautifulSoup, NavigableString, Tag
```

Figure 3.1: Code snippet for importing libraries

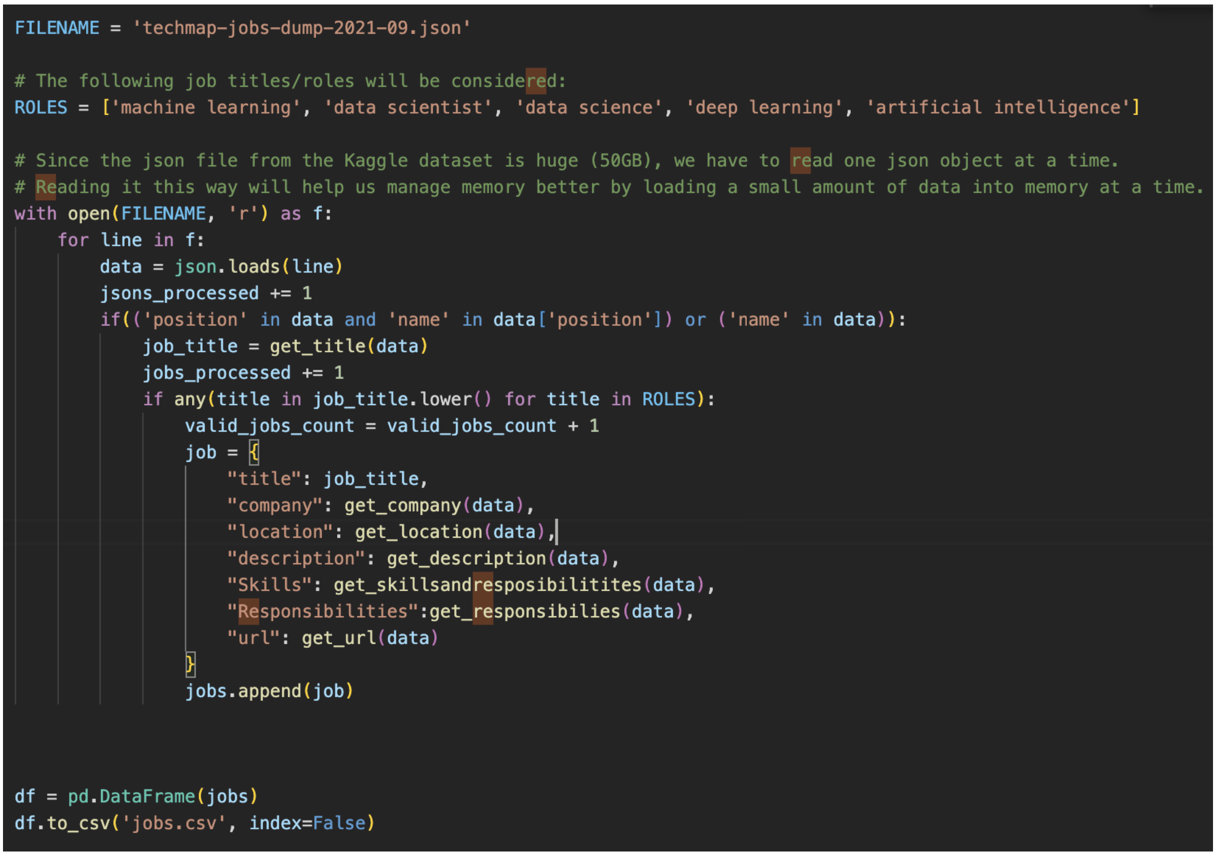
2. **Loading Kaggle dataset:**
   Since the JSON file from the Kaggle dataset is huge (50GB), we read one JSON object at a time. Reading it this way will help us manage memory better by loading a small amount of data into memory at a time.



```
if __name__ =="__main__":

    FILENAME = 'techmap-jobs-dump-2021-09.json'

    # The following job titles/roles will be considered:
    ROLES = ['machine learning', 'data scientist', 'data science', 'deep learning', 'artificial intelligence']

    # Since the json file from the Kaggle dataset is huge (50GB), we have to read one json object at a time.
    # Reading it this way will help us manage memory better by loading a small amount of data into memory at a time.
    with open(FILENAME, 'r') as f:
        for line in f:
            data = json.loads(line)
```

Figure 3.2: Code snippet for importing libraries

3. **Data filtering using Roles:**
   As the dataset contains all the job ads and we require only roles related to machine learning. So we created a list of roles related to machine learning to filter the job titles.



```
FILENAME = 'techmap-jobs-dump-2021-09.json'

# The following job titles/roles will be considered:
ROLES = ['machine learning', 'data scientist', 'data science', 'deep learning', 'artificial intelligence']

# Since the json file from the Kaggle dataset is huge (50GB), we have to read one json object at a time.
# Reading it this way will help us manage memory better by loading a small amount of data into memory at a time.
with open(FILENAME, 'r') as f:
    for line in f:
        data = json.loads(line)
        jsons_processed += 1
        if(('position' in data and 'name' in data['position']) or ('name' in data)):
            job_title = get_title(data)
            jobs_processed += 1
            if any(title in job_title.lower() for title in ROLES):
                valid_jobs_count = valid_jobs_count + 1
                job = {
                    "title": job_title,
                    "company": get_company(data),
                    "location": get_location(data),
                    "description": get_description(data),
                    "Skills": get_skillsandresposibilitites(data),
                    "Responsibilities":get_responsibilies(data),
                    "url": get_url(data)
                }
                jobs.append(job)



df = pd.DataFrame(jobs)
df.to_csv('jobs.csv', index=False)
```

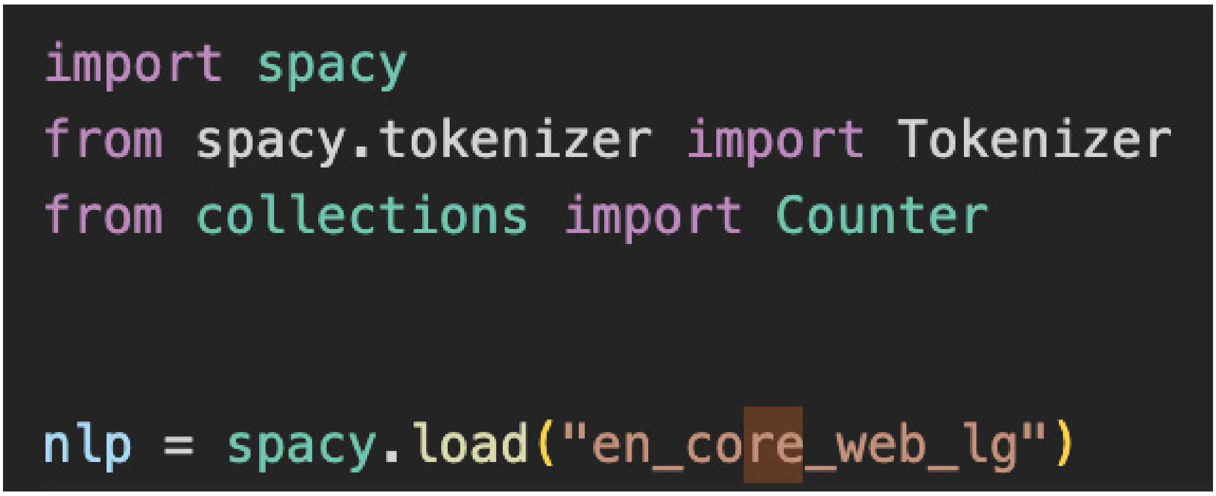
Figure 3.3: Code snippet for filtering the roles related to machine learning

Here for every job ad, if the title contains any one of the roles that particular job ad is selected for fetching the skills and responsibilities. We create a dictionary containing the title, company, location, description, Skills, responsibilities, and URL. We have appended all the filtered job ads to the dictionary.

3.4.3.2 Stage 2: Extracting the required skills and frequency

This stage can be further categorized into the following steps:

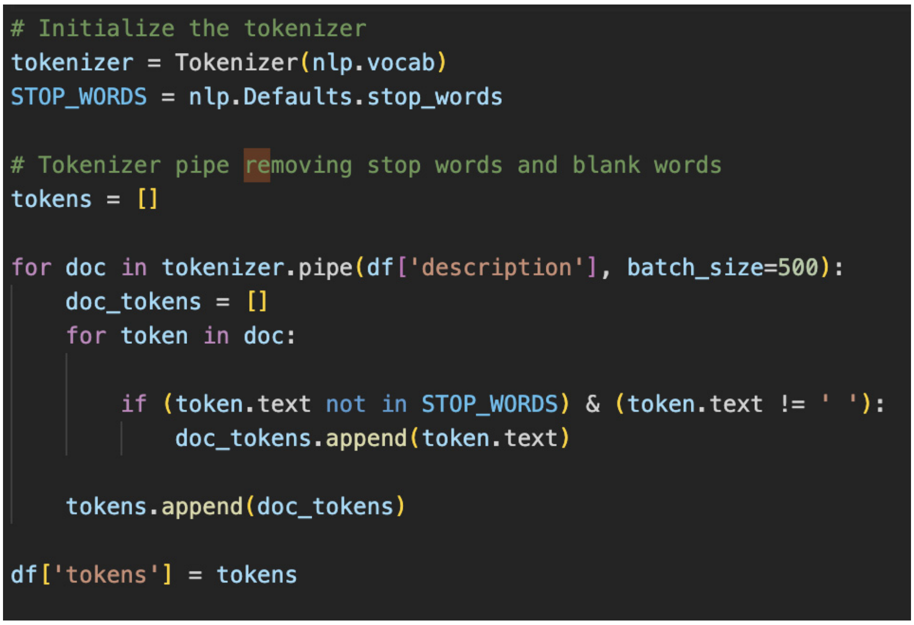
1. **Importing and loading NLP libraries:**
   The Spacy library is imported and offers a number of pre-trained models that may be used to tokenize, tag, and parse text in several languages, including English. A pre-trained spaCy model for the English language is loaded by using nlp.load("en core web lg"). Once loaded, the model may be used to analyze text and extract valuable information from it.



```
import spacy
from spacy.tokenizer import Tokenizer
from collections import Counter


nlp = spacy.load("en_core_web_lg")
```

Figure 3.4: Code snippet for importing NLP libraries

2. **Tokenization:**
   The text in the description column of a Pandas DataFrame is divided into individual words(tokens) using the Tokenizer module. We eliminate stop words and blank words from the resultant tokens and save them in a new column titled tokens. The generated tokens are filtered for each page, excluding stop words and blank words. The token is a list appended to every document containing the generated tokens. With df['tokens'] = tokens, the tokens list is finally inserted as a new column in the original DataFrame.

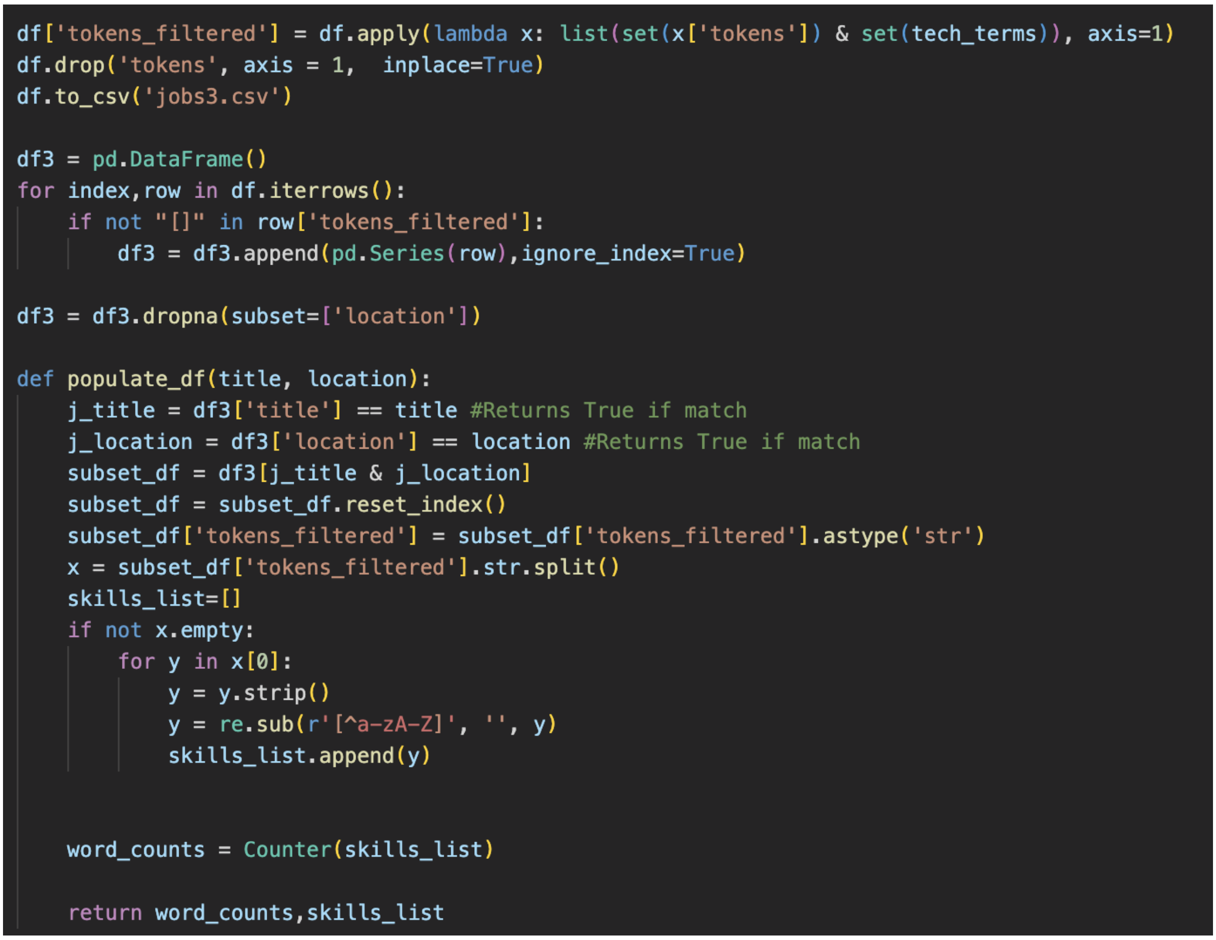


```
# Initialize the tokenizer
tokenizer = Tokenizer(nlp.vocab)
STOP_WORDS = nlp.Defaults.stop_words

# Tokenizer pipe removing stop words and blank words
tokens = []

for doc in tokenizer.pipe(df['description'], batch_size=500):
    doc_tokens = []
    for token in doc:

        if (token.text not in STOP_WORDS) & (token.text != ' '):
            doc_tokens.append(token.text)

    tokens.append(doc_tokens)

df['tokens'] = tokens
```

Figure 3.5: Code snippet for importing NLP libraries

3. **Generated Skills list:**
   We have used the tech terms i.e all the possible list of skills, and taken the

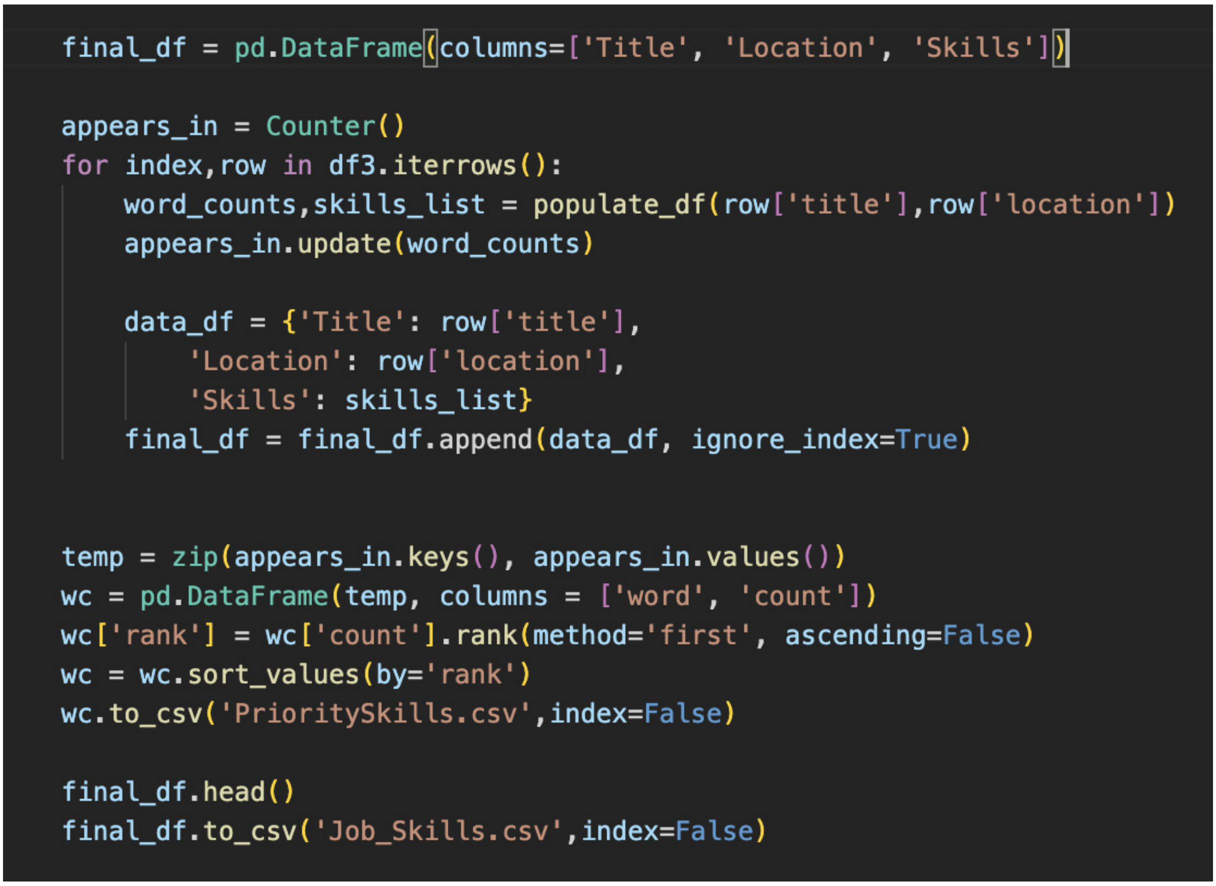
filtered tokens. These filtered token are appended to the skills list to find the frequency of each skill.



```
df['tokens_filtered'] = df.apply(lambda x: list(set(x['tokens']) & set(tech_terms)), axis=1)
df.drop('tokens', axis = 1,  inplace=True)
df.to_csv('jobs3.csv')

df3 = pd.DataFrame()
for index,row in df.iterrows():
    if not "[]" in row['tokens_filtered']:
        df3 = df3.append(pd.Series(row),ignore_index=True)

df3 = df3.dropna(subset=['location'])

def populate_df(title, location):
    j_title = df3['title'] == title #Returns True if match
    j_location = df3['location'] == location #Returns True if match
    subset_df = df3[j_title & j_location]
    subset_df = subset_df.reset_index()
    subset_df['tokens_filtered'] = subset_df['tokens_filtered'].astype('str')
    x = subset_df['tokens_filtered'].str.split()
    skills_list=[]
    if not x.empty:
        for y in x[0]:
            y = y.strip()
            y = re.sub(r'[^a-zA-Z]', '', y)
            skills_list.append(y)


    word_counts = Counter(skills_list)

    return word_counts,skills_list
```

Figure 3.6: Code snippet for importing NLP libraries

4. **Frequency of the skills:**
   We have taken all the skills from the tokens filtered column, and using the counter function, we converted the skills list to a dictionary named Word_count. We created another counter appear_in which updates for every iteration. If the same skill is found in the list, the count of each skill increases or the skill is added.



```
final_df = pd.DataFrame(columns=['Title', 'Location', 'Skills'])

appears_in = Counter()
for index,row in df3.iterrows():
    word_counts,skills_list = populate_df(row['title'],row['location'])
    appears_in.update(word_counts)

    data_df = {'Title': row['title'],
        'Location': row['location'],
        'Skills': skills_list}
    final_df = final_df.append(data_df, ignore_index=True)


temp = zip(appears_in.keys(), appears_in.values())
wc = pd.DataFrame(temp, columns = ['word', 'count'])
wc['rank'] = wc['count'].rank(method='first', ascending=False)
wc = wc.sort_values(by='rank')
wc.to_csv('PrioritySkills.csv',index=False)

final_df.head()
final_df.to_csv('Job_Skills.csv',index=False)
```

Figure 3.7: Code snippet for importing NLP libraries

3.5 Survey

3.5.1 Introduction

According to Shank (2005) [43], qualitative research is a sort of methodical empirical examination. When something is described as systematic, it refers to a planned, organized, and open process that complies with standards endorsed by the qualitative research community. This means that qualitative researchers focus on understanding how people attempt to interpret or make sense of events in their natural settings and aim in the context of the meanings people attribute to them. Survey-based studies have been more prevalent in the last ten years of Software Engineer Research, and several guidelines are available to help individuals eager to do surveys. A qualitative survey is one that collects data to understand people's perspectives, thought processes, motivation and narration in a machine learning field. After the extraction of Skills and responsibilities for machine learning engineers are filled out from the archival research, the research moves into a qualitative survey which is the second phase of the research. It is a less structured approach to research that aim to gain insights of people's opinion and experiences. Qualitative research offers explanations about why and how. From the archival research results, a qualitative survey will be carried out to determine the most crucial skills and responsibilities to becoming a

machine learning engineer. Our main objective to conduct this survey research is to identify skills and responsibilities that are most important for a machine engineer and additional skills and responsibilities that brings benefits to companies.
The following figure gives a brief on the steps to be followed in a survey design according to Morelli et. al., 2020. [36]

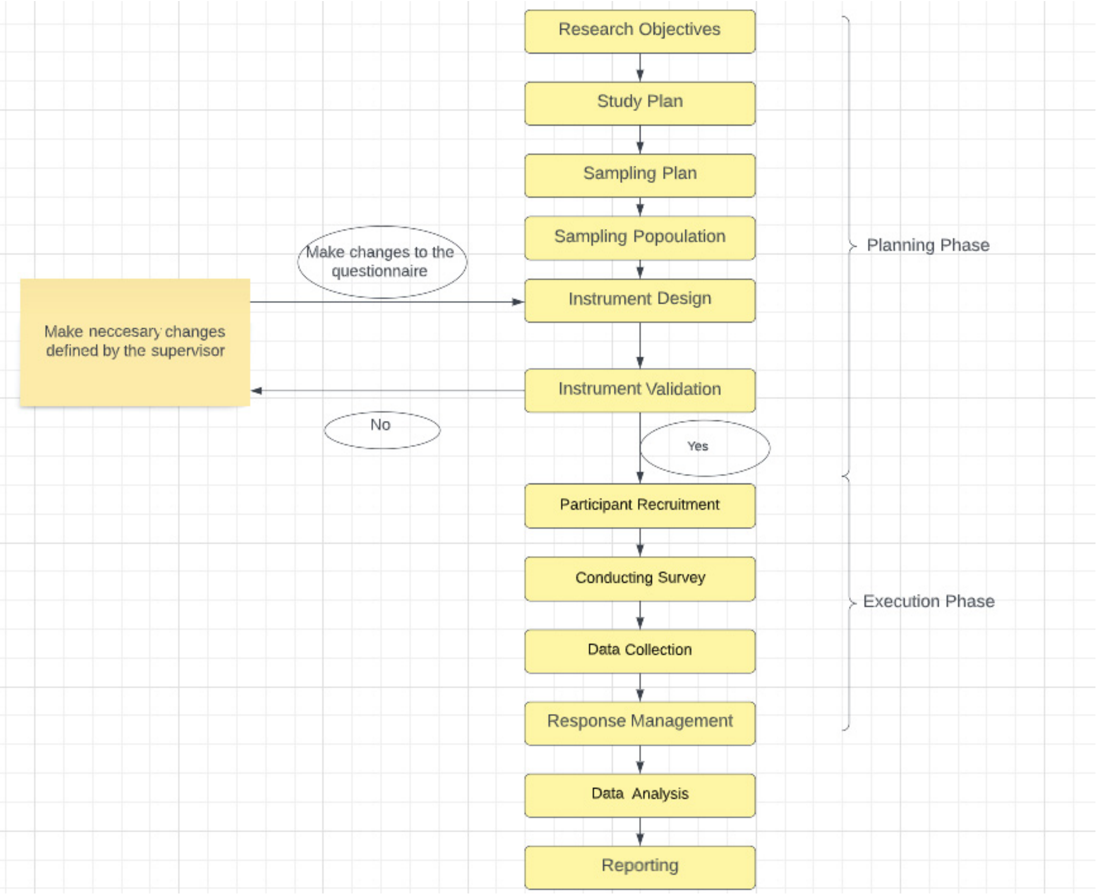


Figure 3.8: Study Design

3.5.2 Stages in Survey

3.5.2.1 Planning Phase

As shown in the Checklist for the survey figure above, this phase has six parts: the research objectives, study plan, population identification, sampling plan, instrument design, and instrument validation. The objectives and study plan have been presented in the previous chapters for this research. The sampling plan, population, and instrument design are discussed in the next sections.

A. Sampling and Sampling Strategy

Since it is impractical to examine every member of the population separately, sampling is frequently utilized. Additionally, it is done to speed up research and conserve resources like money and time. Population sampling is the process of choosing a group of people to represent the entire population. Researchers utilize a range of sampling techniques when dealing with vast populations. Testing the entire community is frequently nearly impossible due to how challenging it is to contact everyone. Researchers employ convenience sampling when extra sources are not necessary for the core research.

The sampling technique in this study is convenience sampling. This phrase describes a method that researchers employ to compile data from a pool of easily reachable respondents for market research. Because it is so rapid, easy, and economical, the sampling procedure is used most frequently. If members choose to be part of the sample, they are typically easy to contact. Because this was an expert assessment, the convenience sampling allowed researchers to concentrate on specific responses. This method works well in this study since the information being collected is very targeted and precise.

B.Sampling Population

The survey was conducted in which the sampling population was experts who are Machine Learning professionals, artificial intelligence professionals, data scientists, and Human resource personnel who handle recruitments for ML Engineer positions. In the current technological world, the most practical way to collect data from the sampling population is by means of online surveys. LinkedIn, Reddit and similar professional discussion marketplaces were used to identify the sampling population based on the position, and field of expertise if in ML, AI, and Data-sciences. For the purpose of this research, online surveys were sent out to 32 professionals, out of which 25 responded back. Thus the population considered for this research is 25.

C.Instrument Design

The instrument design stage involves the preparation of a questionnaire for conducting the survey. The questionnaire is formulated to satisfy the research goal and objectives. The questionnaire was created in google forms; the questionnaire involves close-ended questions where the respondents can rate their answers on the Likert scale of one to five, where five is the highest and one is the lowest rating. The questions in the survey included obtaining expert opinions on the most important skills and responsibilities required by ML engineers. For the survey to determine the important responsibility that ML Engineer has, the experts were asked to rate which responsibilities ML engineers have primarily and then which are not so important. Based on this, the results are populated. The questionnaire contains three main sections:

The questionnaire contains three main sections:

- The first one is to understand the experience level of the respondent for which we have the question designed.
- The second section focuses on the skills that are important to become a ML Engineer. In this section, the skills that have been populated from the archival

research are listed, and the experts rate their importance on a scale of one to five, where five is the most important and one is the least.

- The third section is for the responsibilities that a ML Engineer needs to have. Similar to the previous section, the responsibilities that have been populated from the archival research are listed, and experts rate their importance on a scale of one to five, where five is the most important and one is the least.

Instrument Validation
We understood the guidelines proposed by kitchmen [44], and followed the guidelines to validate the questionnaire. We also received guidance from our thesis supervisor and created the questionnaire for our survey based on the results from our archival research. Our goal was to make the questionnaire easy and straightforward for respondents to answer. We kept their perspective in mind throughout the process, ensuring that the questions were not too complex or difficult to understand. The final questionnaire consists of three close-ended questions; it will take 15-20 minutes for the respondents to answer these questions. We validated the questionnaire with the approval of our supervisor before sending it to the respondents.

The main purpose of this questionnaire was to get the experts to provide their rating on the skills and responsibilities so that the important skills and responsibilities that a ML Engineer needs can be understood.

3.5.2.2 Execution Phase

The execution phase involves participant recruitment and response management. The two stages are explained below.

A.Criteria for the Selection of Participants and Data Collection
Criteria for the Selection of Participants is the process of selecting individuals to take part in a survey. The criteria for selecting the participants is based on the job role the participants are currently working in the industry. To find industry experts in the field of machine learning, we sampled the population and considered several characteristics to select the participants. These characteristics included their industrial experience working with machine learning, any certifications they had related to machine learning, and any achievements they had. We also reviewed their contributions to the industry, if existed, such as any articles or journals they had written. Once we had identified potential participants, we sent them a questionnaire and communicated with them throughout the process. To learn about the experiences of experts in the field of machine learning, we searched for professionals on LinkedIn using the keyword "Machine Learning". We found profiles of people who had job titles such as "Machine Learning Engineer", "data scientist" and "artificial intelligence professional". We considered these professionals suitable participants for our research because implementing machine learning techniques is an important component of data science and artificial intelligence, according to sebastian [45]. We sent a request to 32 of these professionals to participate in our survey, as they have experience in implementing machine learning techniques and tools. The bar chart in the following shows how many of these professionals agreed to participate in our

survey.

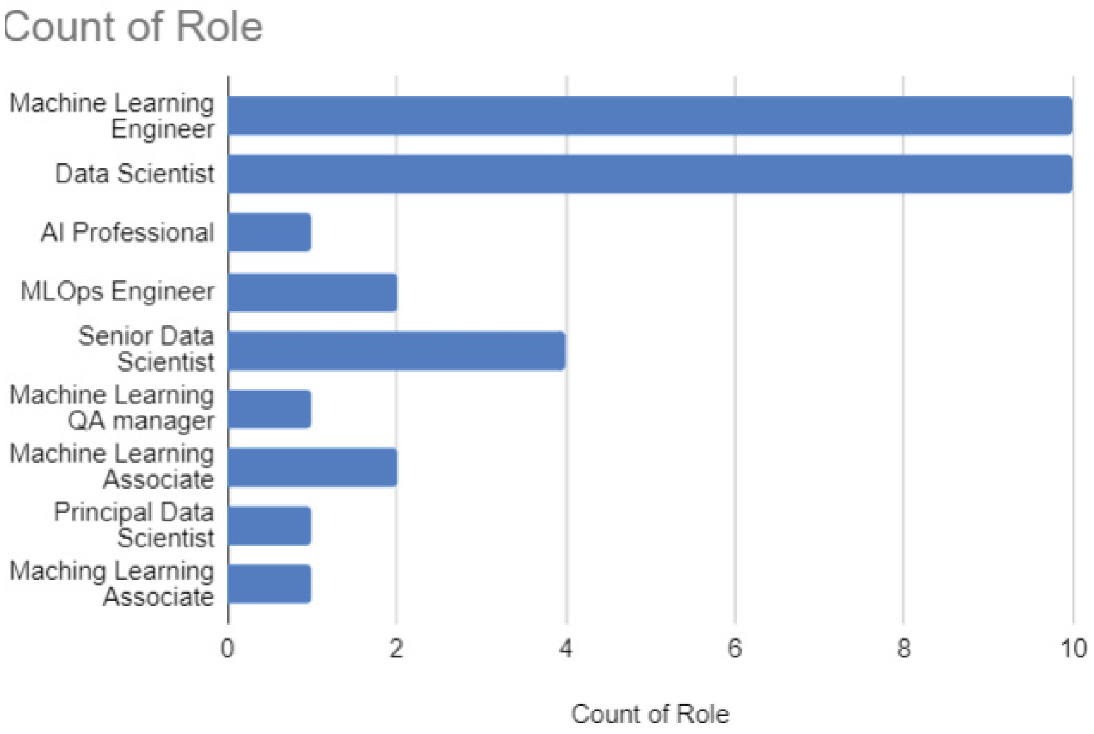


Figure 3.9: Bar Chart of Total number of Practitioners' Role

The following Figure 3.10, column chart defines the industrial experience with machine learning. We gathered data on the industrial experience of these professionals through LinkedIn.

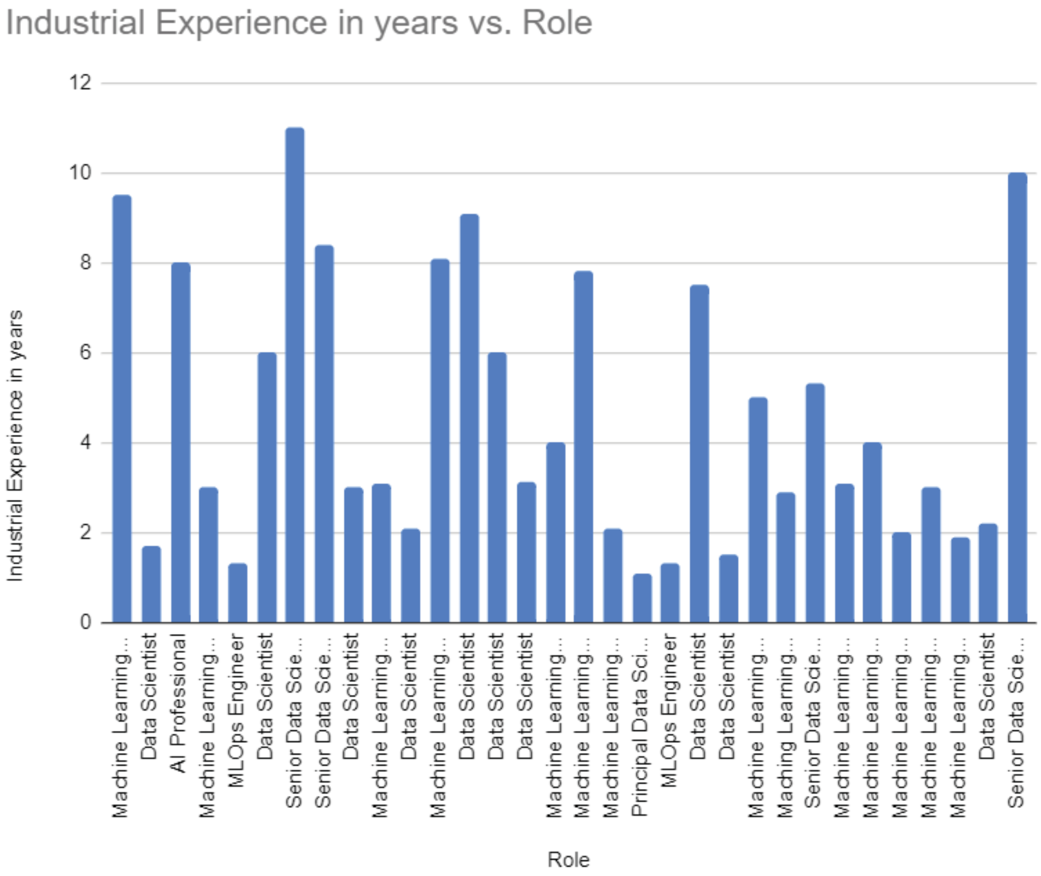


Figure 3.10: Column Chart of Practitioners' Industrial Experience

A qualitative survey consisted of a questionnaire with queries that covered the aspects that can give an insight into the important skills required to be a ML Engineer and the prime responsibilities to look for in a ML engineer. This will help to identify the important skills and responsibilities to become a ML Engineer which is the main aim of this research. The questionnaire was distributed in the form of an e-survey that helped to reach wider expert respondents by not having geographical location as a limitation. The e-survey form was sent to professional IT experts(participants) through LinkedIn, Reddit, and other similar professional marketplaces, and the majority of the respondents responded through LinkedIn.

B.Response Management

In this research, the response is recorded in the form of a Likert Scale, where the respondents can choose on a scale of one to five and give their expert opinion.

Data Processing and Analysis:

This phase involves the processing of the results of the survey. This research analysed the results qualitatively to identify the important skills and responsibilities. The skills that are most repeatedly stressed by the expert respondents will be considered to be those important by inference technique. Usually, surveys can be analyzed using two statistical methods which are descriptive and inferential statistics. Below gives a brief about them:

- **Descriptive statistics:**
  The fundamental metrics used to describe survey data are descriptive statistics. They are made up of brief summaries of single variables and the corresponding survey sample. The frequency and percentage response distributions, measures of central tendency, and dispersion measures such as the range and standard deviation, which describe how close the values or responses are to central tendencies, are a few examples of descriptive statistics for survey data.

- **Inferential Statistics:**
  Inferential statistics uses sample data to draw inferences about the larger population from which the sample was drawn. Since inferential statistics aim to draw conclusions from a sample and generalize them to a population, researchers must be certain that the sample faithfully represents the population. It is not permitted for researchers to choose a nice group. Instead, one can be certain that the sample is representative of the population by employing random selection. The main method for obtaining samples that, on average, represent the population in this process. The mean, one of the statistics generated by random selection, does not commonly have exceptionally high or low values. Inferential statistics can be used to do a more powerful analysis of the findings of an online web survey. As the name would suggest, this branch of statistics is concerned with reaching more broad conclusions concerning social concerns. Examples of this include associations between variables, how well your sample captures the overall population, and cause-and-effect relationships. Examples of inferential statistics that are often used in survey data analysis include t-tests that compare group averages, analyses of variance, correlation, and regression, as well as more complex techniques like factor analysis, cluster analysis, and multidimensional modelling procedures.

Reporting:
The survey results after the statistical analysis of grouping of similar responses and percentage value calculation and frequency analysis will be presented in the form of Pie charts and bar – graphs that will help to understand the results better. The pictorial form of reporting gives a better understanding.

Chapter 4

RESULT AND ANALYSIS

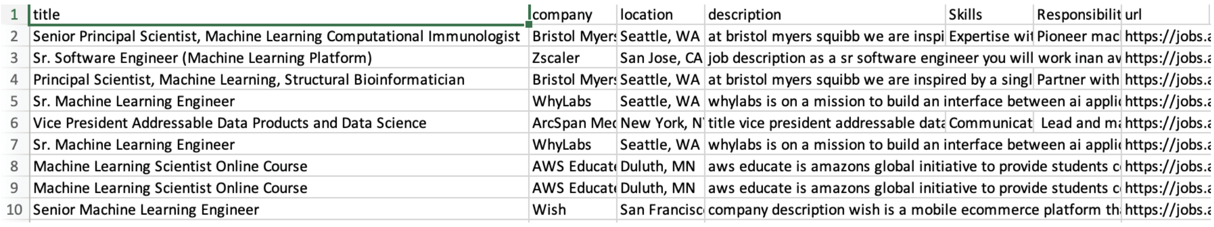
This chapter presents the results obtained from the method employed. Looking at the research questions, the first question entails the skills and competencies that are needed for a SE to become a ML Engineer. The second question deals with the responsibilities that a ML Engineer will have. Both these questions are answered as a part of Archival Research. The results of Archival research are presented in the next section.

4.1 Archival Research Results

The archival research was carried out on job postings related to Data Science, ML from the source

The dataset used in this study is a subset of the 3.4 million global job posts collected through web scraping activities carried out by Techmap.io, a workplace search engine designed to help job seekers find local businesses that use specific technologies. With the ultimate goal of creating a thorough technology knowledge graph, the dataset was gathered in September 2021 and filtered to remove pertinent technologies and workplace traits. The employment market's current situation in connection to technology developments and workplace environments was investigated using the dataset produced as a consequence of this procedure, which includes charting the technologies used in businesses from various sources.

A subset of 6497 job advertisements was found to be specifically related to the roles of "machine learning," "data scientist," "data science," "deep learning," and "artificial intelligence" within the larger dataset of 3.4 million global job postings obtained through web scraping activities by Techmap.io. This subset of job postings was obtained through a filtering process that involved searching for these specific roles in the job titles. Additionally, the skills and responsibilities associated with these job postings were extracted by taking the entire paragraph where they were mentioned in the job description.

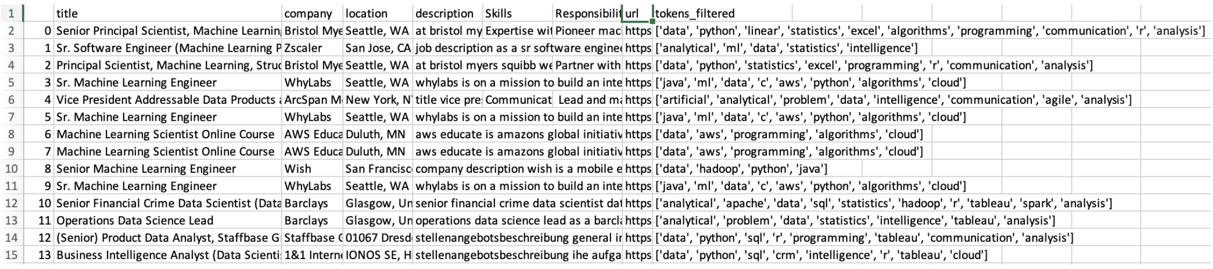


1	title	company	location	description	Skills	Responsibilit	url
2	Senior Principal Scientist, Machine Learning Computational Immunologist	Bristol Myers	Seattle, WA	at bristol myers squibb we are inspi	Expertise wit	Pioneer mac	https://jobs.a
3	Sr. Software Engineer (Machine Learning Platform)	Zscaler	San Jose, CA	job description as a sr software engineer you will		work inan av	https://jobs.a
4	Principal Scientist, Machine Learning, Structural Bioinformatician	Bristol Myers	Seattle, WA	at bristol myers squibb we are inspired by a singl		Partner with	https://jobs.a
5	Sr. Machine Learning Engineer	WhyLabs	Seattle, WA	whylabs is on a mission to build an interface between ai applic			https://jobs.a
6	Vice President Addressable Data Products and Data Science	ArcSpan Mec	New York, N'	title vice president addressable data	Communicat	Lead and ma	https://jobs.a
7	Sr. Machine Learning Engineer	WhyLabs	Seattle, WA	whylabs is on a mission to build an interface between ai applic			https://jobs.a
8	Machine Learning Scientist Online Course	AWS Educate	Duluth, MN	aws educate is amazons global initiative to provide students c			https://jobs.a
9	Machine Learning Scientist Online Course	AWS Educate	Duluth, MN	aws educate is amazons global initiative to provide students c			https://jobs.a
10	Senior Machine Learning Engineer	Wish	San Francisc	company description wish is a mobile ecommerce platform th			https://jobs.a

Figure 4.1: The results initially obtained

4.1.1 Skills

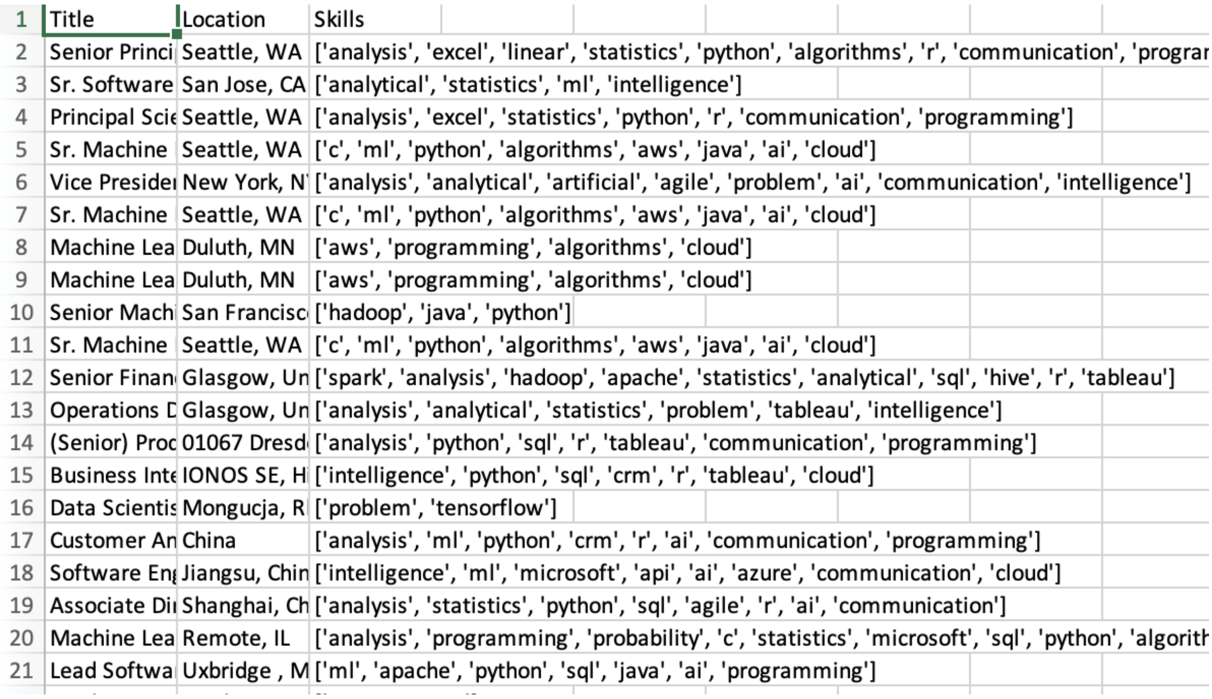
Once the data set was populated, the spacy tool was used to filter the results to obtain the jobs having ML and Artificial Intelligence. Tokenization is also employed to break down the sentences to make them tokens. The code snippet is already outlined in the method section. The results obtained once the filters are applied and processed through spacy are shown below:



1		title	company	location	description	Skills	Responsibilit	url	tokens_filtered
2	0	Senior Principal Scientist, Machine Learning	Bristol Mye	Seattle, WA	at bristol my	Expertise wit	Pioneer mac	https	['data', 'python', 'linear', 'statistics', 'excel', 'algorithms', 'programming', 'communication', 'r', 'analysis']
3	1	Sr. Software Engineer (Machine Learning P	Zscaler	San Jose, CA	job description as a sr software enginee			https	['analytical', 'ml', 'data', 'statistics', 'intelligence']
4	2	Principal Scientist, Machine Learning, Struc	Bristol Mye	Seattle, WA	at bristol myers squibb we		Partner with	https	['data', 'python', 'statistics', 'excel', 'programming', 'r', 'communication', 'analysis']
5	3	Sr. Machine Learning Engineer	WhyLabs	Seattle, WA	whylabs is on a mission to build an inte			https	['java', 'ml', 'data', 'c', 'aws', 'python', 'algorithms', 'cloud']
6	4	Vice President Addressable Data Products a	ArcSpan M	New York, N'	title vice pre:	Communicat	Lead and ma	https	['artificial', 'analytical', 'problem', 'data', 'intelligence', 'communication', 'agile', 'analysis']
7	5	Sr. Machine Learning Engineer	WhyLabs	Seattle, WA	whylabs is on a mission to build an inte			https	['java', 'ml', 'data', 'c', 'aws', 'python', 'algorithms', 'cloud']
8	6	Machine Learning Scientist Online Course	AWS Educa	Duluth, MN	aws educate is amazons global initiativ			https	['data', 'aws', 'programming', 'algorithms', 'cloud']
9	7	Machine Learning Scientist Online Course	AWS Educa	Duluth, MN	aws educate is amazons global initiativ			https	['data', 'aws', 'programming', 'algorithms', 'cloud']
10	8	Senior Machine Learning Engineer	Wish	San Francisc	company description wish is a mobile e			https	['data', 'hadoop', 'python', 'java']
11	9	Sr. Machine Learning Engineer	WhyLabs	Seattle, WA	whylabs is on a mission to build an inte			https	['java', 'ml', 'data', 'c', 'aws', 'python', 'algorithms', 'cloud']
12	10	Senior Financial Crime Data Scientist (Data	Barclays	Glasgow, Un	senior financial crime data scientist dat			https	['analytical', 'apache', 'data', 'sql', 'statistics', 'hadoop', 'r', 'tableau', 'spark', 'analysis']
13	11	Operations Data Science Lead	Barclays	Glasgow, Un	operations data science lead as a barcl:			https	['analytical', 'problem', 'data', 'statistics', 'intelligence', 'tableau', 'analysis']
14	12	(Senior) Product Data Analyst, Staffbase G	Staffbase (	01067 Dresd	stellenangebotsbeschreibung general ir			https	['data', 'python', 'sql', 'r', 'programming', 'tableau', 'communication', 'analysis']
15	13	Business Intelligence Analyst (Data Scienti:	1&1 Intern	IONOS SE, H	stellenangebotsbeschreibung ihe aufga			https	['data', 'python', 'sql', 'crm', 'intelligence', 'r', 'tableau', 'cloud']

Figure 4.2: The results obtained after the filters are applied

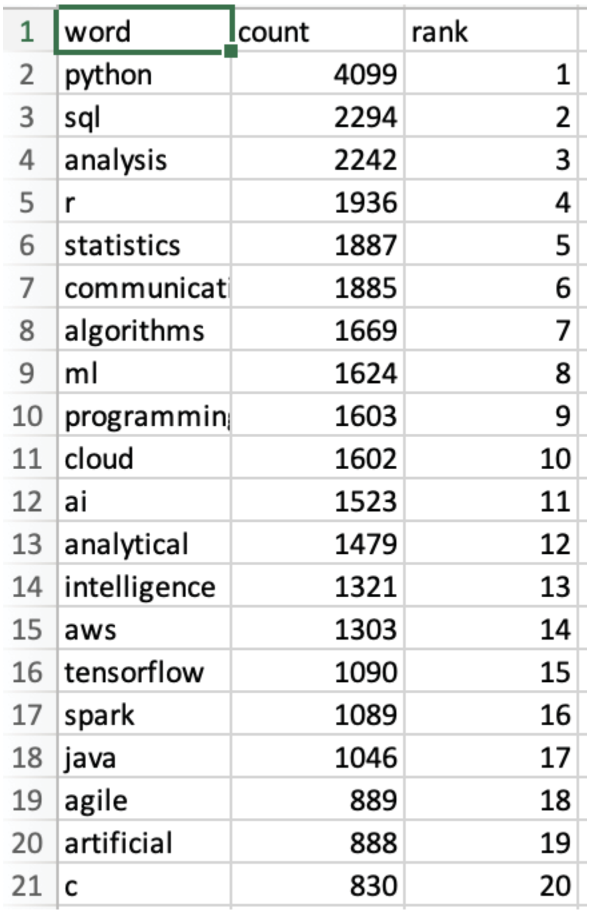
We have created a new data frame and considered the tokens filtered as our skills and removed the empty list from the tokens filtered. On seeing the data set, all the skills are presented in the form of tokens. The results are exported to a .csv file that can be viewed in excel, whose screen snippet is shared in figure 4.3



1	Title	Location	Skills
2	Senior Princi	Seattle, WA	['analysis', 'excel', 'linear', 'statistics', 'python', 'algorithms', 'r', 'communication', 'prograr
3	Sr. Software	San Jose, CA	['analytical', 'statistics', 'ml', 'intelligence']
4	Principal Scie	Seattle, WA	['analysis', 'excel', 'statistics', 'python', 'r', 'communication', 'programming']
5	Sr. Machine	Seattle, WA	['c', 'ml', 'python', 'algorithms', 'aws', 'java', 'ai', 'cloud']
6	Vice Presider	New York, N	['analysis', 'analytical', 'artificial', 'agile', 'problem', 'ai', 'communication', 'intelligence']
7	Sr. Machine	Seattle, WA	['c', 'ml', 'python', 'algorithms', 'aws', 'java', 'ai', 'cloud']
8	Machine Lea	Duluth, MN	['aws', 'programming', 'algorithms', 'cloud']
9	Machine Lea	Duluth, MN	['aws', 'programming', 'algorithms', 'cloud']
10	Senior Machi	San Francisc	['hadoop', 'java', 'python']
11	Sr. Machine	Seattle, WA	['c', 'ml', 'python', 'algorithms', 'aws', 'java', 'ai', 'cloud']
12	Senior Finan	Glasgow, Un	['spark', 'analysis', 'hadoop', 'apache', 'statistics', 'analytical', 'sql', 'hive', 'r', 'tableau']
13	Operations D	Glasgow, Un	['analysis', 'analytical', 'statistics', 'problem', 'tableau', 'intelligence']
14	(Senior) Proc	01067 Dresd	['analysis', 'python', 'sql', 'r', 'tableau', 'communication', 'programming']
15	Business Inte	IONOS SE, H	['intelligence', 'python', 'sql', 'crm', 'r', 'tableau', 'cloud']
16	Data Scientis	Mongucja, R	['problem', 'tensorflow']
17	Customer An	China	['analysis', 'ml', 'python', 'crm', 'r', 'ai', 'communication', 'programming']
18	Software Eng	Jiangsu, Chin	['intelligence', 'ml', 'microsoft', 'api', 'ai', 'azure', 'communication', 'cloud']
19	Associate Dir	Shanghai, Ch	['analysis', 'statistics', 'python', 'sql', 'agile', 'r', 'ai', 'communication']
20	Machine Lea	Remote, IL	['analysis', 'programming', 'probability', 'c', 'statistics', 'microsoft', 'sql', 'python', 'algorith
21	Lead Softwa	Uxbridge , M	['ml', 'apache', 'python', 'sql', 'java', 'ai', 'programming']

Figure 4.3: The results obtained after the filters are applied

The dictionary of skills obtained after the counter subclass is used to analyze how each skill has occurred in the entire description of all the job ads after filtering. The description is used for filtering as there is no particular section for skills in the dataset. In this way, we figured out skills and their frequency in the dataset. By using the rank function of the counter class we obtained the frequency graph for the skills. Here are the list of skills and their rank in .csv file, whose screen snippet is shared in figure 4.4



1	word	count	rank
2	python	4099	1
3	sql	2294	2
4	analysis	2242	3
5	r	1936	4
6	statistics	1887	5
7	communicati	1885	6
8	algorithms	1669	7
9	ml	1624	8
10	programmin	1603	9
11	cloud	1602	10
12	ai	1523	11
13	analytical	1479	12
14	intelligence	1321	13
15	aws	1303	14
16	tensorflow	1090	15
17	spark	1089	16
18	java	1046	17
19	agile	889	18
20	artificial	888	19
21	c	830	20

Figure 4.4: The list of skills and their frequency count

Here are the list of skills based on their frequency count:

1. Python
2. SQL
3. Analysis
4. R
5. Statistics
6. Communication
7. Algorithms
8. Cloud
9. AI
10. Analytical
11. Intelligence
12. AWS

13. TensorFlow
14. Spark
15. Java
16. Agile
17. C
18. Azure
19. NLP
20. Hadoop
21. Tableau
22. Pandas
23. NumPy
24. Microsoft
25. Hive
26. Excel
27. Matlab
28. Apache
29. API
30. Linear
31. Bayesian
32. Probability
33. Deep Learning(DL)
34. Algebra
35. Sci-kit
36. CRM
37. Calculus
38. Neural Networks(NN)
39. Mathematics

The frequency analysis graph is shown below

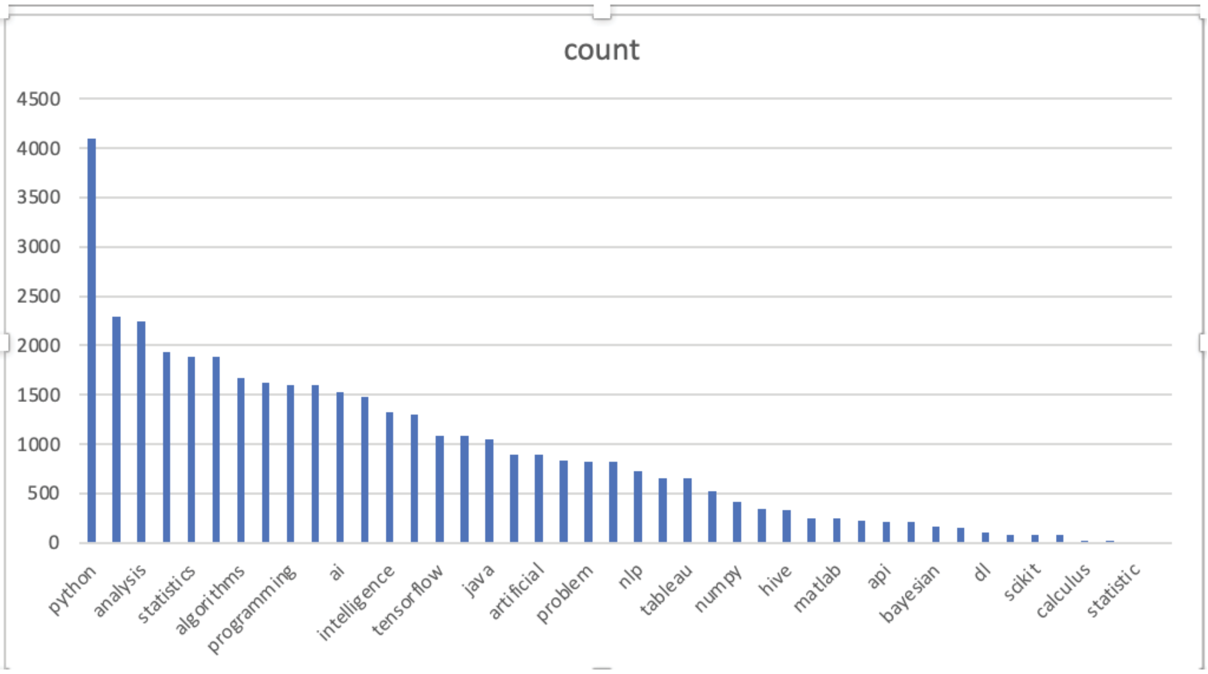
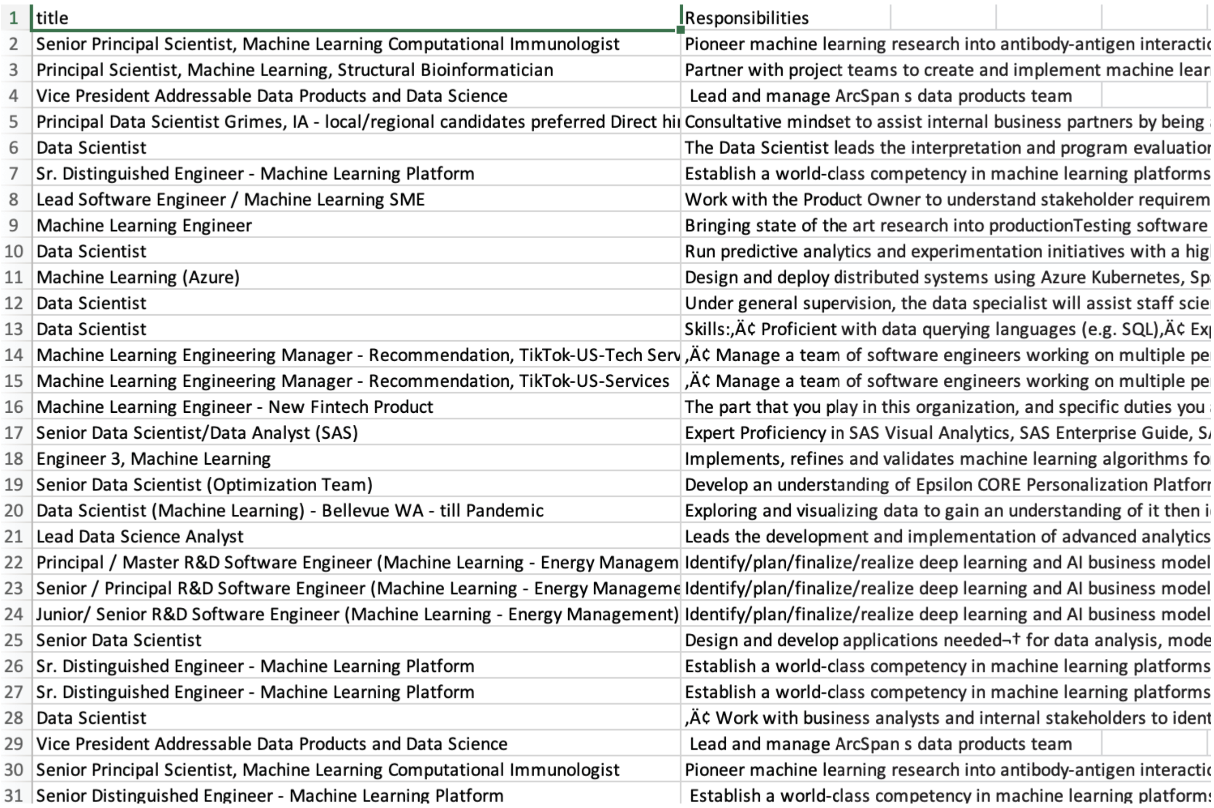


Figure 4.5: Frequency graph of all the skills based on their frequency count

4.1.2 Responsibilities

As figure 4.1 shows the responsibilities column. We have filtered out the responsibilities section by using the beautiful soup library. Beautiful soup is a python library used for fetching data from HTML. As the job ads description in the dataset contains HTML tags, we filtered out the responsibilities by taking the particular tag which contains the keyword "responsibilities" in it. As tokenization is not possible to perform in responsibilities as they are not any particular, unlike skills. We fetched responsibilities for 297 job ads from the dataset. Here are the list of skills and their rank in .csv file, whose screen snippet is shared in figure 4.6



1	title	Responsibilities
2	Senior Principal Scientist, Machine Learning Computational Immunologist	Pioneer machine learning research into antibody-antigen interactic
3	Principal Scientist, Machine Learning, Structural Bioinformatician	Partner with project teams to create and implement machine lear
4	Vice President Addressable Data Products and Data Science	Lead and manage ArcSpan s data products team
5	Principal Data Scientist Grimes, IA - local/regional candidates preferred Direct hi	Consultative mindset to assist internal business partners by being
6	Data Scientist	The Data Scientist leads the interpretation and program evaluatior
7	Sr. Distinguished Engineer - Machine Learning Platform	Establish a world-class competency in machine learning platforms
8	Lead Software Engineer / Machine Learning SME	Work with the Product Owner to understand stakeholder requirem
9	Machine Learning Engineer	Bringing state of the art research into productionTesting software
10	Data Scientist	Run predictive analytics and experimentation initiatives with a hig
11	Machine Learning (Azure)	Design and deploy distributed systems using Azure Kubernetes, Sp
12	Data Scientist	Under general supervision, the data specialist will assist staff scie
13	Data Scientist	Skills:,Ä¢ Proficient with data querying languages (e.g. SQL),Ä¢ Ex
14	Machine Learning Engineering Manager - Recommendation, TikTok-US-Tech Serv	,Ä¢ Manage a team of software engineers working on multiple pe
15	Machine Learning Engineering Manager - Recommendation, TikTok-US-Services	,Ä¢ Manage a team of software engineers working on multiple pe
16	Machine Learning Engineer - New Fintech Product	The part that you play in this organization, and specific duties you
17	Senior Data Scientist/Data Analyst (SAS)	Expert Proficiency in SAS Visual Analytics, SAS Enterprise Guide, S
18	Engineer 3, Machine Learning	Implements, refines and validates machine learning algorithms fo
19	Senior Data Scientist (Optimization Team)	Develop an understanding of Epsilon CORE Personalization Platfor
20	Data Scientist (Machine Learning) - Bellevue WA - till Pandemic	Exploring and visualizing data to gain an understanding of it then i
21	Lead Data Science Analyst	Leads the development and implementation of advanced analytics
22	Principal / Master R&D Software Engineer (Machine Learning - Energy Managem	Identify/plan/finalize/realize deep learning and AI business model
23	Senior / Principal R&D Software Engineer (Machine Learning - Energy Manageme	Identify/plan/finalize/realize deep learning and AI business model
24	Junior/ Senior R&D Software Engineer (Machine Learning - Energy Management)	Identify/plan/finalize/realize deep learning and AI business model
25	Senior Data Scientist	Design and develop applications needed¬† for data analysis, mode
26	Sr. Distinguished Engineer - Machine Learning Platform	Establish a world-class competency in machine learning platforms
27	Sr. Distinguished Engineer - Machine Learning Platform	Establish a world-class competency in machine learning platforms
28	Data Scientist	,Ä¢ Work with business analysts and internal stakeholders to ident
29	Vice President Addressable Data Products and Data Science	Lead and manage ArcSpan s data products team
30	Senior Principal Scientist, Machine Learning Computational Immunologist	Pioneer machine learning research into antibody-antigen interactic
31	Senior Distinguished Engineer - Machine Learning Platform	Establish a world-class competency in machine learning platform

Figure 4.6: Responsibilities after filter are applied

These responsibilities were carefully examined and the list was prepared after the responsibility is read from the csv file.

1. Designing and Modelling of ML systems
2. Research and the implementation of ML systems
3. Datasets and Data Representation
4. Testing of ML systems
5. Training and retraining of systems
6. Algorithm coding
7. Automation of predictive systems
8. Analysing of huge volume of data
9. Solving Problems
10. Extending ML Libraries and frameworks
11. Managerial responsibilities
12. Design and develop applications needed for data analysis, modelling and data visualizations.

13. Participate in the development of both back-end data pipelines and front-end applications

14. Develop and execute innovative machine learning algorithms

15. Generate analytical reports

From the first stage of Archival research, the various skills and responsibilities required for becoming a ML Engineer are collected as presented above. The second stage of research involved the survey, whose results are presented in the next section.

4.2 Survey

ML professionals evaluated the questionnaire's preliminary iteration rigorously. They provided valuable feedback, which helped to refine the questions and clarify the roles and responsibilities of Machine Learning. ML experts who participated in the study were randomly selected through LinkedIn and specialized social media platforms. The participant's expertise was verified through published articles in ML journals, talks at international ML conferences, and sessions at international AI conferences. It is clear that ML's responsibilities and talents are not incompatible. As with classification and prediction, analogy-based learning and resemblance learning are connected. However, people that use ML distinguish between approaches and tasks using analogous concepts, and the survey documents this reality. A total of 25 experts were chosen who are in the ML field, and the following section gives the results of the expert survey that was distributed. The results are presented both in tabular and graphical formats for better understanding.

4.2.1 Survey result for How many years you have worked in the field of ML and Artificial Intelligence?

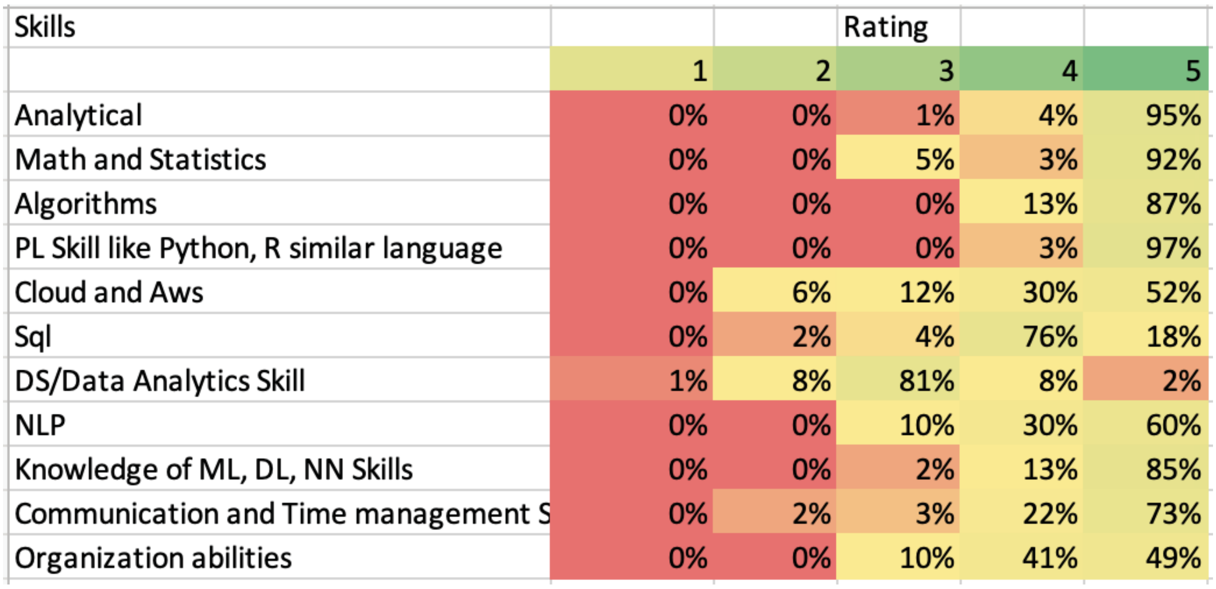
The table below gives the survey result for the years the respondent has been in the ML industry. It can be seen that the respondents are almost equally spread out in the three categories which gives fairly unbiased results.

Experience	Select
0 - 5	16
6 - 10	7
10 +	2

Table 4.1: Result for how many years participants have worked in the field of ML and Artificial Intelligence

4.2.2 Survey Result for: On a scale of 1 to 5 rate the importance of the following skills important for a ML engineer: (5 is the highest and 1 lowest):

The following table gives the result of the survey that was taken to rate the important skills on a scale of one to five. The skill that was of highest importance is given 5 whereas the one in lowest is 1. This way, the experts were able to rate the skill on the level of importance.

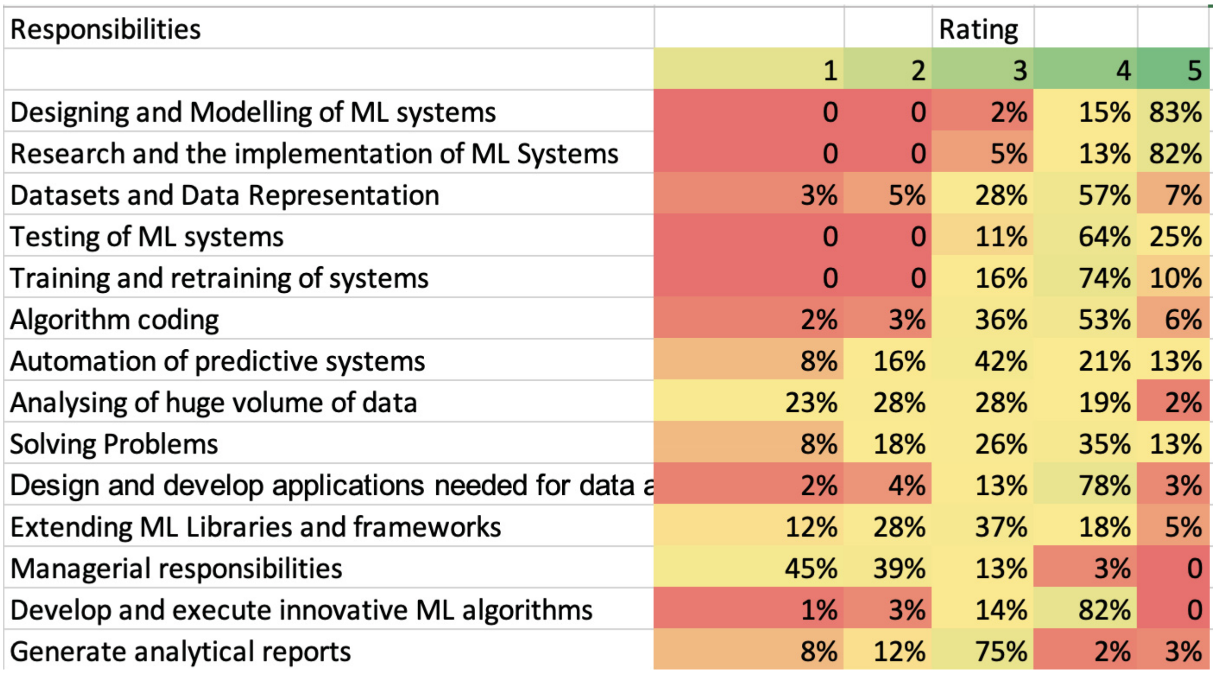


Skills			Rating		
	1	2	3	4	5
Analytical	0%	0%	1%	4%	95%
Math and Statistics	0%	0%	5%	3%	92%
Algorithms	0%	0%	0%	13%	87%
PL Skill like Python, R similar language	0%	0%	0%	3%	97%
Cloud and Aws	0%	6%	12%	30%	52%
Sql	0%	2%	4%	76%	18%
DS/Data Analytics Skill	1%	8%	81%	8%	2%
NLP	0%	0%	10%	30%	60%
Knowledge of ML, DL, NN Skills	0%	0%	2%	13%	85%
Communication and Time management S	0%	2%	3%	22%	73%
Organization abilities	0%	0%	10%	41%	49%

Figure 4.7: Heat Map of Important skill required for ML Engineer

4.2.3 Survey Result for on a scale of 1 to 5 rate the responsibilities that a ML Engineer has: (5 is the highest and 1 lowest):

The following table gives the result of the survey that was taken to rate the responsibilities of ML Engineer requires to do primarily was rated on a scale of one to five, one being least important and five being most important. The responsibilities which the experts feel are most essential are rated as shown below. This way, the experts were able to rate the responsibilities of ML Engineer on the level of importance.



Responsibilities			Rating		
	1	2	3	4	5
Designing and Modelling of ML systems	0	0	2%	15%	83%
Research and the implementation of ML Systems	0	0	5%	13%	82%
Datasets and Data Representation	3%	5%	28%	57%	7%
Testing of ML systems	0	0	11%	64%	25%
Training and retraining of systems	0	0	16%	74%	10%
Algorithm coding	2%	3%	36%	53%	6%
Automation of predictive systems	8%	16%	42%	21%	13%
Analysing of huge volume of data	23%	28%	28%	19%	2%
Solving Problems	8%	18%	26%	35%	13%
Design and develop applications needed for data a	2%	4%	13%	78%	3%
Extending ML Libraries and frameworks	12%	28%	37%	18%	5%
Managerial responsibilities	45%	39%	13%	3%	0
Develop and execute innovative ML algorithms	1%	3%	14%	82%	0
Generate analytical reports	8%	12%	75%	2%	3%

Figure 4.8: Heat map on Results for Responsibilities of a ML Engineer

4.3 Analysis

Based on the results of the Archival research and Expert Survey, the following were the deductions made for the various skills and responsibilities that are required to be a ML Engineer.

4.3.1 Skills required by ML Engineer

The following are the skills With their Weighted Average of Importance:

Here we have calculated the Weighted Average of the results to know which skill is more important. The weighted average is a statistical measure that assigns a weight to each value in a set of data and then calculates the average of these values based on the weights.
In the example given, the weighted average for the skill "Analytical" is calculated as follows:
Step 1: Multiply each rating with its corresponding weight in the data set.
Step 2: Sum up the products from step 1.
Step 3: Divide the sum from step 2 by the total number of ratings (in this case, 5).

So, for the skill "Analytical":
Weighted average = $[(1 * 0) + (2 * 0) + (3 * 1) + (4 * 4) + (5 * 95)]/5 = 98.80$
This calculation gives the weighted average for the "Analytical" skill based on the ratings and weights provided in the data set. Similarly, the weighted average for each

skill can be calculated using the above formula.

Skills			Rating			Weighted avg
	1	2	3	4	5	
Analytical	0%	0%	1%	4%	95%	98.8
Math and Statistics	0%	0%	5%	3%	92%	97.4
Algorithms	0%	0%	0%	13%	87%	97.4
PL Skill like Python, R similar language	0%	0%	0%	3%	97%	99.4
Cloud and Aws	0%	6%	12%	30%	52%	85.6
Sql	0%	2%	4%	76%	18%	82
DS/Data Analytics Skill	1%	8%	81%	8%	2%	60.4
NLP	0%	0%	10%	30%	60%	90
Knowledge of ML, DL, NN Skills	0%	0%	2%	13%	85%	96.6
Communication and Time management Skill	0%	2%	3%	22%	73%	93.2
Organization abilities	0%	0%	10%	41%	49%	87.8

Table 4.2: Skills and weighted average

The weighted average for each skill can be sorted based on importance from high to low by arranging the skills in descending order of their weighted averages. The skill with the highest weighted average is considered the most important and the skill with the lowest weighted average is considered the least important.

Skills	Weighted avg
PL Skill like Python, R similar language	99.40%
Analytical	98.80%
Math and Statistics	97.40%
Algorithms	97.40%
Knowledge of ML, DL, NN Skills	96.60%
Communication and Time management Skill	93.20%
NLP	90%
Organization abilities	87.80%
Cloud and Aws	85.60%
Sql	82%
DS/Data Analytics Skill	60.40%

Table 4.3: Skills ranked from highest to lowest in terms of importance.

Here is a graphical representation of the skills ranked from highest to lowest in terms of importance.

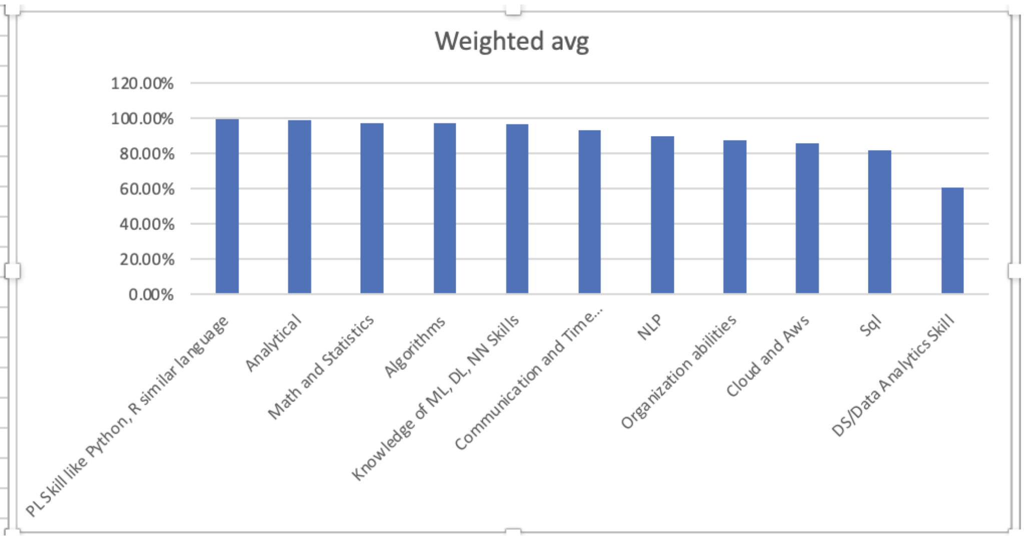


Figure 4.9: Graphical representation of the skills with their importance level

4.3.2 Responsibilities for a ML Engineer

As a deduction from the research, the following are the responsibilities that a ML Engineer require

Here we have calculated the Weighted Average of the results to know which responsibility is more important. The weighted average is a statistical measure that assigns a weight to each value in a set of data and then calculates the average of these values based on the weights.

In the example given, the weighted average for the responsibility "Designing and Modelling of ML systems" is calculated as follows:

Step 1: Multiply each rating with its corresponding weight in the data set.

Step 2: Sum up the products from step 1.

Step 3: Divide the sum from step 2 by the total number of ratings (in this case, 5).

So, for the skill "Designing and Modelling of ML systems":

Weighted average $= [(1 * 0) + (2 * 0) + (3 * 2) + (4 * 15) + (5 * 83)]/5 = 96$

Responsibilities			Rating			Weighted a
	1	2	3	4	5	
Designing and Modelling of ML systems	0	0	2%	15%	83%	96%
Research and the implementation of ML Systems	0	0	5%	13%	82%	95%
Datasets and Data Representation	3%	5%	28%	57%	7%	72%
Testing of ML systems	0	0	11%	64%	25%	82.80%
Training and retraining of systems	0	0	16%	74%	10%	78.80%
Algorithm coding	2%	3%	36%	53%	6%	71.60%
Automation of predictive systems	8%	16%	42%	21%	13%	63.00%
Analysing of huge volume of data	23%	28%	28%	19%	2%	55.40%
Solving Problems	8%	18%	26%	35%	13%	65.40%
Design and dev - data analysis,modeling.	2%	4%	13%	78%	3%	75.20%
Extending ML Libraries and frameworks	12%	28%	37%	18%	5%	55.20%
Managerial responsibilities	45%	39%	13%	3%	0	34.80%
Develop and execute innovative ML algorithms	1%	3%	14%	82%	0	75.40%
Generate analytical reports	8%	12%	75%	2%	3%	56%

Table 4.4: Responsiblities and their weighted average.

The weighted average for each responsibility can be sorted based on importance from high to low by arranging the skills in descending order of their weighted averages. The responsibility with the highest weighted average is considered the most important and the responsibility with the lowest weighted average is considered the least important.

Responsibilities	Weighted avg
Designing and Modelling of ML systems	96%
Research and the implementation of ML Systems	95%
Testing of ML systems	82.80%
Training and retraining of systems	78.80%
Develop and execute innovative ML algorithms	75.40%
Design and develop applications for data analysis, modelling,visualizations.	75.20%
Datasets and Data Representation	72%
Algorithm coding	71.60%
Solving Problems	65.40%
Automation of predictive systems	63.00%
Generate analytical reports	56%
Analysing of huge volume of data	55.40%
Extending ML Libraries and frameworks	55.20%
Managerial responsibilities	34.80%

Table 4.5: Responsiblities ranked from highest to lowest in terms of importance.

Here is a graphical representation of the skills ranked from highest to lowest in terms of importance.

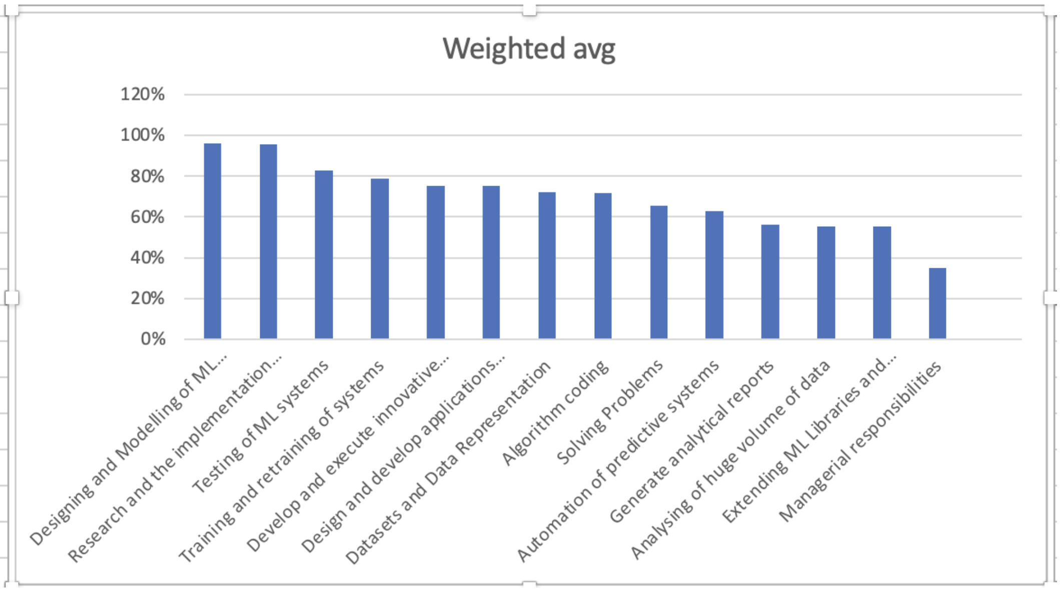


Figure 4.10: Graphical representation of responsibilities with their importance level

Whilst many skills and responsibilities are present, an attempt has been made to combine similar skills into a main category of skill; for example Algebra, statistics, are all Mathematical skills, so they have been categorized as that to give important skills. A similar approach has been carried out for responsibilities as well.

ML algorithms are being created to assess massive amounts of historical data and make predictions. Therefore, being a ML Engineer is an important job requiring many skills and responsibilities. The following chapter presents the discussion and the final chapter, the conclusion of this research.

Chapter 5

Discussion

This chapter briefly discusses various aspects of the previous chapters' results. The statistics demonstrate a growing need for ML jobs and projects. According to a recent projection from Gartner, Inc. [46], the worldwide revenue from AI is expected to reach around sixty-two billion dollars in 2022, an increase of around twenty-one percent from the previous year. The growth of the AI software industry will depend on how far businesses take their AI maturity, according to the senior research director at Gartner.

ML Engineer, DS, and SE with these talents are at the top of the list of professionals that are in a lot of demand, according to a new survey by Indeed [47], the industry leader in online job portals. Numerous well-known corporations, like Univa, Microsoft, Apple, Google, and Amazon, have made significant investments in ML research and development and are utilizing it for their upcoming projects. Given all the developments in ML, it should be no surprise that any enthusiast interested in building a career in software Engineering or technology would be benefitted from machine learning and potentially grow in their career.

One of our thesis's main contributions is providing a skills and responsibilities roadmap for a software engineer to transition into a Machine Learning engineer or work with a ML system. Our study will potentially improve the validity of these articles [21] [33]; we performed an empirical study to identify a set of most important skills and responsibilities that would help software engineers adapt to machine learning and its applications in the respective field. This research focuses specifically on assisting such enthusiasts and provides thorough information on the ML skills required to become a ML Engineer.

5.0.1 What are the skills and competencies that software engineers require to design a ML system?

In this research, it was found that many skills of being a Data Scientist and ML Engineer overlap with one another. ML is attracting a lot of interest from businesses that were seen looking at the number of job postings for the ML engineer, which will open up many chances for those interested in the field. Software engineers adapt to AI capabilities by integrating Machine Learning workflow into the organization's development workflow [17]. There is a reasonable probability that each ML Engineer would give a different answer to the question, "What do you do as a ML engineer?"

Although it seems strange, but this is true. As a result, a ML novice needs to have a comprehensive awareness of the various jobs in that ML expertise can be used. Therefore, the skill set that each of them should have would be different.

We have extracted the list of skills related to machine learning using archival research as our initial method. As our dataset is huge, we first filtered the job title related to machine learning, and then performed NLP techniques to extract the skills. The code snippet of how we extracted the skills is shown in the method section, and all the extracted skills are shown in the results section as a screen snippet of the CSV file. We have uploaded all the CSV files to GitHub.

ML Engineer works with enormous amounts of data to teach a machine and provide it with the information it needs to carry out a certain activity. Being a ML Engineer entails getting ready to handle exciting and difficult problems that will alter how humanity currently perceives the world. Both technical and non-technical expertise is required. In reality, there could be a little bit more to add to this; a ML Engineer will primarily need to know the following skills, as chapter 4 states.

- Background in Programming: Our results state(figure 4.4) that 'C', 'Java', 'R', and 'Python' are the Programming languages a ML Engineer should be proficient in. As each of these skills has a large frequency count, our study says that they are essential for any software engineer aspiring to be a machine learning engineer. The results in the survey section say how important these skills important in terms of percentage. Most of the survey experts have given ratings of 5 out of 5. This clearly states that these skills are important to become a ML engineer.

- Analytical Skill: As our results state, Analytical skill has occurred in 1479 job ads(figure 4.4). This count says that it is one of the important skills for any software engineer aspiring to be a machine learning engineer. Most of the survey experts have given ratings of 5 out of 5, which clearly states that this skill is very important to be a machine learning engineer.

- Statistics and mathematical base: Although mathematics frequency count is less it falls under the category of statistics. It is seen that Statistics have a frequency count of 1887(figure 4.4) in the job ads. It is important to any individual aspiring to be a ML engineer as it helps them to analyze and visualize data to find unseen patterns. While the latter is used to determine the possibility of future happenings, the former deals with the analysis of past events.

- Data-specific knowledge: Data is at the core of ML. Consequently, a skilled ML Engineer is knowledgeable in database management systems, data structures, and data modelling. They can also use visualization software like Tableau, Dash, or Power BI to present their findings.

- Communication and time management Skills: As our results state that communication skill has a frequency count of 1885(figure 4.4). This count clearly says that it is one of the important skills for being a machine learning engineer. When experts were asked about this skill's importance, most rated it important. For instance, ML developers will have to convey ML principles to those who are unfamiliar with the subject.

5.0.2 What are the responsibilities of a Machine Learning Engineer that are required by companies for working with a Machine Learning system?

This is another question commonly put forth by anyone aspiring to enter the ML field. This research has also answered this question by outlining the responsibilities that a ML Engineer will have to carry out. In the method section, we discussed that responsibilities are fetched from the dataset using beautiful soup, which helps to fetch the data in the HTML tags. We could fetch around 300 job ads whose responsibilities are listed(figure 4.6). In the results section, we discussed the responsibilities and their importance based on the survey(table 4.5). These are the responsibilities

1. Designing and Modelling of ML systems
2. Research and the implementation of ML systems
3. Testing of ML systems
4. Training and retraining of systems
5. Develop and execute innovative machine learning algorithms
6. Design and develop applications needed for data analysis, modelling and data visualizations.
7. Datasets and Data Representation
8. Algorithm coding
9. Solving Problems
10. Automation of predictive systems
11. Generate analytical reports
12. Analysing of huge volume of data
13. Extending ML Libraries and frameworks
14. Managerial responsibilities

These are some of the responsibilities that a ML Engineer has. The global ML market is rising as a result of the expanding demand for ML across numerous industrial sectors to enhance customer experience and gain a competitive edge. The job description for a ML Engineer has a lot to offer. ML services are in high demand globally due to the quick uptake of artificial intelligence in several industries, including automotive, healthcare, financial services, government, and others. ML service providers now have a wealth of options as a result of the expansion of ML applications in the manufacturing industry.

Although we have listed many responsibilities, these may not be the most important ones for all the job positions. Responsibilities vary from one job role to another.

This is due to the requirements of any particular company and role.

Each ML position is unique and will have a distinct focus. For instance, some people concentrate on ML, while others concentrate on ML pipelines, large data, and DL. Once the decision to become a ML Engineer is made, the first step is to analyze how to make oneself the best fit. In order to know that, it is essential to know what skills and responsibilities are important to become a ML engineer. The main result that can be taken from this research is the findings on important skills and responsibilities needed to be a ML engineer.

It's critical for anybody thinking about a career in ML to be aware of the needed education and work experience that will ease admission into this industry. Anyone can enter this field; however, the learning path you find most valuable will depend much on where you start. Enrolling in a certification program that walks through the fundamental technical ideas of classical ML, DL, and the more challenging mathematical and statistical concepts commonly employed in the field of ML, represents the most fundamental thing that can be done. Without a basic understanding of these principles, someone pursuing this vocation will find it more difficult and should think about learning them before enrolling in a certification program.

SE, DS, statistician, and people with in-depth knowledge of statistics and advanced mathematics with a comprehension of algorithms are the best candidates for ML certification programs. From the literature, [35] the study was examined on data published between 2012 to 2018, which leads to the drawback where new kinds of ML responsibilities would have been missed; our study identifies the earliest responsibilities from the 2021 dataset required by software engineers. When working in this industry, professionals will spend a lot of time interacting with complex algorithms and vast amounts of data.

5.0.3 What are the most important skills and responsibilities for a Machine Learning Engineer?

This question is the main and important question to conclude this research. This research has also answered this question by highlighting the Skills and Responsibilities that are required to be a ML Engineer. Using the method, we could find the importance of skills and responsibilities. We have taken the survey from the ML professionals who are currently working in the industry. By considering their opinions on the results generated from the archival research adds value to the skills and responsibilities extracted. Here is the list of Skills ranked from high to low:(Based on Table 4.3)

- PL Skill like Python. R or similar language
- Analytical
- Math and Statistics

- Algorithms
- Knowledge of ML, DL, NN Skills
- Communication and Time management Skill
- NLP
- organizational abilities
- cloud AWS
- SQL
- DS/Data Analytics Skills

As a software engineer, it is crucial to understand the most important skills in machine learning (ML). This way, you can concentrate your efforts on developing these skills and becoming a ML engineer. By highlighting the most and least important skills, it provides a clear picture for the reader on which skills to focus on. The results section in the study also includes a weighted average of the importance percentage, which identifies the level of importance for each skill in ML.

By understanding the importance of certain skills in the field of ML, a software engineer can prioritize their efforts and concentrate on developing the most relevant and in-demand skills. The results section of the study clearly indicates the level of importance for each skill by presenting the weighted average of the importance percentage. This information allows the reader to make informed decisions on where to focus their learning and development efforts. By honing these critical skills, software engineers can potentially become an expert in the field of machine learning and advance their careers in this rapidly growing and exciting field.

Here is the list of responsibilities ranked from high to low:(Based on Table 4.5)

- Designing and Modelling of ML System
- Research and ML implementation
- Testing of ML systems
- Training and retraining of systems
- Develop and execute innovative ML algorithms
- Design and develop applications needed for data analysis, modeling, and data visualizations.
- Datasets and Data Representation
- Algorithm Coding
- Solving Problems

- Automation of predictive systems
- Generate analytical reports
- Analysing the huge volume of data
- Extending ML Libraries and frameworks
- Managerial responsibilities

From the studies, [33] [34] [35], their main findings failed to provide a comprehensive list of skills and competencies required for software engineers to work with ML technologies. Thus, our study demonstrates the most important skills and responsibilities a software engineer requires to become a machine learning engineer. As a software engineer, it is imperative to recognize the crucial responsibilities of machine learning (ML). This understanding allows you to concentrate your efforts and develop the necessary skills to become a ML engineer. By highlighting the most important and less important responsibilities, a clear picture emerges for the reader on which skills to focus. The study's results section presents a weighted average of the importance percentage, clearly identifying the importance level for each ML responsibility. By focusing on these key responsibilities and honing their skills, software engineers can likely advance in their careers and potentially become machine learning experts. It is important to note that ML is an ever-evolving field, and staying up to date with the latest developments and advancements is essential to remain at the forefront t of the industry. You can ensure a successful and fulfilling career as a ML engineer by continually developing and refining your skills.

5.1 Validity Threats:

Validity threats can be internal or external or construct or conclusion.

5.1.1 Internal validity threats of the study:

The following are the different internal validity threats.

- History: the particular occurrences in – between measurements. The circumstance can affect the participants and differentiates the result of the research.
- Maturation: change that happens internally due to time lapsed. For example, the participants in the first place were 32; due to a change in participant's attitude internally, our respondents for research were 25.
- Testing: The impact of one test's results on the results of another. If the participants issue subsequent responses, it may affect the previous results.
- Instrumentation: Modifications to the instrument, observers, or scorers that could lead to altered results. Changes made to the questionnaire frequently and sent to the respondents may change the question's meaning.

- Regression to the mean is another name for it in statistics. The choice of subjects based on extreme results or traits is what poses this concern.
- Subject selection: the biases that may influence the choice of comparison groups. The results may be biased if the questionnaire is sent to any subgroup of an organisation or population related to a single organisation.
- Loss of subjects during experiments. If the participants of the population drop out or don't respond, the information will be missing to draw conclusions.
- The choice of comparison groups and maturation can interact, potentially causing misleading results and the incorrect conclusion that the treatment was the source of the impact.

Looking at the different types of internal validity threats, there are two aspects that can be considered are:

Testing:
Since the results of the Programming are what the survey results are going to be based on, getting some altered results from Programming can also alter the survey results marginally.

Instrumentation:
The instrument chosen here is a survey where the questionnaires are sent out to experts; it may be possible that their opinions may vary when sent out to some other experts.

5.1.2 Risk factors for external validity

The following are some of the external validity threats:

- A pre-test may increase, decrease or interact with a subject's receptivity to the experimental variable.
- Effects of inclusion and exclusion standards: Consequences of choosing a sample in accordance with particular standards. This could have affected the study's findings, which wouldn't have happened had there been random sampling.
- Reactive effects of experimental arrangements: If a study's design of experiments had an impact, it would be difficult to extrapolate those results to non-experimental situations.
- Repeated therapy interference: Because the same participants get multiple treatments, it is challenging to account for the effects of earlier therapies.

The external threat that can possibly be said is:
Inclusion and exclusion criteria:
For the requirement of this research, the sampling population has been chosen under the specific criteria of being experts in their field. This may obviously change the results if we were to choose in random population as they may not have the required expertise.

5.1.3 Construct validity

Construct validity defines the extent to which the research methods(survey, archival research) accurately measure and results to answer the research ambition, objectives and theory. Measuring it involves multiple factors such as survey questions, how those questions comply with the research objectives and the relationship between real-world outcomes and the research context [48]. The threats that can occur in our research are during respondent's outcome and research method questions(survey questionnaire). Our research motivation is to identify the skills and responsibilities of being a ML engineer; thus, our objectives for the research were related to identifying important skills and responsibilities to design a ML system. Through archival research, we identified the skills and responsibilities of being a ML engineer; thus, we implemented a survey to validate skills and responsibilities. We created a questionnaire that aimed to point out the research objectives and modified it so that the target audience(participants) could understand it. We designed the survey questionnaire, which complies to answer the research questions with the guidance of our supervisor.

5.1.4 Conclusion validity

The validity of a conclusion refers to the accuracy of the conclusions drawn from research methods, or how trustworthy and credible the results are [48]. In our research, the potential threat to conclusion validity lies in the data collection and drawing of conclusions from the data sources. To address this threat, we sourced our dataset from Kaggle. Furthermore, Kaggle provides reliable and accurate data as it aggregates job postings from popular job sites such as CareerBuilder, EURES, Monster, Seek, etc.

The other threat lies in the validity of data collected from IT experts in the fields of AI, ML and DS by sending a questionnaire and ensuring their career experience in these fields. This helped to improve the validity of the results by getting valuable insights from experts in the field. Thus, by verifying the reliability of the data sources, we were able to draw accurate conclusions and identify the most important skills and responsibilities to become a ML engineer.

Chapter 6

Conclusion and Future Scope

This chapter presents the conclusion of this research. Various aspects are involved in this research, and various findings are outlined in this chapter.

6.1 Conclusion

To develop ML systems that create a large number of ML models, ML engineers must assess, organize, and analyze data, conduct testing, and enhance the learning process. Systems that run by themselves for the automated predictive model are of special interest to ML engineers. By combining SE and data analysis, ML engineers let machines learn without the additional program.

As more industries adopt AI and automation, from the rise of techniques that can recommend in entertainment and retail to the self-learning computers of self-driving cars, ML is at the forefront of fusing data and information science to help business optimize their offerings and services.

This research aimed to determine the most important skills needed to be a ML Engineer. Additionally, the roles and responsibilities required to be a ML Engineer were also analyzed. To obtain these results the job descriptions for the ML Engineer and Data Science roles were examined for this study. Using archival research, skills categories needed were found. Similarly, the responsibilities that ML Engineer requires were also collected. Based on these results, the survey was conducted to rank the required importance of skill and responsibility to analyze the results. The main findings in this research showed the important skills. The research questions are used to determine the skills and responsibilities that are needed to become a ML engineer.

The current research is conducted, identified and ranked the abilities currently needed for AI/ML roles. The findings have three potential applications. First, the existing teams like Data Analysts or DS professionals, might update their current AI/ML offerings and can use these as a reference. Therefore, the design of related undergraduate and graduate degree programs may benefit from this work. Second, recruiters who are on the lookout for skilled AI/ML personnel could make use of the findings. The recruiters and their departments can apply the suggested abilities while making the job profile and description for the open positions to hire. The hiring team might make effective use of their staff in this way. The third main application and the core purpose of this research are to empower software engineers as to what they need to equip themselves with if they would like to become ML Engineers.

The major findings in this research showed that some skills are primary to becoming a ML engineer. The previous chapters have provided insight into what would be needed. Concluding on those, the following is the list of important skills needed for becoming a ML Engineer ranked based on the importance level from high to low.

- PL skills like Python, R or similar language
- Analytical
- Math and Statistics
- Algorithms
- Knowledge of ML,DL, NN skills
- Communication and time management skill
- NLP
- Organizational abilities
- cloud and AWS
- SQL
- DS/Data Analytics skill

Looking at the part of the research involving responsibilities and concluding on those, the following is the list of important responsibilities needed for becoming a ML Engineer ranked based on the importance level from high to low.

- Designing and Modelling of ML System
- Research and ML implementation
- Testing of ML systems
- Training and retraining of systems
- Develop and execute innovative ML algorithms
- Design and develop applications needed for data analysis, modelling , and data visualizations.
- Datasets and Data Representation
- Algorithm Coding
- Solving Problems
- Automation of predictive systems

- Generate analytical reports
- Analysing the huge volume of data
- Extending ML Libraries and frameworks
- Managerial responsibilities

The list presented outlines the comprehensive skills and responsibilities of a ML Engineer; however, it should be noted that this is not an exhaustive list as there are many other aspects to the role. This suggests that the research objectives have been met and the research questions have been answered. In the current era, ML has become a highly sought-after and well-regarded skill, bringing numerous benefits to various industries and improving people's quality of life. Surveys indicate that corporations are focusing heavily on this subject, making it clear that expertise in ML will be crucial for them to achieve their goals.

In conclusion, the Machine Learning (ML) field is constantly evolving and growing. There is no set roadmap to becoming a ML engineer, as success in this field largely depends on an individual's dedication and commitment to continuous learning and staying up to date with the latest advancements and developments. With new discoveries being made every day, it is imperative to stay current in order to excel as a ML engineer. This research aims to assist aspiring Software engineers in navigating the path to success and equip them with the necessary skills and knowledge to achieve their goals.

6.2 Future Scope

The findings of this research provide a valuable starting point for future work in the field of ML Engineer. One potential direction for future research is a longitudinal study to examine changes in the skills and responsibilities of ML engineers over time as the field continues to evolve. Additionally, a cross-cultural study could provide insights into cultural differences in the skills and responsibilities deemed important for ML engineers. A qualitative study could also be conducted to explore the reasons behind the skills and responsibilities deemed important, providing a deeper understanding of industry professionals' perspectives and experiences. Another potential direction is a case study examining the skills and responsibilities of ML engineers in a specific industry or organization, providing practical insights into their implementation. A similar method can be employed to determine the skills and responsibilities for any other job role. Additionally, the various categories involved in the ML Engineer role can be analyzed more deeply. We have uploaded the code and results to our Github which could be used for future work.

Finally, developing a training program or certification process based on the skills and responsibilities identified in this research could provide a framework for professional development in the field and help bridge the gap between ML and SE engineers. These future works can help to build on the findings of this research and contribute to the ongoing development and understanding of the field of ML Engineer.

References

[1] D. Harris. What is artificial intelligence (ai)? how does ai work? https://builtin.com/artificial-intelligence, 2022.

[2] A. S. Gillis. What is a machine learning engineer? what do they do? https://www.techtarget.com/searchenterpriseai/definition/machine-learning-engineer-ML-engineer, 2022.

[3] Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR, 2021.

[4] John D Kelleher, Brian Mac Namee, and Aoife D'arcy. *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT press, 2020.

[5] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[6] Ethem Alpaydin. *Machine learning: the new AI*. MIT press, 2016.

[7] Stefano A Bini. Artificial intelligence, machine learning, deep learning, and cognitive computing: what do these terms mean and how will they impact health care? *The Journal of arthroplasty*, 33(8):2358–2361, 2018.

[8] Jerrold S Maxmen. The post-physician era: medicine in the twenty-first century. 1976.

[9] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

[10] C David Naylor. On the prospects for a (deep) learning health care system. *Jama*, 320(11):1099–1100, 2018.

[11] Iqbal H Sarker. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3):1–21, 2021.

[12] Jonathan Schmidt, Mário RG Marques, Silvana Botti, and Miguel AL Marques. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials*, 5(1):1–36, 2019.

[13] S. Prasad. Machine learning engineer vs data scientist - job role, skills and salary," blogs updates on data science, business analytics, ai machine learning, 02-sep-2020. https://www.analytixlabs.co.in/blog/machine-learning-engineer-vs-data-scientist/, 2022.

[14] João Ricardo Sato, Marcelo Queiroz Hoexter, Pedro Paulo de Magalhães Oliveira Jr, Michael John Brammer, Declan Murphy, Christine Ecker, MRC AIMS Consortium, et al. Inter-regional cortical thickness correlations are associated with autistic symptoms: a machine-learning approach. *Journal of psychiatric research*, 47(4):453–459, 2013.

[15] Du Zhang and Jeffrey JP Tsai. Machine learning and software engineering. *Software Quality Journal*, 11:87–119, 2003.

[16] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. pages 560–575, 2021.

[17] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. Software engineering for machine learning: A case study. pages 291–300, 2019.

[18] Yehia Elkhatib, Jon Whittle, and Bashar Nuseibeh. The impact of machine learning on software engineering education and practice. *IEEE Software*, 37(5):36–43, 2020.

[19] Olimar Borges, Valentina Lenarduzzi, and Rafael Prikladnicki. Preliminary insights to enable automation of the software development process in software startupsan investigation study from the use of artificial intelligence and machine learning. pages 37–38, 2022.

[20] Mark Harman. The future of software engineering: Learning to learn machine learning. *ACM SIGSOFT Software Engineering Notes*, 45(6):36–39, 2020.

[21] Will Ballard. Why every software engineer should learn machine learning. *Medium*, May 2018.

[22] Thomas Zimmermann, Nachiappan Nagappan, and Harald Gall. Why software engineers should learn machine learning: A survey study. *IEEE Transactions on Software Engineering*, 45(6):537–558, 2019.

[23] Romeo J. Ai and machine learning jobs gain growing popularity. `https://hrexecutive.com/in-search-of-artificial-intelligence/`, 2020.

[24] Silvia Fareri, Gualtiero Fantoni, Filippo Chiarello, Elena Coli, and Anna Binda. Estimating industry 4.0 impact on job profiles and skills using text mining. *Computers in industry*, 118:103222, 2020.

[25] Mirjana Pejic-Bach, Tine Bertoncel, Maja Meško, and Živko Krstić. Text mining of industry 4.0 job advertisements. *International journal of information management*, 50:416–431, 2020.

[26] Stefan Debortoli, Oliver Müller, and Jan vom Brocke. Comparing business intelligence and big data skills. *Business & Information Systems Engineering*, 6(5):289–300, 2014.

[27] Adrian Gardiner, Cheryl Aasheim, Paige Rutner, and Susan Williams. Skill requirements in big data: A content analysis of job advertisements. *Journal of Computer Information Systems*, 58(4):374–384, 2018.

[28] Pietro Giorgio Lovaglio, Mirko Cesarini, Fabio Mercorio, and Mario Mezzanzanica. Skills in demand for ict and statistical occupations: Evidence from web-based job vacancies. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 11(2):78–91, 2018.

[29] Amit Verma, Kirill M Yurov, Peggy L Lane, and Yuliya V Yurova. An investigation of skill requirements for business and data analytics positions: A content analysis of job advertisements. *Journal of Education for Business*, 94(4):243–250, 2019.

[30] Eduard Anton, Alina Behne, and Frank Teuteberg. The humans behind artificial intelligence–an operationalisation of ai competencies. 2020.

[31] Henrik Kortum, Jonas Rebstadt, and Oliver Thomas. Dissection of ai job advertisements: A text mining-based analysis of employee skills in the disciplines computer vision and natural language processing. In *Proceedings of the 55th Hawaii International Conference on System Sciences*, 2022.

[32] Leo Mrsic, Hrvoje Jerkovic, and Mislav Balkovic. Interactive skill based labor market mechanics and dynamics analysis system using machine learning and big data. In *Asian Conference on Intelligent Information and Database Systems*, pages 505–516. Springer, 2020.

[33] Daniela Cruzes, John Grundy, and Liming Zhu. Software engineering competencies for the future of machine learning. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 28(4):16, 2019.

[34] David López-Mena, Mario Piattini, and Francisco Ruiz. Machine learning and software engineering competencies for industry 4.0. *IEEE Software*, 37(3):91–97, 2020.

[35] Juliana Soares, Tayana Conte, and Carla Silva. Machine learning competencies for software engineers: A systematic mapping study. *IEEE Transactions on Education*, 62(2):100–108, 2019.

[36] Jefferson Seide Morelli, Kai Petersen, and Emilia Mendes. Cerse-catalog for empirical research in software engineering: A systematic mapping study. *Information and Software Technology*, 105:117–149, 2019.

[37] Timothy C Lethbridge, Susan Elliott Sim, and Janice Singer. Studying software engineers: Data collection techniques for software field studies. *Empirical software engineering*, 10(3):311–341, 2005.

[38] Claes Wohlin and Aybüke Aurum. Towards a decision-making structure for selecting a research design in empirical software engineering. *Empirical Software Engineering*, 20(6):1427–1455, 2015.

[39] S. O'Grady. The redmonk programming language rankings: June 2021. `https://redmonk.com/sogrady/2021/08/05/language-rankings-6-21/`, 2021.

[40] What is python used for? a beginner's guide. `https://www.coursera.org/articles/what-is-python-used-for-a-beginners-guide-to-using-python`, 2022.

[41] B. Lutkevich and E. Burns. Title. https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP, 2022.

[42] S. Kakarla. What is natural language processing? an introduction to nlp. https://www.activestate.com/blog/natural-language-processing-nltk-vs-spacy, 2021.

[43] G Shank. Qualitative research: A personal skills approach. alexandria, 2005.

[44] Barbara Kitchenham, O. Pearl Brereton, David Budgen, Mark Turner, John Bailey, and Stephen Linkman. Systematic literature reviews in software engineering – a systematic literature review. *Information and Software Technology*, 51(1):7–15, 2009. Special Section - Most Cited Articles in 2002 and Regular Research Papers.

[45] Sebastian Raschka, Joshua Patterson, and Corey Nolet. Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *Information*, 11(4):193, 2020.

[46] Gartner, inc. https://www.gartner.com/en/newsroom/press-releases/2021-11-22-gartner-forecasts-worldwide-artificial-intelligence-software-market- 2022.

[47] Indeed. Top jobs in demand: 2021 update. 2021.

[48] Per Runeson and Martin Höst. Guidelines for conducting and reporting case study research in software engineering. *Empirical software engineering*, 14:131–164, 2009.

[49] Thar Htin Shar. Students' perspectives on artificial intelligence in education. 2019.

[50] Adi Shamir and Yael Lenein. Introducing machine learning in computing education: Development of computational thinking skills in 12-year-old students through a learning-by-design course. *Education and Information Technologies*, 26(4):4257–4274, 2021.

[51] Mark Harman, Seyed-Hassan Mirian-Hosseinabadi, Yue Jia, Yuanyuan Zhang, Jens Krinke, Xiaozhe Wang, David Binkley, Xin Yao, Tao Xie, Claire Le Goues, et al. The future of software engineering: Machine learning and other emerging technologies. pages 1204–1218, 2019.

[52] F. Lyzanets. Machine learning in education: Impact on the industry. https://keenethics.com/blog/machine-learning-in-education, 2022.

[53] Machine learning in education. https://aws.amazon.com/education/ml-in-education/#:~:text=Machine%20learning%20(ML)%20is%20transforming,unlock%20new%20discoveries%20and%20insights, 2022.

[54] Gartner forecasts worldwide artificial intelligence software market to reach 62 billion in 2022. https://www.gartner.com/en/newsroom/press-releases/2021-11-22-gartner-forecasts-worldwide-artificial-intelligence-software-market- 2022.

[55] P. Berander and A. Andrews. 4 requirements prioritization," diva-portal.org. `https://www.diva-portal.org/smash/get/diva2:836447/FULLTEXT01.pdf.`, 2022.

[56] Joyce K Fletcher. The paradox of postheroic leadership: An essay on gender, power, and transformational change. *The leadership quarterly*, 15(5):647–661, 2004.

[57] Payne. Nlp with spacy and business tools you can build right now - an introduction to nlp tools. `https://www.width.ai/post/spacy-nlp-business-tools`, 2022.

[58] E. Burns C. Stedman and M. K. Pratt. What is data preparation? an in-depth guide to data prep. `https://www.techtarget.com/searchbusinessanalytics/definition/data-preparation.`, 2022.

[59] Author. spacy 101: Everything you need to know · spacy usage documentation. `https://spacy.io/usage/spacy-101`, 2022.

[60] Patrik Berander and Anneliese Andrews. Requirements prioritization. In *Engineering and managing software requirements*, pages 69–94. Springer, 2005.

[61] Jefferson Seide Molléri, Kai Petersen, and Emilia Mendes. An empirically evaluated checklist for surveys in software engineering. *Information and Software Technology*, 119:106240, 2020.

[62] Fullgence M Mwakondo. *A model for mapping graduates' skills to industry roles using machine learning techniques: A case of software engineering*. PhD thesis, University of Nairobi, 2018.

[63] Bogdan Walek and Ondrej Pektor. Data mining of job requirements in online job advertisements using machine learning and sdca logistic regression. *Mathematics*, 9(19):2475, 2021.

[64] WooJeong Kim, Soyoung Rhim, John YJ Choi, and Kyungsik Han. Modeling learners' programming skills and question levels through machine learning. In *International Conference on Human-Computer Interaction*, pages 281–288. Springer, 2020.

[65] Maud Chassignol, Aleksandr Khoroshavin, Alexandra Klimova, and Anna Bilyatdinova. Artificial intelligence trends in education: a narrative overview. *Procedia Computer Science*, 136:16–24, 2018.

[66] Ido Roll and Ruth Wylie. Evolution and revolution in artificial intelligence in education. *International Journal of Artificial Intelligence in Education*, 26(2):582–599, 2016.

[67] David McArthur, Matthew Lewis, and Miriam Bishary. The roles of artificial intelligence in education: current progress and future prospects. *Journal of Educational Technology*, 1(4):42–80, 2005.

[68] L Pomsar, M Zorkovsky, B Rusinakova, and I Zolotova. Exposing students to ai and industry related skills via openlab high school–industry–university partnership. In *2019 17th International Conference on Emerging eLearning Technologies and Applications (ICETA)*, pages 649–654. IEEE, 2019.

[69] Gilad Shamir and Ilya Levin. Teaching machine learning in elementary school. *International Journal of Child-Computer Interaction*, 31:100415, 2022.

[70] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[71] Ming Xue and Changjun Zhu. A study and application on machine learning of artificial intellligence. In *2009 International Joint Conference on Artificial Intelligence*, pages 272–274. IEEE, 2009.

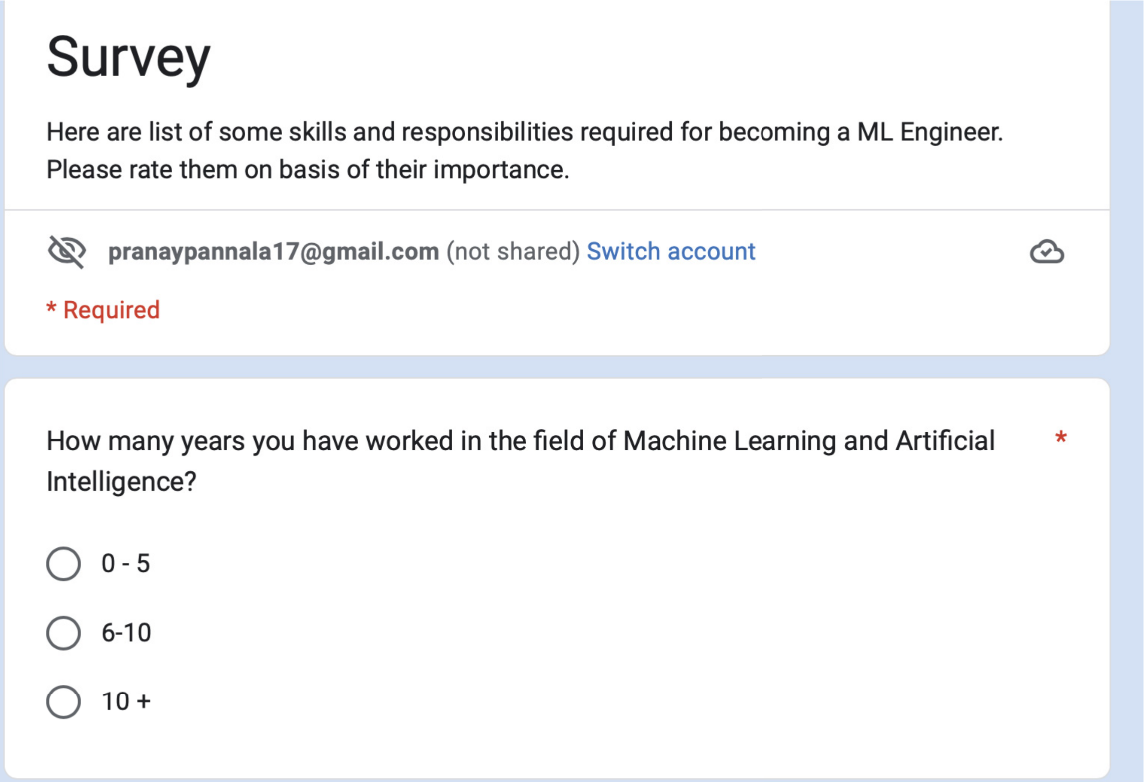
Appendix A

Supplemental Information

A.1 Github Repository

https://github.com/Pranay1718/ML-Jobs-and-Skills-Analysis

A.2 Snapshots of Questionnaire



Survey

Here are list of some skills and responsibilities required for becoming a ML Engineer. Please rate them on basis of their importance.

pranaypannala17@gmail.com (not shared) Switch account

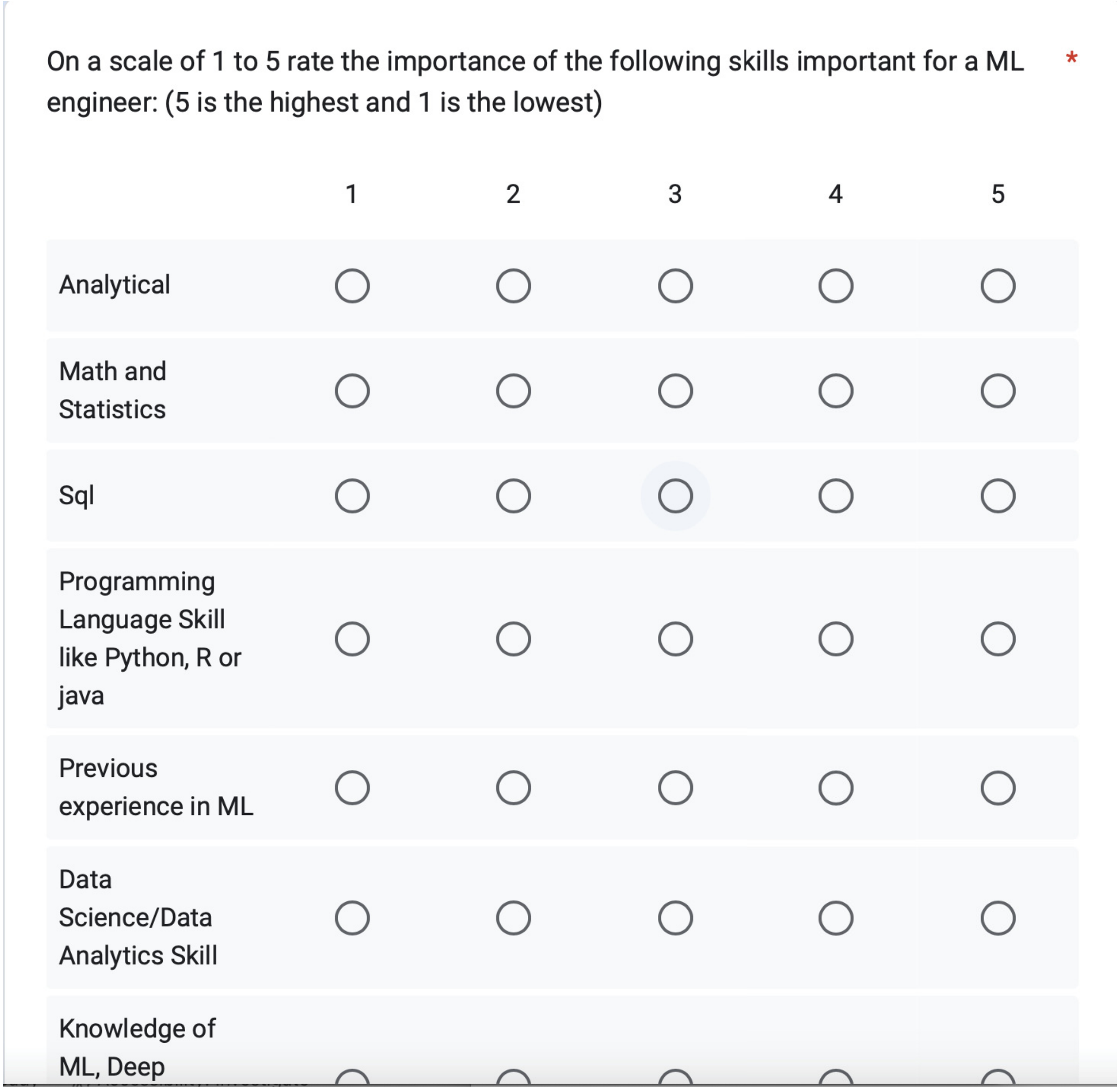
* Required

How many years you have worked in the field of Machine Learning and Artificial Intelligence? *

○ 0 - 5

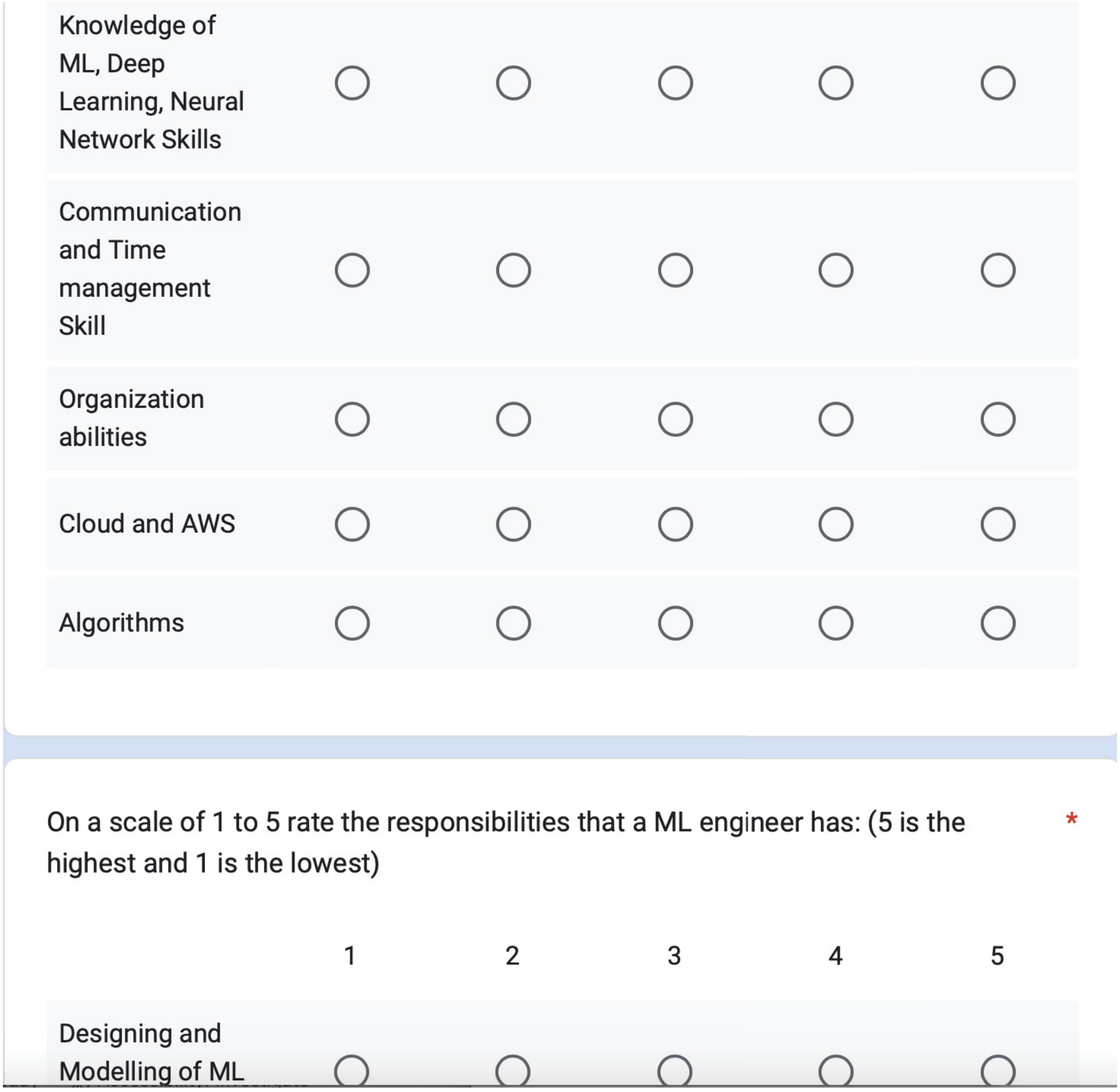
○ 6-10

○ 10 +



On a scale of 1 to 5 rate the importance of the following skills important for a ML engineer: (5 is the highest and 1 is the lowest) *

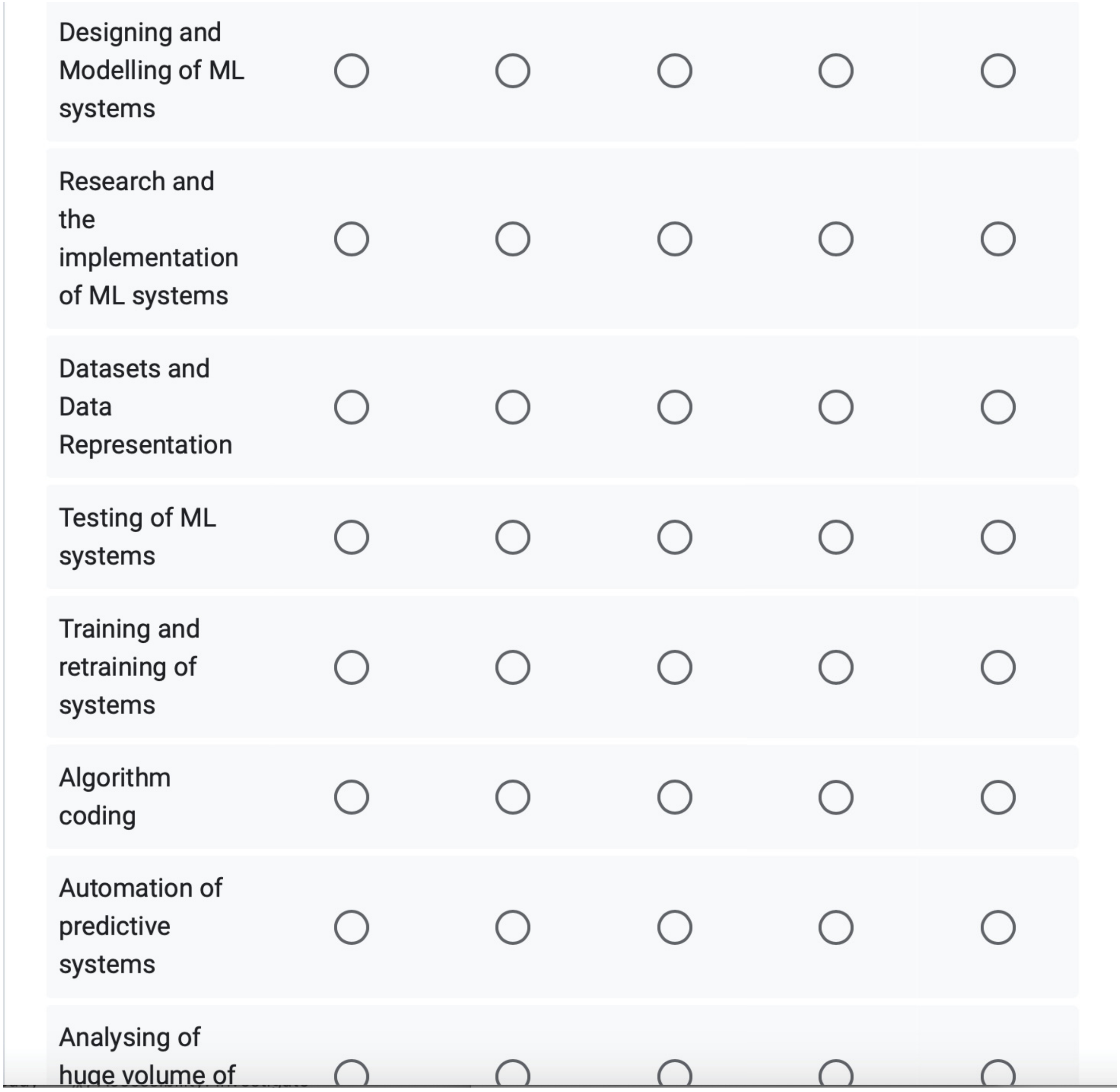
	1	2	3	4	5
Analytical	○	○	○	○	○
Math and Statistics	○	○	○	○	○
Sql	○	○	○	○	○
Programming Language Skill like Python, R or java	○	○	○	○	○
Previous experience in ML	○	○	○	○	○
Data Science/Data Analytics Skill	○	○	○	○	○
Knowledge of ML, Deep	○	○	○	○	○



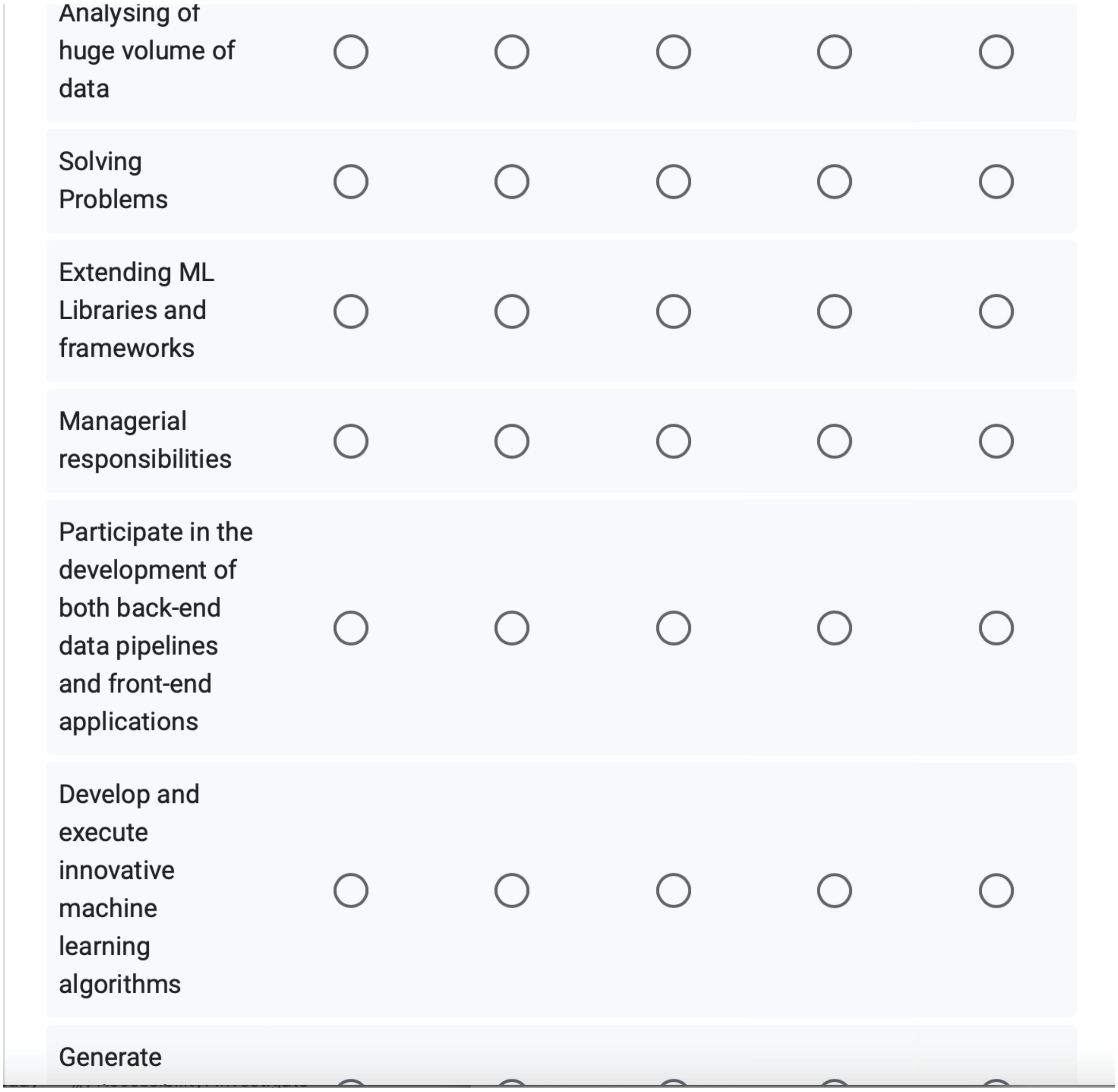
Knowledge of ML, Deep Learning, Neural Network Skills	○	○	○	○	○
Communication and Time management Skill	○	○	○	○	○
Organization abilities	○	○	○	○	○
Cloud and AWS	○	○	○	○	○
Algorithms	○	○	○	○	○

On a scale of 1 to 5 rate the responsibilities that a ML engineer has: (5 is the highest and 1 is the lowest) *

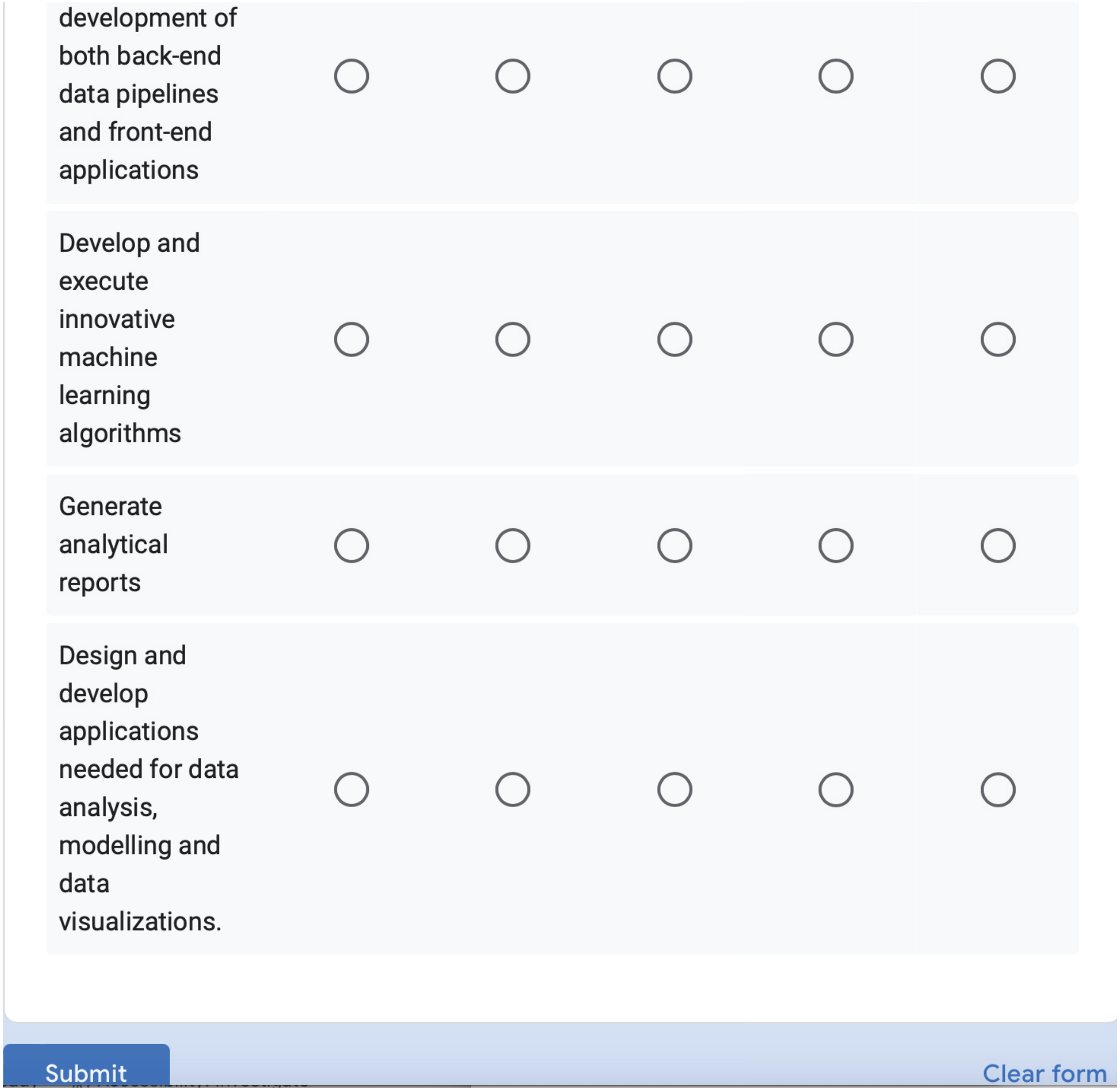
	1	2	3	4	5
Designing and Modelling of ML	○	○	○	○	○



Designing and Modelling of ML systems	○	○	○	○	○
Research and the implementation of ML systems	○	○	○	○	○
Datasets and Data Representation	○	○	○	○	○
Testing of ML systems	○	○	○	○	○
Training and retraining of systems	○	○	○	○	○
Algorithm coding	○	○	○	○	○
Automation of predictive systems	○	○	○	○	○
Analysing of huge volume of	○	○	○	○	○



Analysing of huge volume of data	○	○	○	○	○
Solving Problems	○	○	○	○	○
Extending ML Libraries and frameworks	○	○	○	○	○
Managerial responsibilities	○	○	○	○	○
Participate in the development of both back-end data pipelines and front-end applications	○	○	○	○	○
Develop and execute innovative machine learning algorithms	○	○	○	○	○
Generate	○	○	○	○	○



development of both back-end data pipelines and front-end applications	○	○	○	○	○
Develop and execute innovative machine learning algorithms	○	○	○	○	○
Generate analytical reports	○	○	○	○	○
Design and develop applications needed for data analysis, modelling and data visualizations.	○	○	○	○	○

Faculty of Computing, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden