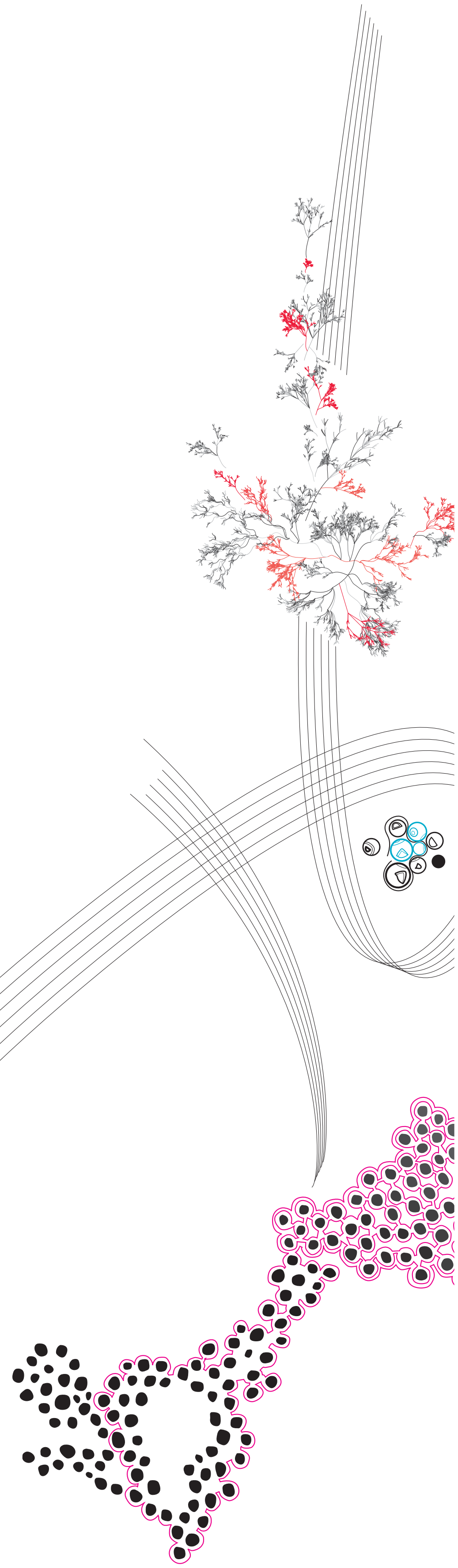


Technology-Driven Career Path Prediction: A Case of Southeast Asia's Public Organization

by

Juwita Pebriana Pasaribu

Supervisors:
Dr. Maya Daneva
Dr. Jeewanie Jayasinghe Arachchige
Dr. Faizan Ahmed

A Master's thesis submitted to the
Faculty of Electrical Engineering, Mathematics
& Computer Science (EEMCS)
in partial fulfilment of the requirements for the degree of
MSc in Computer Science

Department of Computer Science
Faculty of Electrical Engineering, Mathematics
and Computer Science (EEMCS)
University of Twente

August, 2024

UNIVERSITY OF TWENTE.

Acknowledgements

As I reflect on the journey I have undertaken over the past two years, two Bible verses resonate deeply with me: Matthew 21:22, *"If you believe, you will receive whatever you ask for in prayer"* and Philippians 1:3, *"Every time I think of you, I give thanks to my God."* Pursuing a degree in another country has been one of my biggest dreams and a constant prayer since junior high school. God graciously answered that prayer and also blessed me by surrounding me with people who have supported me along this journey. Therefore, I wish to express my deepest gratitude to God the Father, Jesus Christ, my Lord and Savior, and the Holy Spirit for providing everything at the perfect time and for giving me the strength to wait patiently for God's timing.

I would like to express my sincere gratitude to Dr. Maya Daneva, Dr. Jeewanie Jayasinghe Arachchige, and Dr. Faizan Ahmed for their guidance, support, and the much-needed words of encouragement throughout the process of completing this research. I would also like to thank the experts at the National Civil Service Agency for their invaluable insights in helping me understand the background context of this research, as well as the Ministry of Communications and Informatics of the Republic of Indonesia for funding my Master's degree at the University of Twente.

I would also like to extend my heartfelt thanks to my family for their unwavering support, prayers, love, and encouragement. To my mom, thank you for your recipes that have kept me sane in a country where affordable, comforting food is hard to come by. To my dad, thank you for always being proud of me and for providing everything I needed. To Anggi, thank you for being such an understanding sister.

Special thanks to my dear friends who are always there to lend an ear and listen to my endless ramblings. To InDeChrist, which has grown in numbers and will hopefully continue to do so, thank you for every gathering, every laugh, Sunday school jamming sessions, fun picnics, game nights, prayers, and all the memories we have created together that continually reassure me of God's love. To ICF Enschede, thank you for providing a space where I could serve the Lord alongside you. Thanks to my fellow cohort, who have been my sounding board throughout this study journey. And to my dear soul sisters in Indonesia, thank you for being my place to confide in and for helping me heal the wounds in my heart through our countless video calls. I may not mention your names here, but please know that it is you I am truly thankful for.

Lastly, I want to thank myself for striving to be a better version of who I am every single day. Thank you for believing in God and holding on until you could smile, even knowing that tears might come again. Thank you for finishing this degree on time. Thank you for being present in the moment and for your willingness to embrace new people, new places, and new experiences. And thank you for never giving up, even when the journey felt overwhelming. Let's continue to put our faith in God as we step into the next season.

Enschede, 22 August 2024
Juwita

Abstract

In recent years, there has been a growing interest in using machine learning to improve human resource management, such as predicting employee attrition, recruitment needs, and performance. This thesis focuses on predicting career paths for employees in government organizations in Southeast Asia, an area previously overlooked by researchers. The goal is to develop a machine learning-based career path prediction framework for civil servants in this region. To achieve this, the research involves a literature review and gap analysis to identify research opportunities and define system requirements. It also includes implementing and analyzing the most suitable machine learning methods for accurate career path prediction.

The thesis adopts an empirical research method. Given the complexity and importance of career progression in the public sector, we evaluate four machine learning algorithms to identify the most effective model for accurate career path prediction. The models assessed include Decision Tree (DT), Random Forest (RF), XGBoost, and Multilayer Perceptron (MLP). These models were chosen based on a comprehensive literature review that highlighted their effectiveness in similar applications.

Adhering to the CRISP-DM methodology, the research encompasses phases of business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The dataset, sourced from the National Civil Service Agency (NCSA), underwent extensive preprocessing. The evaluation focused on key performance metrics: accuracy, precision, recall, and F1-score. The findings revealed that XGBoost outperformed other models for functional positions, demonstrating superior accuracy and handling of complex data interactions. For managerial positions, while Random Forest exhibited the highest accuracy, MLP showed better precision and F1-score, indicating its effectiveness in minimizing false positives.

The work in this thesis significantly contributes to both research and practice by addressing a critical gap in career path prediction within the public sector, specifically for civil servants in Southeast Asia. By developing tailored predictive models based on real-world data, this research provides valuable insights for optimizing workforce planning, enhancing talent management, and informing policy decisions in public service organizations. Moreover, it enriches the global conversation on career path prediction by integrating region-specific data and methodologies, presenting a validated framework that underscores the technical viability and practical advantages of incorporating machine learning into human resource management systems.

Keywords: Career Path Prediction, Machine Learning, Public Sector, Requirement Engineering.

Contents

Abbreviation List

API	Application Programming Interface
CRISP-DM	CRoss-Industry Standard Process for DataMining
DT	Decision Tree
DTO	Data Transfer Objects
HR	Human Resource
ML	Machine Learning
MLP	Multi-Layer Perceptron
NCSA	National Civil Service Agency
RE	Requirement Engineering
RF	Random Forest
SE	Software Engineering
SHAP	SHapley Additive exPlanations
SLR	Systematic Literature Review
XGBoost	eXtreme Gradient Boosting

Chapter 1

Introduction

1.1 Research Background

The increased accessibility of education has led to a more qualified and skilled workforce, consequently intensifying the competition for job opportunities [51] not only in the private sector but also within the public sector. As the pool of potential candidates grows, individuals often find themselves navigating a complex and competitive job market where standing out becomes crucial. This dynamic is clearly exemplified in Indonesia, where a substantial total of 2,409,882 individuals submitted applications to participate in the civil servant recruitment process conducted in 2023 [61]. Out of this figure, 1,853,617 applicants met the required qualifications and have progressed to the next selection phase to compete for the 572,496 available formations [62]. This phenomenon indicates the enthusiasm of the Indonesian people to pursue a career as a civil servant. However, in contrast to this enthusiasm, during the National Civil Service Coordination Meeting (Rakornas) held in a hybrid format in Batam on Thursday (21/7/2022), Bima Haria Wibisana, the Acting Head of the National Civil Service Agency (BKN), expressed concerns regarding the proficiency and performance of human resources (HR) serving as civil servants in Indonesia, perceiving them to be relatively inadequate [30].

The quest for employment goes beyond mere qualification and education. It extends into the realm of job satisfaction. As individuals pursue their professional aspirations, the realization dawns that securing a job does not inherently guarantee enduring contentment and satisfaction. Increasingly more often, job satisfaction has become an important topic in the public discourse due to its impact on various job-related factors. Job satisfaction demonstrates a positive correlation with motivation, job involvement, organizational citizenship behavior, organizational commitment, life satisfaction, mental health, and job performance [9, 72]. Conversely, it exhibits a negative association with absenteeism, turnover, and perceived stress [31]. Furthermore, it has been observed that enhanced organizational performance is a consequence of dedicated workforces that find satisfaction in various facets of their work within the organization [70].

Understanding and being attuned to one's career trajectories is a pivotal aspect that not only mitigates the risk of job dissatisfaction but also serves as a catalyst for enhancing overall job performance. The importance of matching individuals to appropriate positions for job satisfaction and performance is widely acknowledged, emphasizing the essential requirement to provide employees with timely opportunities for career advancement [16]. From the perspective of an employee, possessing a clear understanding of their potential career trajectory provides a sense of assurance that there exists a structured path for career progression and valuable guidance that will ultimately lead them to success in their professional endeavors. In essence, this foresight becomes a source of motivation, empowering employees with the confidence to navigate their careers with purpose and a strategic vision for future achievements. On the other hand, from the employer's viewpoint, understanding the potential career paths of an employee serves as a valuable tool in the strategic management of human resources. Such knowledge plays a pivotal role in reducing the turnover rate and facilitating optimal Person-Job Fit within the organizational framework. Thus, guiding individuals through the next stages of their careers is beneficial for both employees and their employers.

The issue of limited understanding and access to career path information is also prevalent among civil servants in Indonesia. Many individuals lack a comprehensive understanding of their career paths, including the necessary steps required to pursue specific trajectories and the associated requirements and qualifications needed for advancement. The primary cause of this issue is the limited access to pertinent and detailed information, which is essential for effective career planning and development. As a result, many civil servants find themselves unable to navigate their career patterns effectively, leading to potential stagnation and a lack of progression within their roles. This deficiency in understanding and information access hampers their ability to make informed decisions about their career trajectories, thereby limiting their opportunities for professional growth. In contrast, a thorough recognition and understanding of career patterns are anticipated to play a crucial role in stimulating heightened competency and improving job performance among civil servants. Enhanced career awareness would empower them to align their efforts with the necessary steps for advancement, ultimately fostering a more dynamic and competent workforce. Consequently, addressing this informational shortfall is essential for optimizing the career development processes and enhancing overall job performance within the civil service sector in Indonesia.

In recent years, career path detection systems vitally leveraged machine learning methods as in other fields. These systems serve the purpose of assisting individuals in making well-informed decisions about their careers by forecasting potential career paths based on a range of attributes [71]. Although closely connected to job recommendations, career path prediction doesn't propose particular job advertisements to candidates. Instead, its goal is to anticipate the subsequent role an individual might undertake in their career journey [15]. Understanding career trajectories offers numerous advantages, for both individuals working in the public sector and the organizations employing them. Therefore, empirical research into predicting career paths prediction is considered to have noteworthy implications for areas such as human resource management, talent development, and organizational planning [40, 45].

1.2 Problem Context

Despite the significant enthusiasm among Indonesians to pursue careers as civil servants, as evidenced by the substantial number of applicants for civil servant positions, there remains a critical issue regarding the career development and job satisfaction of those who succeed in securing these roles. The existing career management systems within the Indonesian civil service framework are insufficient, leading to a pervasive lack of understanding among civil servants about their career paths. This deficiency includes limited knowledge of the necessary steps for career progression, the qualifications required for advancement, and the overall trajectory of their professional growth. This problem is exacerbated by inadequate access to detailed and pertinent information, which is essential for effective career planning and development.

As a result, many civil servants are unable to navigate their career patterns effectively, leading to potential stagnation and a lack of progression within their roles. This situation not only hampers their personal and professional growth but also adversely affects their job satisfaction and overall performance. Consequently, the inefficiency in career development processes poses a significant challenge to optimizing human resource management within the Indonesian civil service sector. Addressing this issue is crucial for fostering a more competent and dynamic workforce, thereby enhancing job performance and satisfaction among civil servants. In government organizations, the implementation of career path detection systems, utilizing machine learning methods, could provide a solution by assisting civil servants in making well-informed decisions about their careers, thus enabling a structured and strategic approach to career advancement. While pursuing an ML-based solution direction could be attractive to government organizations, engineering ML-based systems, however, is far from trivial. In fact, it is problematic, from multiple standpoints.

Unlike the systems from the past designed to support various business processes, including Human Resources processes (and also including career development as part of the Human Resources area), software engineering for ML-based systems is recognized to be challenging in ways that the systems from the past never were [4, 21, 22]. One of the key challenges is the uncertainty and variability of the requirements for this type of system. In fact, two recent surveys with industry practitioners [4, 6] revealed RE to be the most difficult activity for the development of ML-based systems. In particular, the 2024 empirical study of Alves et al. with 200+ practitioners from multiple countries on the challenges of implementing ML-based systems concludes: “the problem understanding and requirements phase, is perceived as the most relevant and complex part of the ML life cycle”. If requirements for an ML system do not or can not provide a solid foundation for the system implementation, the organization adopting such a system runs the risk of software implementation that is not meaningful to its users. The next sub-sections indicate why ML systems pose unique software engineering (SE) challenges and, in particular, requirements engineering (RE) challenges and what these challenges are. As this thesis is dedicated to the implementation of a new ML-based system for career path prediction, understanding what is known in prior publications about these challenges from the perspective of RE, has some implications for this research.

1.2.1 ML-based systems from RE perspectives

In the past 40 years, the disciplines of RE – and of SE in general, have accumulated substantial knowledge on the development of business applications that are meant to support the various processes in organizations (e.g. in areas such as Human Resources, Purchasing, Accounting, just to name a few) [13]. Among other software engineering processes, the processes for engineering the requirements for this type of systems, have been extensively investigated by scholars in both Computer Science and Information Systems. The community of researchers of these systems refers to them as sub-systems or modules of enterprise-wide solutions known as “Enterprise Resource Planning (ERP) systems” [12]. Over the years, RE scholars have come up with many proposals for methods, most of which are empirically evaluated and adopted in practice, to elicit, specify, prioritize, validate, and negotiate the requirements for ERP systems. Knowledge about the functional requirements for ERP systems has been codified in reusable reference models (for example, business process models, entity-relationship models, and use case diagrams) and provided by ERP vendors and their consulting partners to client organizations, in order to support RE analysts in their tasks. Unlike these systems – well studied for a few decades, ML systems are a relatively recent phenomenon. Because of their recency, there is no accumulated reusable knowledge, no 'proven' methods and tools, nor there are any standards regarding RE for this type of system. What is more, RE scholars consider these systems not as 'just another technology' but tend to agree that ML-based systems indicate a paradigm shift in light of which RE methods need to be rethought and re-defined. Many researchers indicate that these systems pose some important unique challenges that put in question the very assumptions behind the established RE methods and processes for ERP systems and, in turn, bring doubt in the effectiveness of the RE methods known from the past. As Amershli et al.(2019) indicate, the development process for ML-based systems is characterized by its data-centricity, non-linearity, and multiple feedback loops between stages, each of which adds up to a new level of complexity and uncertainty. Unlike the 'conventional' ERP applications (as those discussed in the publications of e.g. Rolland et al. (2005), Daneva, and Wieringa, 2010), ML-based systems learn from data instead of being programmed with predefined rules. This is to say that a system's behavior no longer emerges from a set of manually coded rules; in contrast, most ML approaches generate rules based on a set of examples (training data) and a specific fitness function [68]. The high dependency on data and the non-deterministic nature of these systems are two properties that did not exist for 'conventional' ERP applications and that complicate a number of RE tasks. As we will see below, the literature review on RE for ML-based systems indicates that the requirements for ML systems are no longer merely a result of an analysis of the stakeholder's needs and desires (as in 'conventional' systems) but also from the analysis of the available data and what could possibly this data be useful for. In turn, the quality of the available data might even become a factor limiting the stakeholders in their user requirements. Furthermore, in the context of ML-based systems development projects, scholars indicate [43] that while RE analysts can define and specify high-level requirements explicitly, the low-level requirements are defined implicitly through the dataset, making the requirements traceability inapplicable. What RE specialists currently do to develop an awareness of the risk due to missing traceability and to partly mitigate this risk, is to understand the descriptive characteristics of the dataset and document them properly. According to Soeren Lauesen (Lauesen, 2001, a RE textbook author, low-level requirements (also known as 'design level requirements' in his textbook) treat specific design choices (in this case, the choice of ML algorithms) and are very important for new (greenfield) systems if one wants to have a full understanding of what the system does and how it does it. However, in the case of ML systems, in order to set specific designs, the RE specialists must first know the range of possibilities (i.e. the applicable ML algorithms) and comparatively evaluate which one performs better in what context and in regard to what prediction task. While such comparative evaluations sound a logical thing to do, and indeed evaluation processes have been proposed by the ML community of practitioners, there are no standards on how to do such evaluations, nor knowledge of the step-by-step approaches that could fit better certain contexts than others. In the next section, we present the characteristics of the RE process that is part of the problem context treated in this thesis. We derive these characteristics based on what so far is known from systematic literature reviews (SLRs) and systematic mapping studies (SMSs) on the topic of ML-based systems from the RE perspective and from the SE perspective. We note that in recent years, a number of such reviews have been published in an effort to understand the challenges these systems pose.

1.2.2 Unique characteristics of RE for ML-based systems based on published literature studies

The research of this thesis has been informed by six recently published literature studies: three SLRs [21, 22, 42] and three SMSs [1, 43, 67]. The SLR of Lwatakare et al. identified 23 challenges in the context of engineering large-scale ML-based systems in industrial settings and motivated that ML-based systems require new methods to overcome these challenges. The authors categorized them as pertaining to four quality attributes, namely, adaptability, scalability, safety, and privacy. The analysis of Lwatakare et al. refers to ML workflow, i.e. data acquisition, training, evaluation, and deployment. A few of the reported challenges, which, in our opinion, concern RE, address (i) the analysis of stakeholders' problems and application domain aspects, and (ii) the uncertainties in defining desired outcome specifications in the experimental stage of ML model development. Next, the SMS of Martinez-Fernandez et al. mapped out the state-of-the-art SE for artificial intelligence-based systems. These authors also identified a prevalence of studies on testing. Regarding RE, while these authors pointed to a few papers treating specific RE-tasks for ML-based systems, they found only one paper (Vogelzag and Borg, 2019) that treated a complete RE process and that was illustrated by using a case study in a real-world organization. The authors indicate that the topic of software requirements is one of the underrepresented areas with great potential for further research. Furthermore, the SLR of Giray (2022) aimed at providing state-of-the-art SE research for engineering ML systems. This review listed a number of challenges referring to the main software life cycle stages. Similarly to Martinez-Fernandez et al. (2022), Giray indicated that most of the research on ML systems in the SE community has focused on testing. Moreover, Giray's analysis found that none of the SE for ML proposals had a mature set of tools and techniques. Of particular relevance for the research in this thesis are the five challenges that the author found in regard to RE. We therefore list them as follows: (1) challenges related to managing users' expectations; e.g. users should be made aware in timely fashion of the fact that ML systems might be imperfect, yet beneficial solutions which could be improved over an extended period of operation time of operation; (2) challenges related to requirements elicitation; e.g. RE-analysts should be prepared for a more intense experimentation and exploration throughout the RE process, compared to the cases of RE 'conventional' systems; (3) requirements specification; e.g. describing ML-systems relies on adopting some specific quantitative measures (such as accuracy, precision, recall, F measure) which stakeholders (users, managers of users and even RE-analysists) might not be able to connect to goal level requirements or domain level requirements (these are the requirements addressing the goal of the system from stakeholders' perspective and the user tasks in the application domain which the system will support); (4) ML-based systems introduce new quality requirements, such as freshness, fairness and explainability for which there are very few proposals on how to do RE for, and whatever proposals are there, there has been very context-specific; and (5) ML-based systems puts in the foreground new types of requirements, those to data quality, as it directly affects the performance of an ML model; data quality requirements have rarely been central to RE in the past. We found three literature studies [1, 22, 67] dedicated specifically to the topic of RE for ML-based systems. The SMS of Villamizar et al. mapped out the landscape of publications on this topic. The RE challenges that these authors found overlap with those of the SLR of Giray (specifically in regard to (1), (2), and (5)). Based on their results, Villamizar et al. hypothesize that defining those requirements reflecting what the ML-enabled system is expected to achieve is challenging, likely due to the abstract nature of translating business problems into ML tasks. These authors also conclude that the overall challenge for RE for ML-based systems is the lack of empirical evaluation as this impedes the accumulation of knowledge of the RE regarding what RE method would work in what context. The 2023 SLR of Ahmad et al. agreed with the finding of Villamizar et al. regarding the lack of empirical evaluation of RE proposals for ML-based systems. In fact, Ahmad et al. observed that 40% of the publications included in their SLR had no evaluation. What is more, according to these authors, existing research on RE for ML covers autonomous systems and cyber-physical systems in sectors such as automotive and self-driving cars. Very little research exists in RE for human-centered AI, i.e. for ML systems embedded in a portfolio of larger information systems such as those in HR, which is the subject of this thesis. Moreover, Ahmad et al. call for more research on a reference map that could capture the key components and attributes needed when specifying AI system requirements. The authors call for more research towards creating a framework to include RE for human-centered AI. Finally, the 2023 SLR of Gjorgjevikj et al. looked into the question of whether 'conventional' RE activities are still relevant to the ML-based systems development process, what challenges these systems pose to RE, and what new quality requirements are to be considered, and what necessary adjustments to are needed in 'conventional' RE practices to better fit into this process. In addition to the SLR, the authors also used a case study to find some possible adjustments in current RE practices that could possibly work in a real-world case. The challenges these authors indicate are similar to those already found by other authors [6, 21]. The adjustments proposed are to address six types of quality requirements: interpretability, fairness, robustness, safety, security, and privacy. Our review of these previously published SLRs and SMSs indicates an agreement in scholars' understanding that ML introduces new technical and ethical challenges of which software project stakeholders must be fully aware "even before the project begins" [22]. These previously published sources allow us to derive the following characteristic aspects of RE for ML-based systems:

(i) from stakeholders' perspective, there are new stakeholders involved in the RE process: data scientists and legal specialists [22].

(ii) the feasibility of the stakeholders' requirements depends on the available data [1,6,21]

(iii) RE is necessarily exploratory in nature with plenty of experiments to expect along the way, as the design-level requirements for a particular system can be known only after comparative evaluation of ML algorithms based on the available data and chosen prediction task [1, 21, 28].

(iv) related to the previous point, comparative evaluation itself becomes in fact an integral part of the exploratory RE process [1, 6].

(v) an ML model can be considered as a requirements specification based on training data since the data can be considered as a learned description of how the ML model shall behave [28].

(vi) new quality requirements, such as freshness and fairness need to be included [21, 22].

(vii) operationalizations of those quality aspects characterizing "the goodness" of ML algorithms to match a human task is in terms of metrics and has so far been technical in nature and needs "translation" into the terms of stakeholders' goals [21, 42, 67]

1.3 Research Goal and Specific Research Objectives

The overall goal of this thesis is to develop a career path prediction framework for civil servants in Southeast Asia. The initial hypothesis underlying this research posits that applying machine learning algorithms to career path prediction can provide civil servants with a clearer understanding of their potential career trajectories, thereby enhancing job satisfaction and boosting their work performance. To achieve this primary research goal, the study will focus on two specific objectives: the first refers to developing a deeper understanding of the application of ML technology in the particular domain of interest, career path prediction; whereas, the second objective refers to designing, implementing, and validating a comprehensive framework that integrates the most appropriate methods, features, and requirements to accurately predict the next career path and ensure its practical applicability. We elaborate on these objectives as follows:

Firstly, through a thorough literature review and gap analysis, the thesis aims to identify existing gaps and research opportunities in the application of machine learning for career path prediction. By examining a wide range of academic articles and research papers, this comprehensive review will provide a solid foundation for understanding the current state of research and highlight areas where further investigation is needed.

Secondly, the study will focus on the design, implementation, and validation of a comprehensive career path prediction framework for civil servants. This process will involve identifying and specifying the key requirements, selecting appropriate methods, and ensuring that the framework is both accurate and practically applicable. Conducted within the scope of an exploratory requirements engineering (RE) process (as described in the previous section), this approach ensures that the resulting system is robust, adaptable, and well-aligned with the specific needs of civil servants. By achieving these objectives, the research aims to contribute valuable insights and practical tools to enhance career development processes within the civil service sector.

For the purpose of this research, in order to illustrate the Southeast Asian context, we will focus on one specific country, namely Indonesia. We will focus use this country as an example. However, it is the understanding of the author that the proposed solution should be possible to be useful also in other countries that have more or less similar public administration structure to Indonesia and are located in Southeast Asia.

1.4 Research Question

To effectively address the research goal, a main research question has been developed.

Main Research Question How to predict the next career path of employees in a public organization in Southeast Asia?

In order to reach a well-grounded answer to the main research question, the following sub-questions have been formulated:

Sub-Research Question

Sub-RQ1: What are the fundamental motivations and factors driving the research to predict career paths?

Sub-RQ2: What type of dataset is utilized, and which specific attributes within the dataset are employed to predict the career paths?

Sub-RQ3: What are the Machine Learning methods and techniques utilized for predicting career paths?

Sub-RQ4: What models/frameworks are proposed in the literature in career path prediction?

Sub-RQ5: What functional and non-functional requirements need to be satisfied by the proposed framework?

Sub-RQ6: What framework is suitable for implementing the career path prediction of employees in a public organization in Southeast Asia?

Sub-RQ7: Which machine learning algorithm demonstrates the best predictive performance in determining the next career path of employees in a public organization in Southeast Asia?

Sub-RQ8: Which features contribute the most to the predicting results?

Sub-RQ9: To what extent is the proposed framework considered useful and usable by practitioners in the field?

1.5 Research Process

The research process is divided into two main stages: Preliminary Research and Main Research. This division ensures a comprehensive understanding of the theoretical background and practical application.

The Preliminary Research phase involves a systematic literature review to understand the current state of knowledge in the field of career path prediction and machine learning applications in human resource management. This phase aims to answer the sub-research questions 1-4. The insights gained from the literature review inform the selection of algorithms and the overall direction for the main research phase.

The Main Research phase will begin with interviewing key stakeholders to gather qualitative insights that inform the research design and model requirements. This will be followed by collecting relevant data, including civil servant records, to build a comprehensive dataset. The data will then be processed to ensure it is suitable for machine learning models. Various machine-learning algorithms will be trained and evaluated using performance metrics to identify the most effective model. The selected model will then be integrated into a prototype application providing an API for predictions. The final steps involve summarizing findings, drawing conclusions, and offering recommendations based on the research results. The main research will answer the last four sub-research questions.

The overview of the complete research is presented in Figure 1.1.

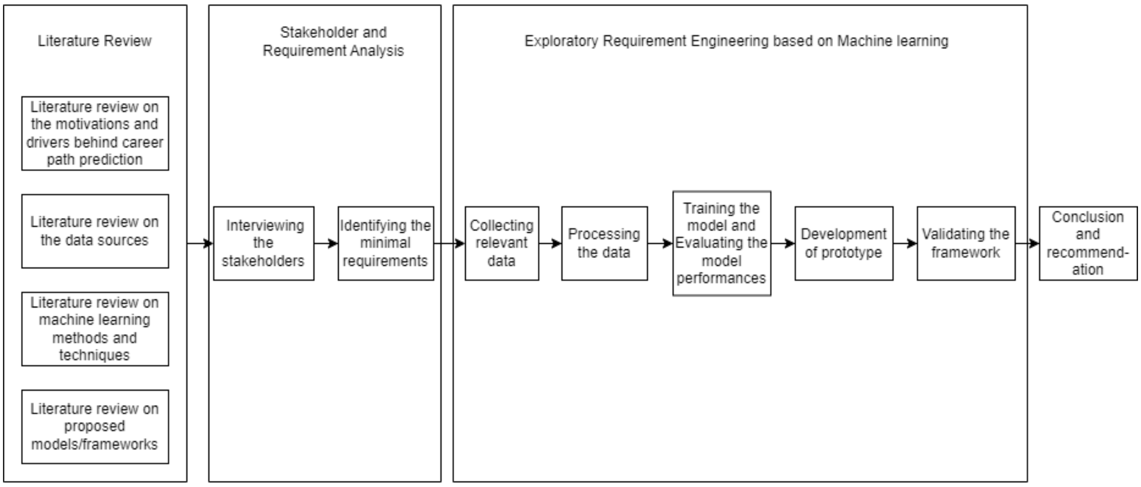

Literature Review
Literature review on the motivations and drivers behind career path prediction
Literature review on the data sources
Literature review on machine learning methods and techniques
Literature review on proposed models/frameworks
Stakeholder and Requirement Analysis
Interviewing the stakeholders
Identifying the minimal requirements
Exploratory Requirement Engineering based on Machine learning
Collecting relevant data
Processing the data
Training the model and Evaluating the model performances
Development of prototype
Validating the framework
Conclusion and recommend-ation

Figure 1.1: Research Process

1.6 Knowledge Areas Required for this Research

Throughout the research process in Figure 1.1, to achieve the goal of this research, the right knowledge has to be collected. For the purpose of this work, the following types of knowledge are deemed instrumental:

1. knowledge about requirements engineering (RE) for ML-based systems;
2. knowledge about the stakeholders
3. knowledge about the data
4. knowledge about ML models
5. knowledge about the application domain, in this case, career path prediction
6. methodological knowledge

These areas of knowledge are presented in Figure 1.2 and are described as follows. The first type includes knowledge about RE methods, processes, and techniques for specific RE tasks. The second type of knowledge is about stakeholders, their needs, goals and requirements, and the techniques for stakeholder analysis. The third type is knowledge about the data available for the research project and the feasibility of using this data for the purpose of experimentation by means of ML models. The fourth type is knowledge that concerns the ML models, their underlying assumptions, and the evaluation criteria for ML algorithms. The fifth type of knowledge is about the application domain itself and refers to the work processes of the practitioners (i.e. the Human Resource Manager in government agencies as well as the employees) who will embed the system in their day-to-day work life within the organizations. The sixth type is knowledge about methodologically sound processes to be used to develop the framework for career path prediction and to plan and execute the comparative empirical evaluation of a set of ML models using the dataset available for this project. Below, we elaborate on each knowledge type.

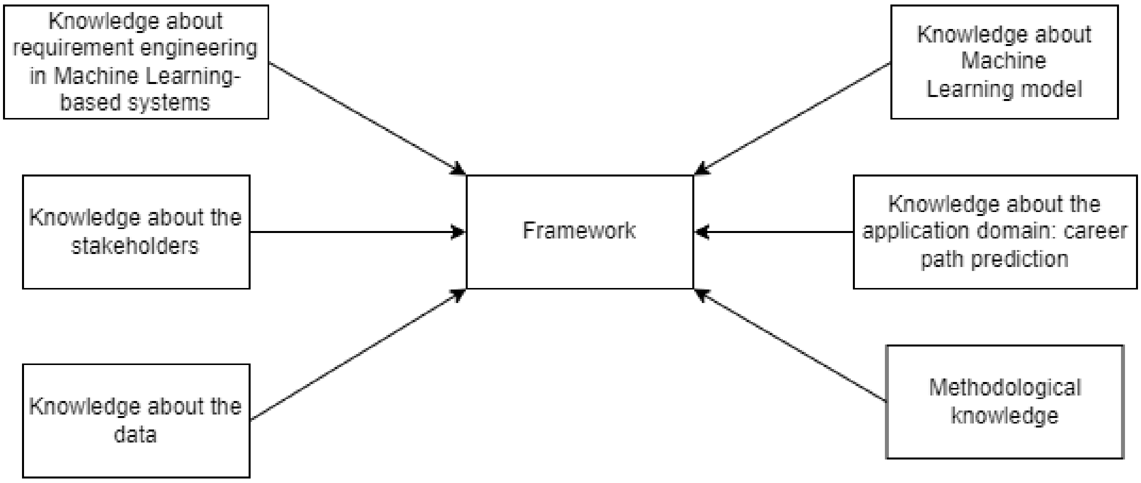

Knowledge about requirement engineering in Machine Learning-based systems
Knowledge about Machine Learning model
Knowledge about the stakeholders
Framework
Knowledge about the application domain: career path prediction
Knowledge about the data
Methodological knowledge

Figure 1.2: Areas of knowledge required in this research

1.6.1 Knowledge about Requirements Engineering for ML-based Systems

Requirements Engineering (RE) is the early phase in the software development process that focuses on understanding and defining what a system should do and how. It involves several activities and processes aimed at eliciting, analyzing, specifying, and managing the requirements of a system. For the completion of this research, we use the following sources of knowledge: (1) Alexander's book on stakeholders' analysis [2] and the RE textbook of Soeren Lauessen (2001) [36] which provide a broad range of RE techniques that a researcher or a RE specialist can choose from based on his/her context; (2) the existing SLRs and SMSs of other authors; these are those discussed in Section 1.2.2 of this chapter.

1.6.2 Knowledge about the stakeholder

Stakeholder analysis is a crucial process that involves identifying, understanding, and managing the expectations and influence of individuals or groups who have a vested interest in the project's outcome [2, 36]. For the purpose of this thesis, as part of the research, stakeholder knowledge will be collected by applying stakeholder analysis techniques in a real-world context, namely in an organization in the public sector in Southeast Asia. As per the recommendation of Lauesen (2001), interview-based techniques will be applied to discover, elicit, and understand the needs and wishes of representatives of various stakeholder groups relevant to this research project.

1.6.3 Knowledge about the data

This type of knowledge concerns the understanding of the available data in the organization which would be used for the purpose of this research. It includes knowledge about the quality of data, its relevance, and usefulness, which helps to determine those features important for the employed ML model and those that could be ignored. As part of feature selection, one needs to know what factors affect the career path prediction, which data fields are needed to predict the career progression available, and how the data is structured.

1.6.4 Knowledge about ML models

This type of knowledge encompasses both theoretical and practical knowledge. It concerns the different types of ML (supervised, unsupervised, and reinforcement learning) and when to use each, the key algorithms, and the metrics for evaluating model performance, such as accuracy, precision, recall, and F1 score, to name a few. For the purpose of this work, an SLR was carried out in order to understand the type of ML approaches employed in our area of interest, career path prediction. This SLR informed the selection of ML algorithms to be employed in this master thesis research. Specifically, we chose to use 4 ML algorithms: Decision Tree [48], Random Forest [8], XgBoost [11], and Multilayer Perceptron [53] will be used.

1.6.5 Knowledge about the application domain (career path prediction)

This knowledge concerns the real-world organizational processes, key information entities (i.e. data entities, attributes, relationships), and organizational roles (employees, functional managers, human resources managers) of the environment in which the ML-based system is considered to be implemented. This knowledge is acquired through the personal work experience of the author of this thesis in the organization in Southeaster Asia that is interested in the present research. It also is complemented by knowledge collected through the stakeholders' interviews which will be done for stakeholder analysis. The knowledge is important for understanding the decision-making process and the decision-making norms that the ML-supported career path prediction should have to comply with.

1.6.6 Methodological knowledge

This type of knowledge guides the creation of the framework in this thesis. It concerns the application of methodologies, processes, and established practices for the systematic execution of research tasks. This thesis uses the CRISP-DM methodology for guiding the work in research projects that include a data mining component. Drawing upon this methodology, an experimental set-up was designed which assured our comparative evaluation of ML algorithms was done in a systematic and disciplined way and could support the exploratory RE for the ML-based system for career path prediction.

1.7 Thesis Structure

This thesis consists of nine chapters and is structured as follows: Chapter 1 outlines the background and motivation for the study. Chapter 2 conducts a Systematic Literature Review to address sub-research questions 1-4, synthesizing existing theoretical knowledge on relevant topics and identifying research gaps. Following this, Chapter 3 details the data mining methodology employed in the study, emphasizing the theories underlying the machine learning algorithms essential to its execution. Next, Chapter 4 delves into business and data understanding, illustrates the proposed framework, and specifies the minimum functional and non-functional requirements for the solution based on the business understanding. Chapter 5 explains the design aspects of the study. Chapter 6 explains the implementation of each algorithm using real-world datasets (NCSA's data), details the feature selection process, and provides a comparative analysis of their performance. Chapter 7 discusses the validation of the results. Chapter 8 then provides the answers to sub-research questions 5-9, critically examines the study's limitations, and suggests directions for future research. Lastly, Chapter 9 concludes the study and highlights its academic and practical contributions. Figure 1.3 offers a visual overview of the research structure and indicates the specific sub-research questions addressed in each chapter.

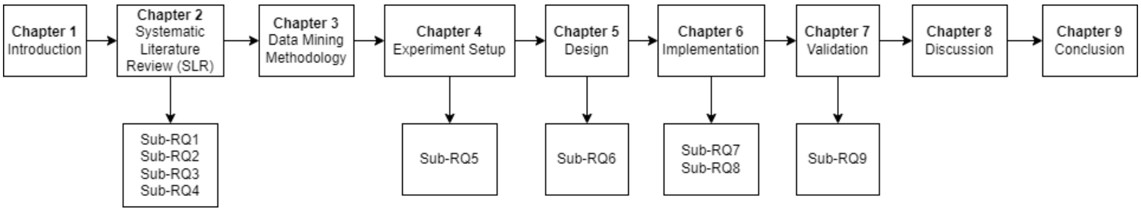


FIGURE 1.3: Thesis Structure

Chapter 2

Systematic Literature Review

In this study, a systematic literature review (SLR) was conducted based on the comprehensive guidelines proposed by Kitchenham and Charters [33]. Their framework stands as a beacon of methodological excellence in the realm of research synthesis, particularly within the domain of software engineering and allied fields. Embracing their systematic approach ensures not only the credibility but also the comprehensiveness of the review process, assuring that no pertinent stone is left unturned in the pursuit of knowledge. The SLR methodology provided a structured approach to identify, evaluate, and synthesize relevant research studies in the field of machine learning applications for career path prediction. Initially, a detailed search strategy was formulated, specifying the databases to be searched, the search terms to be used, and the inclusion and exclusion criteria for selecting relevant studies. This was followed by a meticulous screening process where titles, abstracts, and full texts of the identified articles were reviewed against the predefined criteria. Data extraction was then carried out, where key information from the selected studies was systematically recorded, focusing on research gaps, methodologies, findings, and limitations. The extracted data was subsequently analyzed and synthesized to provide a comprehensive overview of the current state of research, identify trends, and highlight areas requiring further investigation. By adhering to the rigorous guidelines of Kitchenham and Charters, the SLR ensured a high level of reliability and validity in the findings, thereby laying a solid foundation for the subsequent phases of the research.

2.1 Planning the Review

2.1.1 Scientific Databases and Search Query Formulation

In this research, two scientific databases are selected for the retrieval of relevant articles, namely Scopus[1] and IEEE[2]. Scopus was chosen due to its reputation for having the largest collection of scholarly publications from various fields, including peer-reviewed journals and conference papers, making it an ideal choice for a comprehensive exploration of literature relevant to the research topic. Moreover, the inclusion of the IEEE is based on its specialized focus on computer science, data science, and closely related disciplines. Leveraging the rich resources offered by IEEE ensures that the research is anchored in authoritative sources specific to the fields directly aligned with machine learning and prediction techniques, thereby enhancing the depth and relevance of the literature review.

[1] https://www.scopus.com/search/form.uri?display=advanced

[2] https://ieeexplore.ieee.org/search/advanced

The formulation of the search query is a crucial step in the systematic literature review, designed to fetch pertinent literature addressing both the main and sub-research questions. This search query consists of keywords and terminologies meticulously chosen to closely align with the research goals. To augment the thoroughness of the research, five primary groups of queries have been devised, each comprising the main query and its pertinent synonym. Table 2.1 provides a detailed compilation of all keywords associated with the main query.

TABLE 2.1: Search Query Keywords

Career	**Machine Learning**	**Path**	**Prediction**	**Algorithm**
Career	"Machine Learning"	Path	Prediction	Algorithm
	"Artificial Intelligence"	Trajectory	Advisory	Model
	"Data Mining"	Planned	Guidance	Technique
	"Neural Network"	Prospect	Suggestion	Methods
	"Fuzzy Logic"			"Decision Tree"
	Transformer			

Based on the keywords presented by Table 2.1, the search queries were formulated. Each keyword and its synonym were connected using the 'OR' logical operator, while each keyword group was connected using the 'AND' logical operator. The queries were customized to meet the format specifications of each scientific database, concentrating on particular sections such as title, abstract, and keywords. The queries utilized on the aforementioned databases are presented below:

Scopus (advanced):
TITLE-ABS-KEY (
(career) AND
("machine learning" OR "artificial intelligence" OR "data mining" OR "neural network" OR "fuzzy logic" OR transformer OR "Long Short-Term Memory") AND
(path OR trajectory OR planned OR prospect) AND
(prediction OR advisory OR guidance OR suggestion) AND
(algorithm OR model OR technique OR methods OR "decision tree")

IEEE:
((career)
AND
("machine learning" OR "artificial intelligence" OR "data mining" OR "neural network" OR "fuzzy logic" OR transformer OR "Long Short-Term Memory")
AND
(path OR trajectory OR planned OR prospect)
AND
(prediction OR advisory OR guidance OR suggestion)
AND
(algorithm OR model OR technique OR methods OR "decision tree"))

2.1.2 Inclusion and Exclusion Criteria

Following the exploration of Scopus and IEEE, specific inclusion and exclusion criteria were established to guide the selection of papers. This step is crucial not only for maintaining research quality and honing the collection but also for streamlining the subsequent data extraction process. It guarantees that only relevant or the most relevant studies are included and given due consideration. The inclusion and exclusion criteria defined in this study are outlined in Table 2.2.

Table 2.2: Search Query Keywords

Inclusion Criteria	**Exclusion Criteria**
Literature was written in English and peer-reviewed	Duplicate literature
Literature was published in conference proceedings and research journals in Computer Science, Mathematics, Engineering, and Decision Science areas	Irrelevant literature based on the title, abstract, and keyword to the main and sub-research questions of this study
Literature was published in the last 10 years	Full-text unavailability
Literature offered one or more answers to at least one of the research questions	

Referring to Table 2.2, the selected literature must be composed in English to ensure clarity for the research team. Additionally, it should undergo international peer review to uphold the standard of this study. Furthermore, the collected literature is required to be published in the form of research journals or conference proceedings, which are accepted platforms for academic discourse. Then, to avoid expanding the collection with materials irrelevant to the study's focus, the literature-gathering process was constrained by specific disciplinary boundaries, including Computer Science, Mathematics, Engineering, and Decision Science

In addition to the inclusion criteria, the exclusion criteria were applied in order to eliminate redundant and unavailable literature. Duplicate studies were identified based on the title and removed from the results obtained from the selected databases after applying inclusion criteria. Later, to enhance the efficiency of the review process and concentrate on relevant research, a manual evaluation was performed to pinpoint and eliminate literature considered irrelevant based on its abstract. Moreover, it is possible that a study may not be openly accessible to the public; consequently, such studies will not be taken into consideration.

2.2 Conducting the Review

2.2.1 Selection

The process of selecting literature in this study involved several stages. It began with implementing the formulated search query in the chosen scientific databases. Recognizing the potential inclusion of irrelevant studies in the search results, it was necessary to apply the inclusion and exclusion criteria specified in Section 2.1.2 to the title, abstract, and keywords of the literature. This step is crucial not only for maintaining the quality of the research but also for optimizing the literature collection procedure. Furthermore, this stage ensures the inclusion and consideration of only the most relevant studies. Guided by the inclusion and exclusion criteria, it resulted in the selection of a total of 26 articles. A more detailed illustration of the selection process is presented in Figure 2.1.

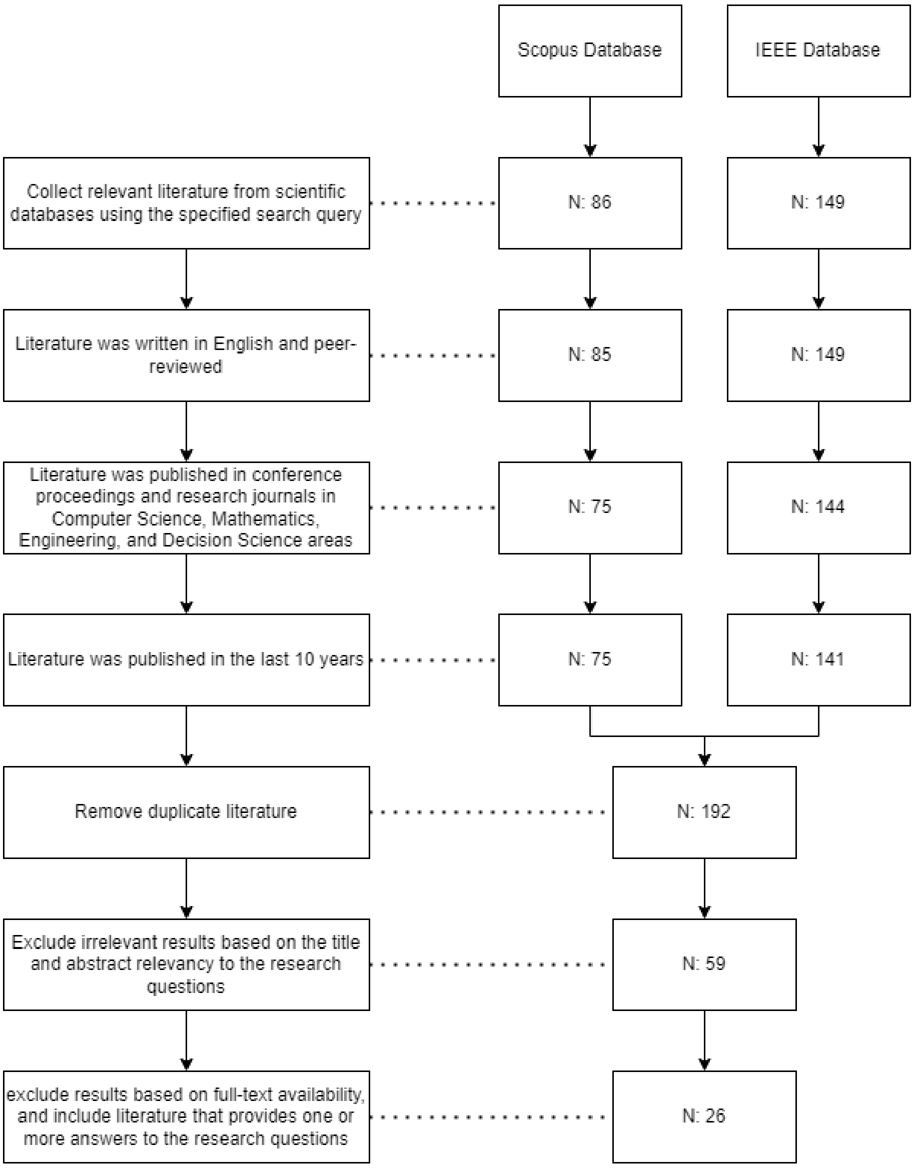

Scopus Database
IEEE Database
Collect relevant literature from scientific databases using the specified search query
N: 86
N: 149
Literature was written in English and peer-reviewed
N: 85
N: 149
Literature was published in conference proceedings and research journals in Computer Science, Mathematics, Engineering, and Decision Science areas
N: 75
N: 144
Literature was published in the last 10 years
N: 75
N: 141
Remove duplicate literature
N: 192
Exclude irrelevant results based on the title and abstract relevancy to the research questions
N: 59
exclude results based on full-text availability, and include literature that provides one or more answers to the research questions
N: 26

Figure 2.1: Literature Selection Process

2.2.2 Data Extraction

After the literature selection process, the data extraction process was conducted by reading the 26 pieces of literature. In this study, the extracted data will be synthesized into quantitative and qualitative analyses to answer and address the research question as described in sub-section ??. This process is crucial for determining the consistency or inconsistency of results across studies [33], thereby contributing to a more nuanced understanding of the research landscape.

As illustrated in Table D.1 in the appendix, the quantitative analysis covers exploration in four categories: Machine Learning (ML), Source of Data and Feature (SF), Data Type (DT), and Implementation (IM). The ML category is employed to determine the incorporation of machine learning methods or techniques in the literature that addresses the third sub-research question. The SF and DT categories distinguish the type of dataset and specify the attributes used in the literature, answering the second sub-research question. Finally, the IM category assesses whether the literature focuses on real-case implementations rather than solely theoretical or analytical discussions, answering the fourth sub-research question.

The qualitative analysis conducted in this SLR entails a comprehensive exploration of the literature. In Table A.2 in the appendix, a detailed description of the quality assessment is provided, encompassing purposes that answer the first sub-research question, outcomes, limitations, challenges, and recommendations discussed within the reviewed literature, addressing the fifth research question.

2.3 Result

2.3.1 Demographic Information About the Included 26 Papers

In examining the country-based trends of literature concerning career path prediction, valuable insights into the global distribution of research contributions in this field emerge. The findings, depicted in Figure 2.2, underscore the significance of geographical context in shaping research agendas and highlight the diverse perspectives and approaches driving innovation and advancement in the study of career path prediction.

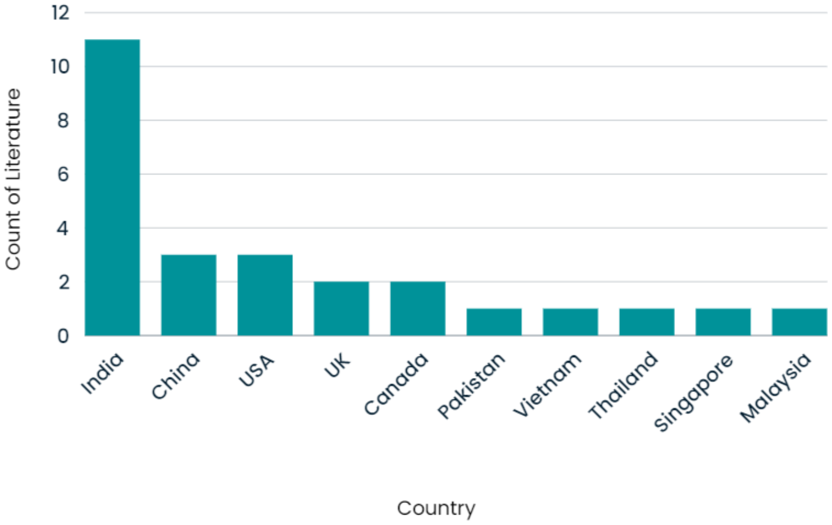


Figure 2.2: Country-based Trends of the Reviewed Literature

Among the 26 selected papers, India stands out with 11 papers dedicated to this subject [5, 24, 34, 35, 49, 52, 56, 60, 65, 66, 71], indicating a significant focus on career trajectory forecasting within its academic discourse. Conversely, China [25, 26, 45] and the USA [20,39,47] follow with 3 papers each, reflecting a comparatively lower but still notable level of scholarly attention. The UK [27,57] and Canada [38,59] contribute 2 papers each, suggesting a moderate presence in this area of study. Notably, Pakistan [29], Vietnam [63], Thailand [58], Singapore [40], and Malaysia [50] each offer 1 paper, highlighting a sparse representation in the literature on career path prediction. Overall, the demographic distribution of the included papers underscores the significance of international collaboration in advancing research efforts aimed at leveraging machine learning techniques for career path forecasting.

2.3.2 Motivation and Factors Driving Career Path Prediction (RQ1)

The first sub-research question (RQ1) intricately examines the motivations and underlying drivers that inspire researchers to undertake studies in the domain of career path prediction. Through this exploration, several factors contributing to the motivation for Career Path Prediction were identified.

To begin, having a profound grasp of one's career trajectory is essential in securing a job that is both suitable and fulfilling [5, 25, 34, 35, 58, 60, 71], laying the foundation for future success [5, 20, 35]. By comprehending the entirety of their career trajectory, individuals can evaluate their qualities [52, 56], concentrate on acquiring essential skills [20, 34, 35], and steer clear of potential distractions [29]. This level of comprehension empowers them to make well-informed decisions regarding their career paths, thereby optimizing the benefits derived from their professional pursuits [24, 26, 49]. Furthermore, it is imperative for employers and recruitment entities to provide comprehensive career development support. This proactive measure not only contributes to the overall satisfaction and development of their workforce but also plays a crucial role in retaining highly skilled and capable employees [38]. This guidance becomes instrumental in pinpointing opportune moments for promotions and salary increases, ensuring an efficient and strategic approach to talent management within the organization [40, 45, 63].

Secondly, machine learning holds significant potential in developing advanced systems for predicting career paths [27, 49, 52, 59, 71], leveraging vast amounts of data to analyze patterns and trends. These systems have the capacity to offer personalized insights, guiding individuals toward suitable career choices based on their skills, preferences, and the evolving job market. Despite this potential, the current body of research on automated career prediction systems remains relatively scarce [5, 27, 29]. More comprehensive studies and exploration in this field are essential to harness the full capabilities of machine learning and ensure the development of accurate and effective tools that can benefit individuals in navigating their career trajectories. As technology continues to advance, increasing research in this area could contribute to the refinement and widespread adoption of machine learning applications in career planning and guidance.

2.3.3 Type of Dataset and Features Utilized in Career Path Prediction (RQ2)

Among the reviewed studies, a discernible trend emerges in the choice of datasets, with 14 papers opting for student datasets [5, 24, 29, 34, 35, 49, 50, 52, 56–58, 60, 65, 66], and 15 papers directing their focus toward employee datasets [5, 20, 25–27, 38–40, 45, 47, 56, 59, 60, 63, 71]. This deliberate selection provides a nuanced exploration of career paths across different demographic segments. The breadth and depth of research findings are accentuated by the strategic utilization of varied data collection methods, as visually depicted in Figure 2.3. Notably, prominent online platforms such as About.Me[3], LinkedIn[4], Indeed[5], Facebook[6], and Twitter[7] emerged as the most commonly used sources for collecting workforce data [20, 39, 40, 47, 56, 59, 63]. Additionally, a subset of studies delves into the realm of competition data, harnessing platforms like DataCastle[8] [25, 26] and Kaggle[9] [71] to extract valuable

[3] https://about.me/

[4] https://www.linkedin.com/

[5] https://indeed.com/

[6] https://www.facebook.com/

[7] https://www.twitter.com/

[8] https://www.dclab.run

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

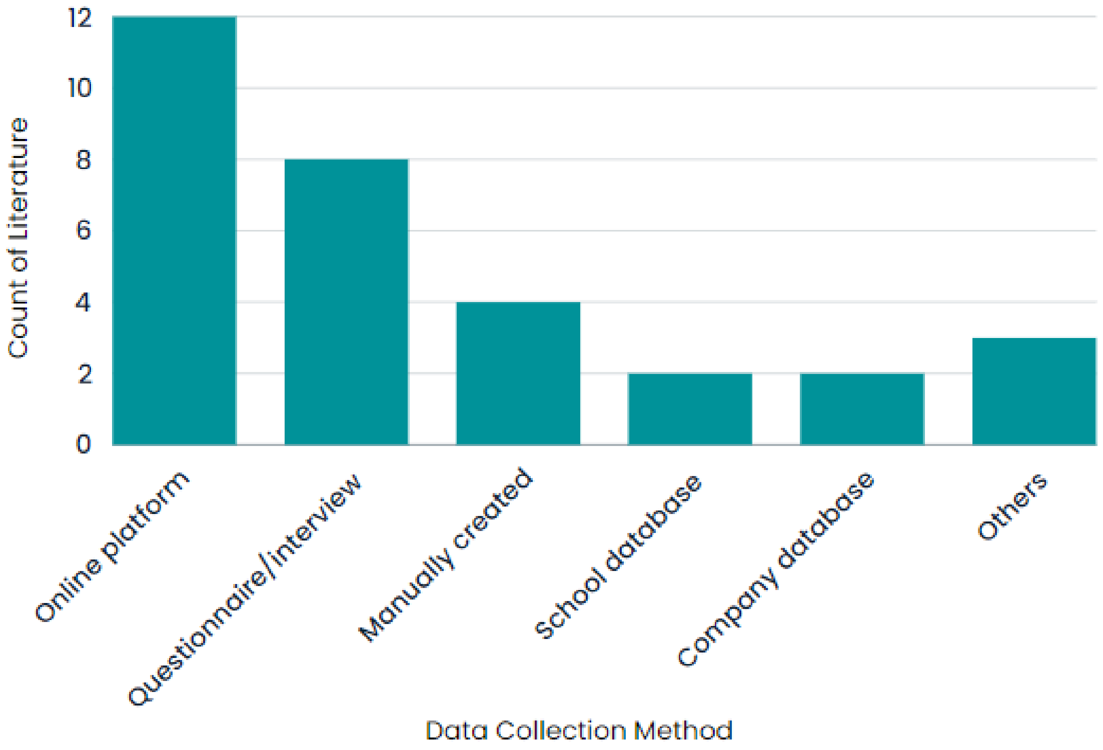


FIGURE 2.3: Data Collection Method

insights. We expect all papers to describe how they collected the data and the source of the data, but some only include the methodologies, while others [5, 24] mention the source of the data. The diversity in data collection methodologies is further underscored by the incorporation of questionnaires [27, 29, 35, 49, 52, 60, 65], interviews [50, 65], and data requisition from official agencies such as the Higher Education Statistics Agency (HESA) [57], schools/colleges [56,58], and companies of the employee [38,56]. Some papers also reused data that was utilized in other studies [47, 71].

The importance of effective feature selection in addressing the challenges posed by the curse of dimensionality is clearly noticeable in the examined papers [23]. A number of features have been meticulously employed to illuminate the predictive landscape in career research. Standout among these features are work experience and skills, meticulously analyzed to gauge professional expertise and illuminate trajectories [20, 25–27, 29, 38, 45, 47, 50, 59, 60, 63, 65, 66, 71]. The academic terrain is robustly explored through features such as educational background and academic performance, offering insights into qualifications and potential career trajectories [20, 24, 29, 34, 40, 52, 56, 58, 60]. Furthermore, a paradigm shift is observed as the spotlight turns towards individual characteristics, with personality traits and behavioral patterns emerging as indispensable features. This evolving interest underscores a deepening inquiry into the intricate interplay of personal attributes in shaping career choices and trajectories [5, 29, 47, 50, 65]. Beyond the traditional metrics, features like career aspirations and interests are also given due consideration, adding layers of complexity to the predictive models employed in unraveling the intricate tapestry of career paths [29, 35, 40, 47, 49, 52].

As for the data type, the papers did not clearly mention explicitly whether the dataset is structured, unstructured, or semi-structured. However, based on the description of the source of the database and the data collection method, we could safely assume that 16 papers used structured data [24–26, 29, 34, 38, 47, 49, 50, 52, 57, 58, 60, 65, 66, 71], 3 papers utilized unstructured data [5, 40, 45], and 7 papers combined structured and unstructured data [20, 27, 35, 39, 56, 59, 63].

2.3.4 Machine Learning Methods and Techniques in Career Path Prediction (RQ3)

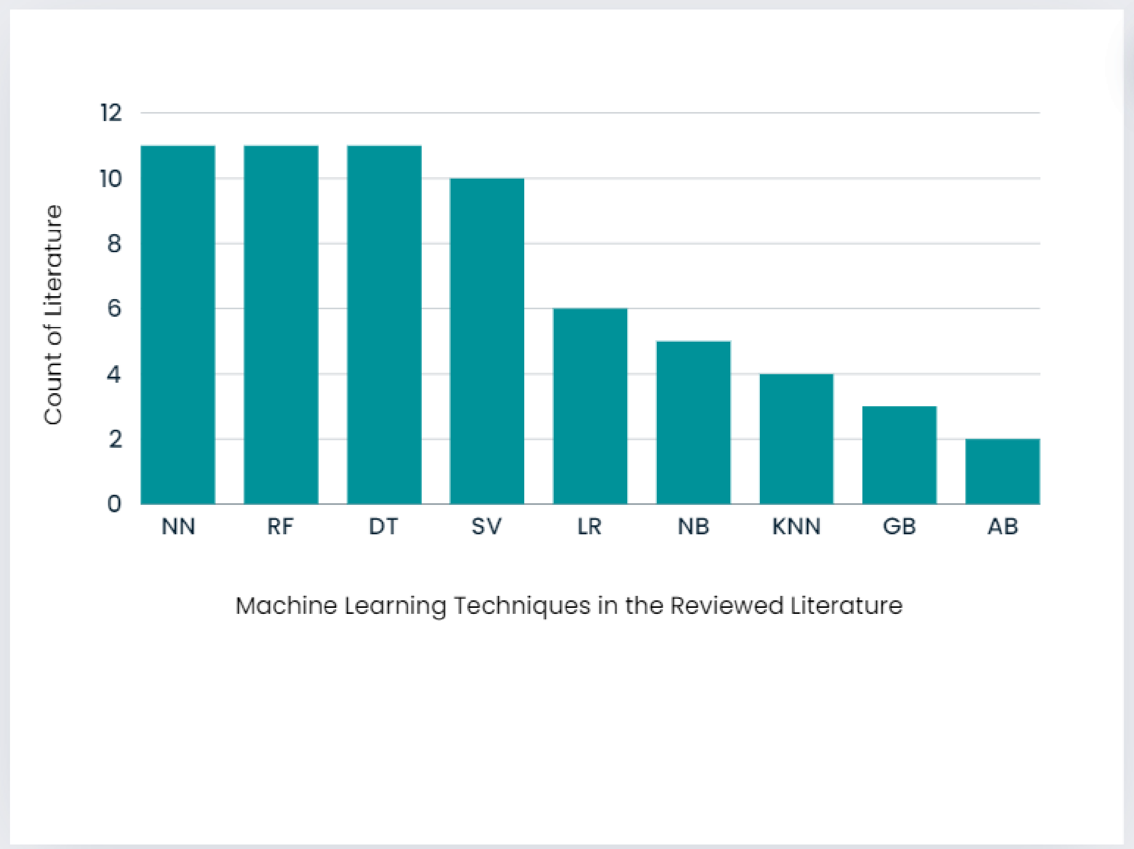


FIGURE 2.4: Implemented Machine Learning Techniques in the Reviewed Literature

In the expansive realm of research dedicated to predicting career paths, a diverse array of machine learning techniques[10] has been utilized to unravel the complexities of individuals' professional trajectories. Notably, tree-based models such as Random Forest (RF) [5, 24, 25, 29, 34, 35, 45, 52, 57, 63, 71], Decision Tree (DT) [5, 24, 25, 34, 35, 45, 49, 52, 56, 57, 60], and Gradient Boosting (GB) [25, 29, 34] have risen to prominence as the most frequently employed, garnering attention in 14 scholarly works, followed by the utilization of deep learning methods, exemplified by Neural Network (NN) [5, 20, 25, 26, 35, 39, 45, 59, 60, 63, 71] in 11 studies. Support Vector (SV) [5, 24, 34, 40, 47, 52, 56, 60, 66, 71] follows closely with a total of 10 papers. Moreover, a comprehensive suite of models, including Logistic Regression (LR) [26, 45, 57, 60, 63, 71], Naive Bayes [57, 59, 60, 66, 71], K-Nearest Neighbour (KNN) [49, 60, 66, 71], along with the application of AdaBOOST [52, 60], has collectively

[10] RF = Random Forest; DT = Decision Tree; SV = Support Vector; NN = Neural Network; LR = Logistic Regression; NB = Naive Bayes; KNN = K-Nearest Neighbour; GB = Gradient Boosting; AB = AdaBoost

played pivotal roles in confronting the multifaceted challenges inherent in predicting career paths.

2.3.5 Proposed Models or Frameworks in Career Path Prediction (RQ4)

In the comprehensive review of 26 papers dedicated to career path prediction, a notable subset comprising 16 papers [5,20,24–27,29,34,35,39,45,50,63,65,66,71] distinguishes itself through the integration of real-case implementations. In the execution of the proposed systems, neural network models have assumed a crucial role, featuring prominently in a collective total of 9 papers. These papers exemplify the diverse applications of various neural network architectures, including Convolutional Neural Networks (CNN) [5, 25, 26], Fully Connected Neural Networks (Fully Connected NN) [20], Long Short-Term Memory networks (LSTM) [26,35,39,45], Recurrent Neural Networks (RNN) [35], Artificial Neural Networks (ANN) [71], and Multi-Layer Perceptrons (MLP) [63, 71].

Delving into the nuances of model utilization, these papers not only underscore the prevalence of neural networks in career path prediction but also highlight the specific architectures employed to address distinct challenges. For example, convolutional Neural Networks, for instance, excel in capturing spatial dependencies, while Long Short-Term Memory Networks prove adept at modeling sequential patterns [18].

Furthermore, although not exclusively machine learning methods, multi-task learning and fuzzy logic techniques have also been utilized in the implementation of career path prediction systems. These techniques involve the simultaneous learning of multiple related tasks to enhance generalization performance [38, 40]. Fuzzy logic, which enables the representation of linguistic variables and fuzzy sets and is suitable for handling uncertainty in career paths, has also been integrated into the career path prediction system [50, 65].

While the proposed models and frameworks mark significant advancements in career path prediction, they also introduce specific challenges and limitations that need addressing. One of the most common issues is the increased computational demands, particularly with complex models such as those involving Convolutional Neural Networks (CNNs) models [5]. This approach, while effective in capturing complex patterns, can significantly escalate computational requirements, potentially hindering the scalability of the prediction system. As a result, there is a need for optimization strategies or alternative model architectures to mitigate computational demands without compromising predictive accuracy.

Another limitation arises from the specificity of the implemented systems, often confined to a singular field or domain [5]. For instance, some studies focus exclusively on engineering students, limiting the generalizability of the predictive models. A more comprehensive approach that encompasses diverse fields and professions is essential for developing universally applicable career path prediction systems.

Data bias is a recurring challenge in the reviewed papers, exemplified by the inclusion of data from high-skilled users with proficient digital skills [20]. This bias may lead to skewed predictions that do not accurately reflect the broader workforce. Addressing this bias involves incorporating a more diverse dataset that encompasses individuals with varying skill levels, ensuring a more representative sample for robust predictions. The dependence on user-provided data introduces another challenge, as individuals may neglect to complete certain data fields [27, 47]. Incomplete data can compromise the accuracy of predictions, emphasizing the importance of developing strategies to handle missing data effectively or encouraging more comprehensive data submission.

Limited access to datasets poses a significant constraint, impacting the system's accuracy [29]. The scarcity of diverse and extensive datasets restricts the model's ability to capture the nuances of different career paths comprehensively. Efforts to enhance dataset accessibility and diversity are crucial for improving the reliability and applicability of career path predictions.

Moreover, the dynamic nature of career paths and the evolving nature of job markets challenge the assumption of static data [39]. In reality, career trajectories are subject to change, and job markets are dynamic. Addressing this challenge involves developing models that can adapt to evolving career landscapes, providing more accurate and timely predictions.

Beyond the features traditionally considered in career path prediction, external factors [45], such as environmental influences and individual historical patterns, also play a crucial role. The reviewed studies often neglect these factors, highlighting the need for a more holistic approach that incorporates a broader range of features to improve the predictive power of career path models. In summary, while the novel models and frameworks present promising avenues for career path prediction, addressing the associated challenges and limitations is crucial for their successful implementation and broader applicability.

2.4 Discussion on the Answer to the Research Questions

2.4.1 Motivation and Factors Driving Career Path Prediction

The reviewed studies reveal a compelling narrative that underscores the importance of cultivating a profound understanding of one's career trajectory, not only for individuals but also for employers. This understanding emerges as a pivotal factor contributing significantly to overall satisfaction, professional development, and the efficient management of talent within organizational settings. The nuanced exploration of this theme provides valuable insights into the reciprocal relationship between individual career planning and the role of employers in fostering a conducive work environment.

Additionally, the papers illuminate the transformative potential of machine learning in predicting career paths. It signals the need for a concerted effort to fill the existing research gap and emphasizes the importance of comprehensive studies to unlock the full capabilities of machine learning for the benefit of individuals navigating their professional trajectories. As technology continues to advance, the research conducted in this field could pave the way for the widespread adoption and refinement of machine learning applications, ushering in a new era of personalized and effective career planning and guidance.

2.4.2 Type of Dataset and Features Utilized in Career Path Prediction

Upon reviewing the 26 selected literature, it became apparent that not all papers included comprehensive details regarding the derivation of their datasets. While certain papers delved into the specifics of how their datasets were derived, including the methods employed (interviews or surveys), others chose to focus primarily on identifying and discussing the data sources themselves [5, 24]. Notably, the widespread use of professional networking platforms such as LinkedIn for data collection purposes stood out, which is unsurprising given the availability of authentic work experience data on such platforms. However, in contrast, acquiring data from official agencies may present its own set of challenges, potentially complicating the data collection process. Furthermore, collecting data from multiple sources and employing various data collection methods were observed in some papers [27, 47, 56, 60, 71].

In terms of the features chosen for predictive modeling, our analysis revealed a mixed approach among the papers. While some papers explicitly conducted feature selection as part of their predictive modeling process [57, 65], a significant portion did not provide explicit mention of this aspect. Within the realm of predictive modeling for career path prediction, certain features consistently stood out as the most commonly utilized. Work experience and skills, for instance, emerged as the most frequently employed features in the 26 papers in this SLR. This prevalence underscores the substantial predictive power of these features and their intrinsic relevance when forecasting career trajectories. Additionally, while less prevalent, individual characteristics were also identified as features in a subset of papers. This inclusion suggests a nuanced approach to predictive modeling, acknowledging the multifaceted nature of individual attributes that contribute to career outcomes. The variation in the selection and utilization of features across the reviewed literature reflects the diverse methodological approaches and research objectives pursued by different studies in their quest to address the complex dynamics of career prediction.

2.4.3 Machine Learning Methods and Techniques in Career Path Prediction

This rich tapestry of machine learning approaches, woven into the fabric of these research endeavors, underscores a shared commitment within the academic community. This commitment extends to the exploration and exploitation of a diverse set of techniques for predicting career paths. The collaborative effort enriches the collective understanding of the intricate factors influencing career trajectories, highlighting the significance of employing a variety of models. The nuanced dynamics of professional development are thus captured comprehensively, emphasizing the multifaceted nature of the endeavor to unravel the complexities within the field.

The findings from the SLR indicate that neural networks, random forest, decision trees, and support vector machines (SVM) have emerged as the most prominent and extensively researched machine learning techniques in the field of career path prediction. These techniques have gained considerable attention and adoption due to their robust performance and flexibility across various datasets and predictive tasks. Neural networks, renowned for their ability to capture intricate relationships and patterns within data, have been extensively explored in forecasting career trajectories using a diverse range of input features such as skills, education, and work experience. Similarly, random forests and decision trees are valued for their interpretability and ease of implementation, making them popular choices for analyzing and predicting career paths by identifying crucial decision points and feature interactions. Additionally, support vector machines have been recognized for their capability to handle high-dimensional data and nonlinear relationships, thereby serving as valuable assets in career path prediction tasks where data may display complex patterns.

It is worth noting that in some instances, certain papers have chosen to utilize specific models not as primary contributors to the proposed systems but rather as comparative tools. These models serve as benchmarks for evaluating the efficacy and performance of the novel systems presented in the research, adding a layer of robustness to the comparative analyses.

2.4.4 Proposed Models or Frameworks in Career Path Prediction

After examining the 26 papers, it becomes evident that the majority of them delve into illustrating the framework in a technical manner [5, 20, 24–27, 39, 45, 63, 65, 66], akin to describing its architecture. However, there seems to be a gap in presenting the logical progression of steps involved in developing the framework. What appears to be lacking is a comprehensive flowchart or structured outline detailing the entire process of framework development [29, 34, 50, 71]. This omission could hinder the clarity and accessibility of the research, making it challenging for readers to grasp the methodology behind the framework effectively.

It is worth noting that among the reviewed 26 papers, there is a subset that primarily focuses on comparing existing machine-learning models. These papers aim to identify which models perform better in predicting career paths without necessarily contributing a novel framework of their own. While comparative studies are valuable for benchmarking and understanding the performance of different models, they may not significantly advance the field in terms of introducing innovative methodologies or frameworks. As such, there seems to be a disparity between papers that offer technical insights into framework architecture and those that concentrate solely on model comparison without introducing novel contributions to the field.

While these models/frameworks represent significant advancements in the field of career path prediction, they also bring to light various challenges and limitations that could hinder their broader application. Primarily, the computational expense associated with fine-tuning and optimizing the model architecture presents a significant hurdle. This is particularly pronounced in complex models requiring extensive training data and computational resources. Moreover, the scalability of these models is often limited, necessitating careful consideration of resource allocation and optimization techniques to ensure efficient processing.

In addition to computational costs, the specificity of the implemented system poses another challenge. Many models are tailored to specific tasks or domains, which can limit their applicability to broader contexts. This restricts the generalizability of findings and may necessitate additional adaptation or customization when applied to different scenarios or datasets. Moreover, advancing feature selection methodologies is crucial in effectively capturing relevant predictors for career path prediction. This involves identifying and incorporating informative features while minimizing noise and irrelevant variables, thus improving the model's predictive accuracy and interpretability.

Furthermore, data bias and limited access to diverse datasets emerge as recurring challenges in model development and evaluation. Biases present in training data can significantly impact model performance and generalization, leading to skewed predictions and unreliable outcomes. We think that this finding is unsurprising, as fairness in ML models has been a central research topic in the ML scientific community (e.g. [10]). Additionally, the limited availability of diverse datasets hampers the robustness and inclusivity of model training, hindering its ability to effectively capture the complexities of real-world scenarios. Addressing these challenges requires a concerted effort to enhance dataset diversity and mitigate biases through comprehensive data collection and preprocessing techniques.

In summary, the challenges surrounding computational cost optimization, system specificity, data bias, and feature selection underscore the need for continued research and innovation in the field of computational modeling. By addressing these limitations through interdisciplinary collaboration and methodological advancements, researchers can develop more robust and inclusive models for predicting career paths and informing decision-making processes in various domains.

Chapter 3

Exploratory Methodology

This chapter outlines the exploratory methodology employed in this thesis, guided by the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework. CRISP-DM provides a comprehensive and widely adopted methodology for structuring and executing data mining projects, ensuring alignment with business objectives throughout the project's lifecycle. This robust and flexible framework supports various phases of data mining projects, making it adaptable to different industries and problem domains. In this project, CRISP-DM facilitated a systematic approach to developing the career path prediction model for civil servants in Southeast Asia. Additionally, this chapter provides an in-depth explanation of the four machine learning algorithms utilized in the study: Decision Tree (DT), Random Forest (RF), XGBoost, and Multilayer Perceptron (MLP). These algorithms were selected based on their effectiveness in similar applications, with their respective strengths and implementations thoroughly analyzed to determine the most suitable model for accurate career path prediction.

3.1 CRoss-Industry Standard Process for Data Mining (CRISP-DM)

The primary method used in this project was the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, a comprehensive and widely adopted methodology for guiding data mining projects. CRISP-DM was adopted because it serves as an effective framework for structuring, arranging, and executing data science (or machine learning) projects, ensuring that every step is systematically addressed and that the project remains aligned with the business objectives throughout its lifecycle [3].

CRISP-DM provides a robust and flexible structure that supports various phases of data mining projects, making it adaptable to different industries and problem domains, which is essential to tailor the methodology to meet specific project requirements while maintaining a coherent and consistent approach. Following the CRISP-DM framework, the significant stages of this project as shown in Figure 3.1 are outlined below:

3.1.1 Business Understanding

The Business Understanding phase is the first and arguably the most critical step in the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, laying the foundation for the entire project. This phase involves comprehensively understanding the project's objectives to ensure alignment with the strategic goals of the sector. The initial step in this phase is determining the business objectives. This involves identifying

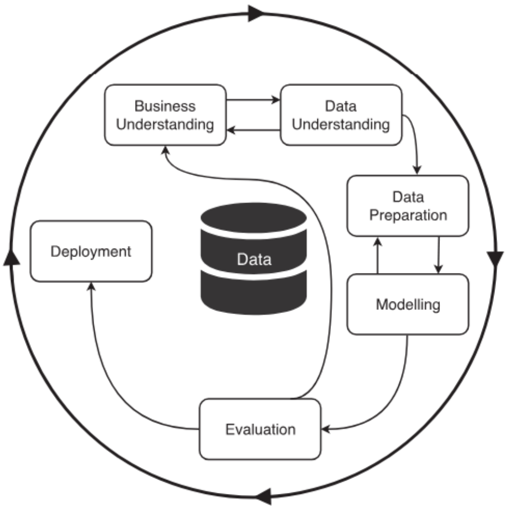


FIGURE 3.1: The CRISP-DM process model of data mining [44]

and engaging with stakeholders to gather their insights, expectations, and specific needs related to the project. Stakeholders can include executives, managers, analysts, and end-users who have a vested interest in the project's outcomes. Through meetings, interviews, and surveys, stakeholders provide valuable input on what they hope to achieve, which helps in defining clear and actionable objectives.

In addition to understanding the business objectives, this stage also involves gathering information on resource availability. This includes assessing the personnel, budget, and technological resources that are at the project's disposal. Understanding the availability and constraints of these resources is crucial for realistic planning and execution.

Detailed requirements need to be collected to ensure that all aspects of the project are addressed. This involves specifying the data needs, desired outcomes, timelines, and any other pertinent details that will guide the project's development. These requirements help in creating a roadmap for the project, ensuring that all necessary steps are taken to meet the objectives.

Moreover, understanding any constraints or limitations is critical for defining success criteria. Constraints can include budget limitations, technological restrictions, data accessibility issues, and time constraints. Identifying these constraints early on allows for better risk management and contingency planning. For example, ensuring data privacy and compliance with regulations is a critical consideration in many projects. Addressing these considerations from the beginning ensures that the project can proceed without legal or ethical issues.

3.1.2 Data Understanding

Building on the Business Understanding phase, the Data Understanding phase in the CRISP-DM framework is crucial for accomplishing the project objectives. The focus of this phase is to identify, collect, and analyze the necessary datasets to ensure that the data collected is suitable for the analysis and modeling required to meet the project goals.

The initial step in this phase involves collecting the relevant data from various sources. These sources can include databases, spreadsheets, external APIs, or even manual entries. It's essential to gather all pertinent data to ensure a comprehensive analysis. Once the data is collected, it needs to be integrated into the chosen analysis tool. This might involve importing data into a data warehouse, a data lake, or directly into analysis software like Python, R, or specialized data analytics platforms.

Following data collection and integration, the data is then examined to document its basic properties. This includes identifying the data format (e.g., CSV, JSON, SQL), counting the number of records (rows), and listing the field identities (columns). This initial examination helps to understand the scope and structure of the data, providing a foundation for more detailed analysis.

Finally, the quality of the data is assessed to ensure its reliability for analysis. This involves checking for missing values, duplicates, and inconsistencies. Data cleanliness issues such as incorrect formats, outliers, and irrelevant data points are documented. Addressing these issues is essential as they can significantly impact the accuracy and validity of the analysis and subsequent modeling. This quality assessment often leads to data cleaning tasks such as correcting errors, filling in missing values, and removing duplicates to prepare the data for robust and accurate analysis.

3.1.3 Data Preparation

The Data Preparation phase, often synonymous with feature engineering, in the CRISP-DM framework, is an essential step that transforms raw data into a format suitable for analysis and modeling. This phase involves meticulous steps to ensure the data is clean, consistent, and ready for predictive modeling. Unlike the initial Data Understanding phase, which aims to familiarize and explore the dataset, data preparation focuses on refining and enhancing the data for optimal use in predictive models.

The initial step in the Data Preparation phase is selecting the data from various sources. This involves determining which datasets might be relevant to the project and documenting the reasons for their inclusion or exclusion. This selection process is crucial because it ensures that only pertinent data is used, which can significantly affect the quality of the analysis and the accuracy of the predictive models.

Once the relevant data is selected, the next step is data cleaning. This step is often the most time-consuming but is critical to avoid the 'garbage-in, garbage-out' problem, where poor-quality data leads to unreliable models. Data cleaning involves several sub-tasks: correcting errors, imputing missing values, or removing erroneous data.

Following data cleaning, data construction is performed. This step involves deriving new attributes from the existing dataset that will be helpful for the analysis. This step often involves domain knowledge to create features that better represent the underlying patterns in the data and improve the model's predictive power.

Then, data integration took place, where information from the selected sources is combined to create a comprehensive dataset. This step includes merging/combining data from different tables or sources into a single, cohesive dataset and ensuring consistency in formatting and coding standards.

After integration, data transformation is carried out to facilitate consistent processing and analysis. It involves normalization and scaling to ensure that no single feature dominates the model due to its scale and encoding categorical variables into a numerical format that can be used by machine learning algorithms.

3.1.4 Modeling

The Modeling phase in the CRISP-DM framework is a critical step that involves selecting and applying various modeling techniques to the prepared dataset. This phase requires careful consideration of different algorithms, hyperparameter tuning, and iterative testing to develop the most accurate and reliable predictive models.

The first step in the modeling phase was to select appropriate machine learning algorithms by carefully reviewing the available algorithms to determine which ones were most suitable for the prediction task at hand. The selection process was guided by several factors, including the nature of the data, the complexity of the relationships between features, and the specific requirements of the prediction task. By exploring the different algorithms, the goal was to identify which model or combination of models would best capture the patterns in the data and provide accurate and reliable predictions.

Once the models were selected, the next step was to perform hyperparameter tuning to optimize their performance. Hyperparameters control the behavior of the training process and the structure of the model. This involved adjusting the settings of each algorithm to find the combination that yielded the best results by utilizing search methods such as Grid Search, Random Search, or Bayesian Optimization.

After tuning the hyperparameters, the dataset is split into a training set and a test set. The model is then trained using the training data. During this training process, the models learned the underlying patterns and relationships between the features and the target variable. The test set was then used to evaluate the model's performance. To ensure that the evaluation process is robust and not overly dependent on a specific train-test split, a cross-validation technique like k-fold cross-validation or Stratified Cross-Validation is applied. During cross-validation, the dataset was split into multiple folds, and the model was trained and evaluated on each fold. This process provided multiple estimates of the model's performance, allowing for a more reliable assessment.

3.1.5 Evaluation

The Evaluation phase in the CRISP-DM framework is a crucial step where the performance of the predictive models is thoroughly assessed to ensure they meet the project's objectives. This phase involves several detailed tasks aimed at determining whether the models are reliable, accurate, and useful for the intended purpose. To evaluate how well the models perform on the classification task, several key performance metrics are calculated, including accuracy, precision, recall, and F1-score.

Accuracy measures the proportion of correct predictions made by the model out of all predictions. While accuracy provides a general sense of how well the model performs, it can be misleading in cases of imbalanced datasets, where one class is much more frequent than the others. Precision measures the accuracy of positive predictions. It is defined as the number of true positive predictions divided by the total number of positive predictions (true positives plus false positives). High precision indicates that the model has a low false positive rate. Recall measures the model's ability to identify all relevant instances (true positives). It is defined as the number of true positive predictions divided by the total number of actual positives (true positives plus false negatives). High recall indicates that the model has a low false negative rate. The F1-score is the harmonic mean of precision and recall. It provides a balance between the two, especially useful when you need to find an optimal balance between precision and recall. It is particularly valuable in situations where the class distribution is uneven and both false positives and false negatives need to be minimized.

Another critical aspect of the Evaluation phase was comparing the performance of different models. Since multiple algorithms were used it was important to determine which model performed best across the evaluation metrics. Each model's strengths and weaknesses were considered, and the model that offered the best balance of accuracy, precision, recall, and F1 score was selected for deployment.

Furthermore, Feature importance analysis was held during the Evaluation phase to ensure that the most significant features were correctly identified and utilized by the models. This process is critical because the accuracy and reliability of the models depend heavily on the selection of meaningful features. In this thesis, SHAP (SHapley Additive exPlanations) values were employed for this purpose. SHAP values provide a unified measure of feature importance by distributing the prediction among the features in a fair manner, making it easier to interpret how each feature contributes to the model's output. This step ensured that the models were making predictions based on meaningful and relevant features, which is crucial for their interpretability and usefulness in practical applications. By understanding which features influence the model's predictions the most, we can not only improve the model's performance but also gain insights into the underlying patterns and relationships within the data. This is particularly important in fields where interpretability and transparency are essential, such as healthcare, finance, and public policy. Moreover, the use of SHAP values helps validate the model by providing a clear rationale for its predictions, thereby increasing trust and confidence in the model's decisions.

The results of the evaluation were documented comprehensively in Section 6.3. This documentation included detailed reports on the performance metrics and comparisons between different models. Additionally, the rationale for selecting the final model was clearly articulated. This thorough documentation was essential for ensuring transparency and for gaining the trust and support of stakeholders. It also provided a reference for future projects and potential improvements.

3.1.6 Deployment

The Deployment phase in the CRISP-DM framework is the final step where the predictive model, which has been thoroughly developed and evaluated, is put into operational use. This phase involves implementing the model in a real-world environment where it can provide actionable insights to support career planning and progression strategies within the civil service sector.

The first task in the deployment phase is preparing the necessary infrastructure. This involves configuring the required hardware and software, such as setting up servers, databases, and potentially cloud services that will host the model and handle data inputs and outputs. Ensuring that the infrastructure is robust and scalable is crucial for the seamless operation of the model. The model is then integrated with existing human resource management systems and other relevant platforms within the civil service sector to facilitate a smooth flow of data between the model and operational systems.

Once the infrastructure is ready, the model is deployed into the production environment. This process typically involves creating APIs (Application Programming Interfaces) that enable other systems to interact with the model, allowing for predictions based on input data. The API allows the model to be easily accessed and utilized by the existing systems within the organization, ensuring that predictions can be made in real time or as needed by the current processes.

Overall, the Deployment phase is a critical step that ensures the predictive model for civil servants' career paths is operational and delivering value in a real-world setting. By following a structured deployment process, the project can provide reliable and actionable insights, ultimately supporting better career planning and progression strategies within the civil service sector.

3.2 Machine Learning Algorithms

This section explores the theoretical foundations of the machine learning algorithms employed in this study to address the research questions. The selection of these algorithms was carefully guided by the insights from the SLR, as shown in Figure 2.4, which identified the most effective techniques for predicting career trajectories. The chosen algorithms include three tree-based algorithms, namely decision tree, random forest, and XGBoost (as a variant of gradient boosting), and one neural network algorithm - the multilayer perceptron (MLP). Each of these algorithms offers a unique approach to classification and prediction, providing distinct advantages that align with the study's specific objectives. By examining the principles and functionalities of these algorithms, this section aims to provide a comprehensive understanding of the methods used to develop the machine learning models within the research framework. The theoretical foundations discussed here not only justify the algorithm selection but also underscore their relevance to the domain of career path prediction. This foundational knowledge is crucial for understanding how the models were built and why they are well-suited to achieve the study's objectives, ultimately leading to more accurate and insightful predictions of career trajectories for civil servants.

3.2.1 Decision Tree

Decision tree is a widely used machine learning algorithm that is applicable to both classification and regression tasks. It is a powerful and intuitive tool for predictive modeling. It represents decisions and their potential outcomes in a structure that resembles a tree. The key strength of decision trees lies in their interpretability and ease of use, making them an excellent choice for a variety of applications, including career path prediction. A decision tree consists of nodes and branches where the topmost node of a decision tree is called the root node. Decision nodes are the internal nodes where the data is split based on certain conditions or attributes, with each decision node representing a test or a decision on an attribute, resulting in branches that lead to other decision nodes or leaf nodes. Leaf nodes, also known as terminal nodes, represent the final output or decision after traversing through the decision nodes [54]. Figure 3.2 illustrates the visualization of a decision tree. The process begins at the root node, where the dataset is divided into subsets based on the value of a selected attribute. At each node, a splitting criterion is applied to partition

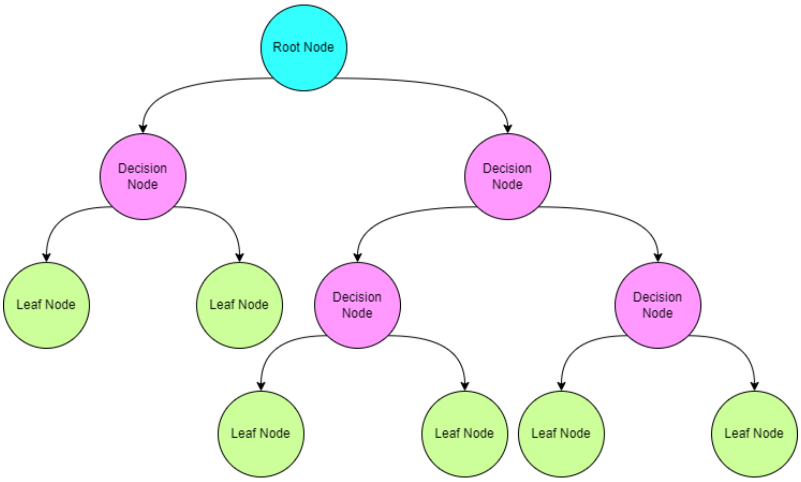


FIGURE 3.2: Visualization of Decision Tree

the data. This criterion selects the feature and threshold that best separates the classes. Common criteria include Gini impurity and information gain where Gini impurity quantifies how much uncertainty is in a node and Information gain quantifies how much that question is reducing error. The formula to calculate the Gini impurity is

$$G = 1 - \sum_{i=1}^{n} p_i^2$$

Where:

- G is the Gini Impurity of the node
- p_i is the proportion of samples belonging to the class i at the node

For calculating the information gain, the formula is given by:

$$IG(T, a) = H(T) - \sum_{v \in \text{Values}(a)} \frac{|T_v|}{|T|} H(T_v)$$

$$H(T) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

Where:

- $IG(T, a)$ is the Information gain
- T is the set of data
- a is the attribute or feature
- $H(T)$ is the entropy of the entire set T
- $Values(a)$ is the set of all possible values that the attribute a can take.
- T_v is the subset of T for which the attribute a has the value v
- $|T_v|$ denotes the number of examples in the subset T_v
- $|T|$ denotes the total number of examples in the original set T
- $H(T_v)$ is the entropy of the subset T_v

The process of splitting continues recursively at each decision node, creating branches and forming new decision nodes or leaf nodes until one of the stopping criteria is met. These criteria could include reaching a maximum tree depth, having a minimum number of samples at a node, or achieving a node purity threshold. The recursion terminates when further splitting is not feasible or does not add significant value. At this point, the leaf nodes provide the final output, which could be a class label for classification tasks or a numeric value for regression tasks.

3.2.2 Random Forest

Although decision trees offer several advantages due to their understandability and interpretability, they also have some disadvantages. Decision trees can become overly complex and overfit the training data, capturing noise instead of the underlying patterns. Ensemble methods like random forest can help mitigate these issues by constructing multiple decision trees during training and aggregating their results to make predictions, thus enhancing the model's robustness and accuracy.

Random forest employs a technique known as bootstrap aggregation or bagging. Bagging involves generating multiple subsets of the training data and combining them into a single aggregated predictor [7]. In classification tasks, this aggregation is achieved by taking a majority vote of the most frequently predicted class (mode) after generating a large number of trees [8]. Mathematically, it can be written as:

$$\hat{y} = \text{mode}\{T_{(b)}(x)\}$$

Where $T_{(b)}(x)$ is the prediction of the b-th tree.

To optimize the performance of a random forest, it is crucial to tune its key parameters. One of the most important parameters is the number of trees in the forest, denoted as **'n_estimators'**. Increasing the number of trees generally improves the model's performance because it allows for a more comprehensive aggregation of predictions, reducing variance and enhancing accuracy. However, a higher number of trees also increases computational cost and training time, so a balance must be struck based on the available resources and the specific requirements of the task. Another critical parameter is the maximum depth of each tree, specified by **'max_depth'**. Limiting the depth of the trees helps prevent overfitting by ensuring that the trees do not become overly complex and capture noise in the training data. Shallower trees might underfit, missing important patterns, while very deep trees might overfit, so it is important to find an optimal depth that balances bias and variance.

Additionally, the parameter **'min_samples_split'** sets the minimum number of samples required to split an internal node. Higher values prevent the model from learning overly specific patterns that may not generalize well to new data. Similarly, **'min_samples_leaf'** determines the minimum number of samples required to be at a leaf node. Increasing this parameter can smooth the model's predictions, making it less sensitive to variations in the training data. Finally, the **'random_state'** parameter is used to ensure the reproducibility of results. By setting a random seed, the same results can be obtained across different runs of the algorithm, which is essential for debugging and comparing models. Tuning these parameters appropriately can significantly impact the performance of a random forest model, making it crucial to understand their roles and how they interact. Proper parameter tuning helps to balance the model's complexity, computational efficiency, and predictive accuracy, leading to better performance on unseen data. —

3.2.3 eXtreme Gradient Boosting

Gradient boosting is another ensemble learning method, distinct yet complementary to bagging methods like random forest. It sequentially builds an ensemble of 'weak learners' or 'base learners', typically classification trees [17], where each new tree attempts to correct the errors of the previous trees thus incrementally improving the model's performance by focusing on the residuals of the prior models, effectively reducing both bias and variance. Mathematically, this can be represented as:

$$F_m(x) = F_{m-1}(x) + \eta h_m(x)$$

Where

- $F_m(x)$ is a new model
- $F_{m-1}(x)$ is the current model
- $\eta h_m(x)$ is the contribution of the m-th model

XGBoost, short for eXtreme Gradient Boosting, is a highly popular and efficient implementation of the gradient boosting framework. Its effectiveness has been well acknowledged across various machine learning and data mining competitions. The key strength of XGBoost is its capability to manage large-scale datasets and intricate feature interactions efficiently [11]. XGBoost employs regularization techniques (L1 and L2) that penalize complex models, thereby enhancing its ability to generalize to new data and prevent overfitting.

$$\text{ObjectiveXGBoost} = \sum_i l(y_i, F_t(x_i)) + \sum_k \Omega(f_k) + \gamma T$$

Where

- $l(y_i, F_t(x_i))$ is the loss function at the t-th iteration
- $\Omega(f_k)$ is the regularization function Ω penalizes the complexity of each tree f_k
- γ is a regularization parameter that controls the complexity of the model
- T is the number of leaves in the tree

3.2.4 Multilayer Perceptron

Artificial Neural Networks (ANNs) are computational models inspired by the human brain's structure and function, designed to recognize patterns and solve complex problems. ANNs consist of interconnected neurons, each generating real-valued activations. Input neurons are activated by environmental sensors, while other neurons are activated through weighted connections from previously active neurons [37]. They excel at handling tasks involving non-linear relationships and large datasets. ANNs operate through forward propagation, passing input data through the network layer by layer. Each neuron computes a weighted sum of its inputs and applies an activation function to introduce non-linearity. This process continues until the final prediction is made. During training, optimal weights are learned through backpropagation, which adjusts the weights to minimize the difference between predicted and actual outputs. Figure 3.3 illustrates a single neuron in an ANN, which is a fundamental unit of the computation in these networks. $X_0, X_1, X_2, .., X_(n-1)$, and X_n are the input features fed into the neuron where each X_i represents a different feature of the data. Each X_i is associated with a weight W_i. These weights are parameters that the network learns during the training process. They determine the importance of each input feature in the computation of the neuron's output. Bias (b) is an additional parameter that allows the activation function to be shifted to the left or the right. It helps the neuron learn the optimal position for activation, improving the model's flexibility. The output of the neuron is represented by y and can be calculated as:

$$y = f(\sum_{i=1}^{n} W_i X_i + b)$$

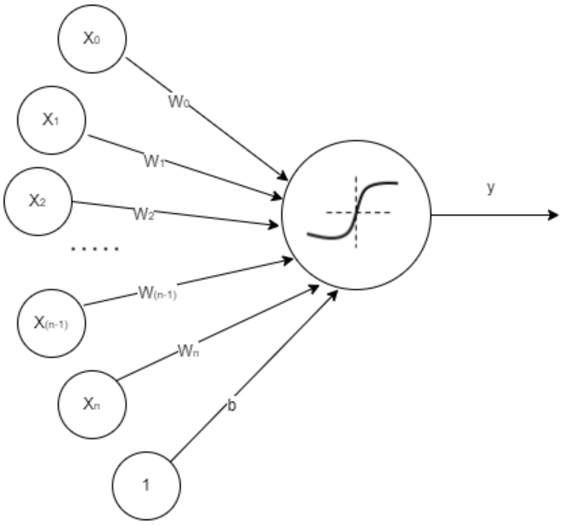


Figure 3.3: A single neuron in an ANN

where f is the activation function.

A Multilayer Perceptron (MLP) is a type of artificial neural network characterized by its architecture of multiple layers of interconnected neurons, enabling it to capture intricate relationships between input features and output classes [71]. MLPs are capable of modeling complex relationships in data, making them suitable for a wide range of applications, including classification tasks [63, 71]. The structure of an MLP includes an input layer, multiple hidden layers, where each neuron applies an activation function to introduce non-linearity, and an output layer that typically uses a softmax activation function for classification tasks to produce a probability distribution over the possible classes. Forward propagation in an MLP involves computing weighted sums of inputs and applying activation functions layer by layer until the final prediction is made.

MLPs are particularly well-suited for predicting career paths due to their ability to handle complex relationships and achieve high accuracy. The flexibility and adaptability of MLPs are also significant advantages. They can automatically learn relevant features from the input data during training, reducing the need for extensive manual feature engineering. This is particularly beneficial when dealing with diverse and complex datasets in career prediction. Additionally, MLPs can process and learn from large-scale datasets efficiently, which is crucial for career path prediction where extensive historical data on employees may be involved. MLPs also can be updated with new data, allowing the model to adapt to changes in career progression trends and patterns. This makes them robust for long-term use in dynamic environments. Furthermore, MLPs provide probabilistic outputs for different career paths, helping HR professionals understand the likelihood of various career transitions and make informed decisions.

Chapter 4

Experiment Setup

This chapter outlines the experimental setup for the study, providing a comprehensive summary of the business and data understanding. Following the data mining methodology described in the previous chapter, this section aims to bridge the gap between theoretical concepts and practical implementation. It includes stakeholder analysis to identify needs and requirements and explains the minimal requirements for the proposed solution, covering both functional and non-functional aspects. The chapter details the essential capabilities and qualities the solution should embody, such as scalability, maintainability, interpretability, and performance. These requirements were derived from interviews with six stakeholders in a real-world government organization in Southeast Asia.

4.1 Business Understanding

In order to gain a comprehensive understanding of the business requirements and ensure effective stakeholder engagement, a stakeholder analysis was conducted. This analysis identified and categorized key stakeholders based on their power and interest, as shown in Figure 4.1, ensuring that the perspectives and needs of all relevant parties were considered in the project.

4.1.1 Stakeholder Analysis

The stakeholder analysis revealed the following key stakeholders in a large government organization in Southeast Asia, specifically in Indonesia:

1. **National Civil Service Agency (NCSA)**: High power and high influence, responsible for regulatory oversight and implementation of civil service policies.

2. **Human Resources (HR) Department of Government Organizations**: High power and high influence, involved in managing civil servant career paths.

3. **Data Scientists**: High power and high influence, responsible for implementing and maintaining the machine learning models.

4. **Programmers**: High power and high influence, involved in the development and integration of the system.

5. **Ministry of Administrative and Bureaucratic Reform**: High power but lower influence, interested in the outcomes and strategic benefits of the project.

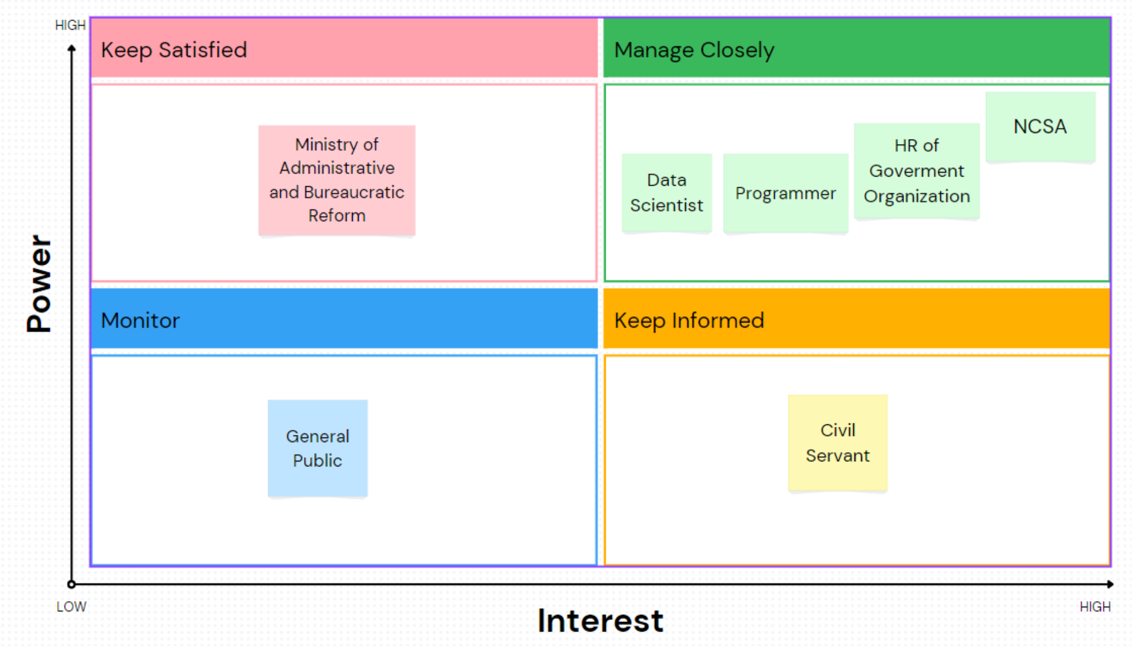


FIGURE 4.1: Stakeholder Analysis

6. **Civil Servants**: Low power but high influence, primary users who will benefit from the career path prediction system.
7. **General Public**: Low power and low influence, indirect beneficiaries interested in improved public administration efficiency.

4.1.2 Requirement Elicitation

To gather qualitative insights and inform the research design and model requirements, a series of interviews were conducted with key stakeholders. The participants included:

1. Deputy of the Civil Servant Information System
2. Director of Development and Enhancement of the Civil Servant Information System
3. Three Senior Software Engineers
4. An Intermediate Data Scientist

These participants were selected to provide a diverse range of perspectives, ensuring a holistic understanding of the current systems, challenges, and future needs. Each interview offered unique insights that were crucial for defining the functional and non-functional requirements of the system.

To develop our interview questions, the guidelines of King and Horrocks [32] were followed. The interview included six types of questions: Importance and Rationale, Current Structure and Trends, Existing Methods and Tools, Challenges and Innovations, Data Collection and Management, and API Integration and Standards. The detailed transcripts of these interviews can be found in the appendix B. The collected interview data was analyzed using the coding techniques of Saldana [55], which allowed the identification of conceptual categories and themes concerning the business requirements as illustrated in Figure 4.2. The outcome of the interviews with the six stakeholders served as the

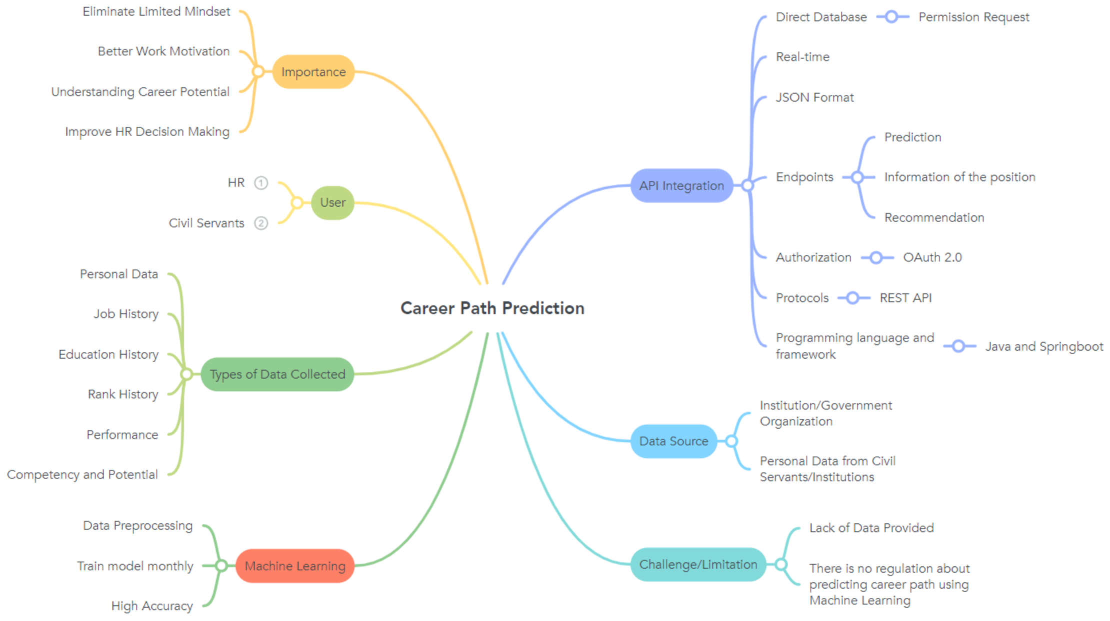


FIGURE 4.2: Mindmap of the Career Path Prediction

foundation to formulate the minimal functional and non-functional requirements. These are presented in Section 4.1.3

From the six interviews, it is evident that predicting career paths holds significant importance. For the organization, it ensures that civil servants are positioned where they can best utilize their skills and talents, enhancing overall performance and efficiency. This process also plays a crucial role in identifying and nurturing future leaders. For the individual civil servant, having a clear career path provides motivation and direction, encouraging continuous personal and professional development.

The current career path structure for civil servants in Indonesia, as described by the interviewees, is influenced by a combination of job function, qualifications, and self-driven development. While there is a framework in place, the alignment between education and job positions is not always straightforward. The system allows for flexibility, enabling civil servants with various educational backgrounds to explore different career opportunities within the organization. However, this flexibility can sometimes lead to misalignments where individuals do not find themselves in roles that fully utilize their potential.

Challenges with traditional methods of career path planning are highlighted through the interviews. These methods are often subjective and susceptible to biases such as nepotism and political influence. Decisions made without comprehensive data can lead to inefficiencies and suboptimal placements. Furthermore, the quality and completeness of historical data are major concerns, as incomplete or inaccurate data hampers effective decision-making.

Machine Learning (ML) offers a promising solution to these challenges, as evidenced by the insights from the interviews. By analyzing large datasets, ML can uncover patterns and provide insights that traditional methods cannot. ML can evaluate competencies, performance, and potential, providing a more accurate and comprehensive view of a civil servant's suitability for various roles. This capability can significantly enhance the accuracy of career path predictions and help identify the best candidates for vacant positions, thereby optimizing the talent management process.

In conclusion, by implementing ML in predicting career paths for civil servants, the system can be revolutionized to be more data-driven, transparent, and efficient. The insights from the interviews indicate that a data-centric approach will ensure that civil servants are placed in roles that align with their skills and aspirations, enhancing performance and job satisfaction. This transition will also mitigate issues related to subjectivity and bias, paving the way for a fairer and more effective talent management system. The deployment of the civil servant career system marks a significant step towards modernizing the career management of civil servants. From the results of the interviews, we can find that there is a strong consensus on the potential benefits of integrating a predictive model into this system. This integration is anticipated to bring about a data-driven approach that will enhance the accuracy and fairness of career path prediction.

There are three types of positions in the civil servant domain in Indonesia: structural/ managerial, functional, and administration. Administration positions are positions with lower position classes and, thus, will not be predicted in this thesis. Since civil servants have an equal chance to choose between functional or managerial career paths, two machine learning models will be developed to predict each of these position types.

4.1.3 Minimal Requirements

In this section, functional and non-functional requirements were determined, detailed, and compiled based on insights from the Systematic Literature Review (SLR) and qualitative interviews and discussions with key stakeholders to ensure they accurately reflect the needs and expectations of the project, as well as ensure their feasibility.

Functional Requirements

The functional requirements describe the behaviors (functions or services) that the proposed solution should encompass. These requirements are fundamental for attaining the desired functionality and performance of the system. The subsequent functional requirements are defined for the implementation of the proposed solution:

- FR1: The solution must be capable of collecting relevant data and include preprocessing mechanisms to clean, transform, and format the collected data for further analysis and model training.

- FR2: The solution must integrate machine learning algorithms for training predictive models using the collected data. It should accommodate a range of machine-learning techniques tailored to meet specific application requirements.

- FR3: The solution must offer mechanisms for evaluating the trained models. It should encompass performance metrics designed to evaluate the pertinent indicators of the model's predictive abilities.

- FR4: The solution must be capable of performing real-time career path prediction based on the trained models.

- FR5: The solution must support interoperability with existing systems, such as databases or APIs.

- FR6: The solution must provide endpoints for predicting the next career path for functional positions and managerial positions, and retrieving position information, including the requirements to fulfill those positions.

Non-Functional Requirements

Besides the functional requirements, non-functional requirements are needed to establish the quality features and limitations that the proposed solution must meet. These requirements are crucial for guaranteeing the system's availability and reliability. The subsequent non-functional requirements are defined for the implementation of the proposed solution:

- NFR1: The solution needs to demonstrate strong performance, ensuring the swift and effective execution of machine learning algorithms, with an access time of less than 5 seconds for users.

- NFR2: The solution must be scalable. It should allow for the addition of new data sources, handle increasing amounts of data without degradation of performance, and support the growth in the number of users accessing the system simultaneously.

- NFR3: The solution must prioritize interpretability, ensuring that the output generated is understandable to end-users.

- NFR4: The solution must be designed for ease of maintenance and updates, with a particular focus on the monthly retraining of machine learning models to maintain accuracy and relevance. Additionally, the solution should include comprehensive documentation to assist developers in understanding and managing the retraining process, ensuring that updates can be performed efficiently and effectively with minimal disruption.

- NFR5: The solution should ensure high availability and reliability, meaning it can be accessed from anywhere at any time, with any downtime well-explained and justified.

- NFR6: The solution must include compliance with OAuth 2.0 for authorization.

- NFR7: The solution must comply with relevant regulations and standards related to data privacy and security in civil service data management.

- NFR8: The solution must ensure non-discrimination by excluding features related to race, gender, age, and other protected characteristics from the training data to support equitable opportunities for all civil servants.

4.2 Data Understanding

4.2.1 Data Sources

The dataset used in this research was derived from the NCSA database, the most comprehensive source of civil servant data in Indonesia. It includes personal data, position history, rank history, and education history for the civil servants of the NCSA as of April 30, 2024, with each type of data stored in separate tables. Information regarding data usage and access restrictions can be found in Appendix E. Initially, performance and competency records were considered due to their potential relevance. However, upon review, it was determined that the data in these tables were insufficient for robust analysis, leading to their exclusion from the final dataset. The data were collected through several information systems provided by the NCSA, which can be updated by human resources or the civil servants themselves. Additionally, the data can be updated through the institution's information systems, which are integrated into the NCSA's database.

4.2.2 Data Description

The collected data is subsequently analyzed by tallying the number of records: 4.770 personal data records, 17.252 position history records (including 3.927 managerial position history records and 5.152 functional position history records), 16.162 education history, 21.795 rank history records, 2.517 competency history records, and 17,806 performance history records. Despite the large volume of data, not all records are usable. For instance, there are only 1016 personal data records related to rank and managerial position history and 2.124 records for functional position history.

Another aspect to consider is the nature and completeness of the competency history and performance data. Most of the competency history data is relatively new and does not accurately reflect the competence of the civil servant at the time they held a position. This means that the competency records, which are supposed to indicate the skills and abilities of civil servants, might not provide a true picture of their actual competencies during their tenure in specific roles. The lack of historical competency data undermines the ability to assess how their skills have evolved over time and their suitability for past positions.

Additionally, much of the performance data is incomplete. Incomplete performance records pose significant challenges for analysis. Missing entries or partially filled records can lead to gaps that prevent a comprehensive evaluation of a civil servant's performance. This incompleteness can result from various factors, such as inconsistent data entry practices, data loss, or changes in performance evaluation criteria over time. These gaps in the performance data hinder the ability to conduct thorough performance assessments, compare historical performance accurately, and draw meaningful insights about the effectiveness and productivity of civil servants in their respective roles.

The features used in this research are derived from several columns in each table (personal data, job history, rank history, and education history). Table 4.1 lists the data provided in the database. In the position_history table, position_id can be found in multiple columns: functional_position_id, general_functional_position_id, and organizational_unit_id, depending on its position type.

There are 998 managerial positions documented in the database, with 239 currently active, while for functional positions, 115 are recorded in total, with 90 of them being active. The inactivity of some positions may result from the deactivation of the organizational unit/position or changes in nomenclature necessitating an update with new data. In addition to positions, the feature education_id also has many unique classes.

A large number of unique classes combined with a limited dataset can lead to several issues, such as overfitting, class imbalance, and high variance, which ultimately cause difficulties in model training and reduce accuracy. Furthermore, if only active positions are used as output, the number of records in the dataset will be reduced. In fact, there are both active and inactive positions that have similarities. These similarities can be observed in position type, position group, and position class. For example, the positions of Head of Regional Offices 1 to Head of Regional Offices 14 may share the same combination of these three columns but are defined by different IDs.

To conclude, the data available in the database is still very raw and not suitable for model training. Therefore, data preparation is necessary and will be explained in section 6.1.

TABLE 4.1: Columns of the Civil Servant Dataset

Table Name	**Column Name**	**Description**
civil_servant	cs_id	The identifier of the civil servant
civil_servant	pcs_start_date	The starting date of the person becoming prospective civil servant
civil_servant	first_rank	Civil servant rank when first employed
position_history	position_history_id	The identifier for the record of position changes
position_history	administration_position_id	The identifier of the administration position
position_history	functional_position_id	The identifier of the functional position
position_history	organizational_unit_id	The identifier of the organizational unit (can be used for managerial position)
position_history	position_start_date	The starting date of the new position
position_history	position_type	Type of the position. 1 = managerial; 2 = functional; 4 = administration
education_history	education_history_id	The identifier for the record of education degree
education_history	education_id	The identifier of the education degree
education_history	level_of_education_id	The identifier of the level of education
education_history	graduation_date	The graduation date of the degree
rank_history	rank_history_id	The identifier for the record of rank changes
rank_history	rank_start_date	The starting date of the person in the new rank
rank_history	rank	Civil servant rank
position	class	Position class
position	is_position_active	Active/inactive status of the position

Chapter 5

Design

This chapter presents the comprehensive design of the proposed solution, encompassing various critical components essential for effective implementation. The design process is structured to ensure that the system meets the identified requirements and performs optimally in a real-world context. Key elements covered in this chapter include the use case diagram, BPMN (Business Process Model and Notation), machine learning framework design, and system architecture design. Each of these components plays a vital role in illustrating the workflow, functionalities, and structural layout of the system. By detailing these elements, the chapter provides a clear blueprint for the development and integration of the machine learning models ensuring the framework's scalability, accuracy, and effectiveness in predicting career paths within the public sector.

5.1 Use Case Diagram

In order to clearly define and visualize the interactions between users and the career path prediction system, a use case diagram is essential. This diagram helps in identifying the key functionalities that the system must support and clarifies how users will interact with these features. The use case diagram illustrated in Figure 5.1 represents the primary interactions between the users and the career path prediction system. In this context, the "Actor" can be either an HR professional or a civil servant, both of whom interact with the system to access and utilize its functionalities.

At the core of the diagram is the "View Prediction" use case, which serves as the central feature of the system. This use case allows the user—whether HR or the civil servant themselves—to access the predicted career paths generated by the system based on the available data and inputs. The "View Prediction" use case is directly associated with the "Actor," indicating that this is the primary action that users engage in when interacting with the system.

Additionally, the diagram shows that the "View Prediction" use case is extended by two optional functionalities: "View Position Information" and "View Recommendation." The «extend» relationships signify that these two use cases are extensions of the "View Prediction" use case. They provide additional, optional information that enhances the user's understanding of the predicted career paths. Specifically, "View Position Information" allows the user to obtain detailed information about the positions associated with the predicted career path, while "View Recommendation" offers suggestions or recommendations that can help the user make more informed career decisions.

Furthermore, the "Login" use case is included in the "View Prediction" use case, as indicated by the «include» relationship. This inclusion suggests that the user must be authenticated before they can access the prediction functionality. The "Login" use case ensures that only authorized users—whether they are HR professionals or civil servants—can view predictions, thereby maintaining the security and privacy of the system.

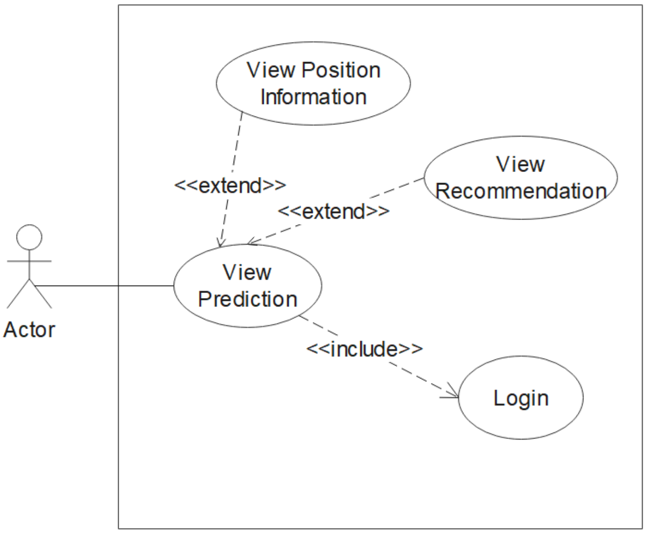


FIGURE 5.1: Use Case Diagram

5.2 Business Process Model and Notation

In order to provide a standard way to visualize the steps of a business process, a BPMN diagram is created, as shown in Figure 5.2. The proposed solution integrates four critical components: the Authentication System, the User, the Career Path Prediction System, and the Machine Learning Script. This diagram effectively illustrates the workflow and interactions between these components, ensuring clarity and coherence in understanding the entire prediction process.

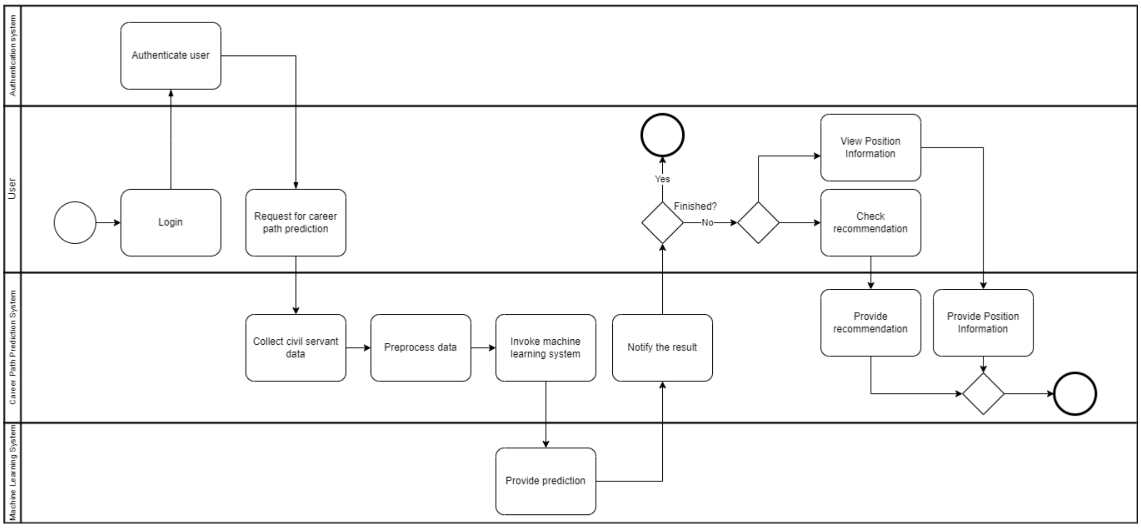


FIGURE 5.2: Career Path Prediction and Recommendation Process

It is important to note that while the diagram includes the Authentication System and the user's interactions, the scope of this thesis does not cover their implementation in the prototype. The focus of this research is primarily on the core components of the career path prediction system, which involves data collection, preprocessing, machine learning model development, and the generation of career path recommendations.

5.3 Machine Learning Framework Design

Based on the business and data understanding, the design for the model of career path prediction is presented in Figure 5.3. This figure outlines a comprehensive workflow divided into four main phases: Data Preparation, Modeling, Evaluation, and Deployment. Each phase encompasses critical steps necessary to transform raw datasets into actionable career path predictions. The figure also indicates how the proposed framework matches the functional requirements FR1, FR2, FR3, FR4, and FR5

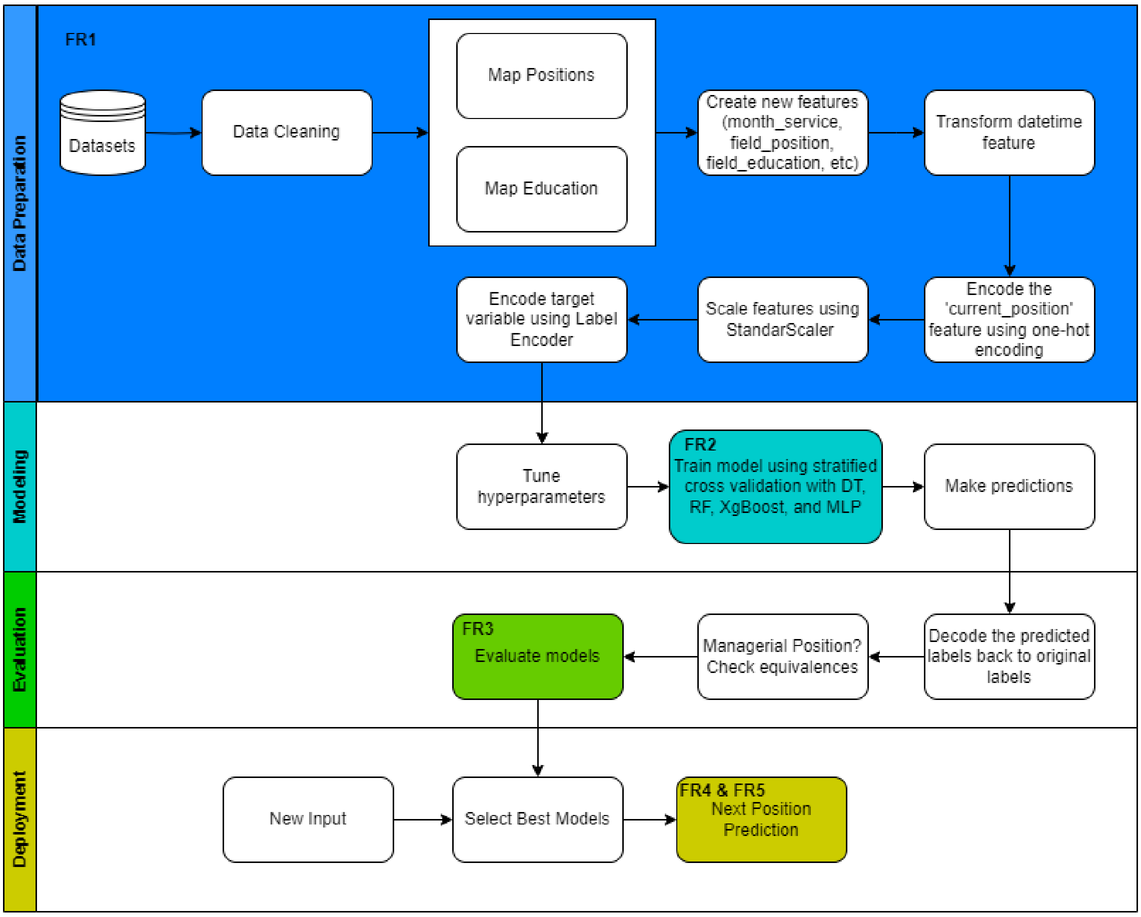

Data Preparation
FR1
Datasets
Data Cleaning
Map Positions
Map Education
Create new features (month_service, field_position, field_education, etc)
Transform datetime feature
Encode the 'current_position' feature using one-hot encoding
Scale features using StandarScaler
Encode target variable using Label Encoder
Modeling
Tune hyperparameters
FR2
Train model using stratified cross validation with DT, RF, XgBoost, and MLP
Make predictions
Evaluation
Decode the predicted labels back to original labels
Managerial Position? Check equivalences
FR3
Evaluate models
Deployment
New Input
Select Best Models
FR4 & FR5
Next Position Prediction

FIGURE 5.3: Career Path Prediction Framework

The **Data Preparation** phase begins with a data cleaning process by handling missing or null values and duplicate records to avoid potential bias or inaccuracies. Following this, position and education data were mapped to each position and education group. This step is crucial because the position and education data are extensive, with several entries being similar to one another and most of the position data is inactive due to changes in its nomenclature or organizational structures. Including these inactive positions could lead to unreliable outcomes while omitting them would reduce the overall dataset size. Mapping helps categorize and organize the data, ensuring consistency and relevance. Creating new features is the next step, where additional information is derived to enhance the dataset's value. The transformation of datetime features is then performed, converting dates into numerical values that represent time differences, making them usable by the model. Subsequently, the 'current_position' feature is encoded using one-hot encoding to convert categorical data into a binary format that the model can process. To ensure that each feature contributes equally to the model's predictions, features are scaled using StandardScaler. Finally, the target variable, representing the next career path, is encoded using Label Encoder to convert categorical labels into numerical form.

In the **Modeling** phase, hyperparameters of the model are tuned to optimize performance. This involves selecting the best set of parameters to enhance the model's accuracy and efficiency. The model is then trained using stratified cross-validation, ensuring it performs well across different subsets of the data. This technique helps validate the model's effectiveness and robustness. Once trained, the model makes predictions based on the input data.

The **Evaluation** phase involves assessing the performance of the models using various metrics. This step evaluates how well the model predicts career paths. For managerial positions, an additional step is taken to check equivalences, ensuring that predictions align with similar managerial roles. Finally, the predicted labels are decoded back to their original labels to interpret the results meaningfully.

In the **Deployment** phase, the best-performing models are selected based on the evaluation. The model is then prepared to handle new input data for future predictions. The final step involves predicting the next career position based on the new input data.

5.4 System Architecture Design

In designing the career path prediction system prototype, the Controller-Service-Repository (CSR) pattern was employed to ensure a clear separation of concerns, maintainability, and scalability. This pattern is widely used in enterprise applications and web development to organize code into distinct layers, each with specific responsibilities. Figure 5.4 illustrates the overall structure of the system.

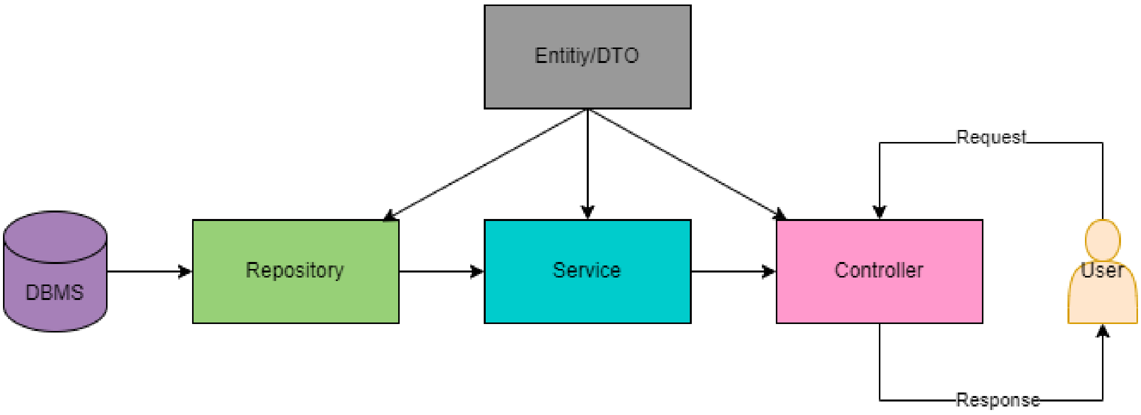


Figure 5.4: Controller-Service-Repository Pattern

Components

1. **Controller Layer:** The Controller layer manages incoming HTTP requests related to career path predictions. For instance, when HR or civil servants request a pre-

diction, the Controller directs the input data (like employee ID, months of service, etc.) to the Service layer. The Controller also determines which API endpoints (e.g., View Prediction, View Recommendation, or View Position Information) should be invoked based on the user's request.

2. **Service Layer:** The Service layer contains the core business logic that drives the prediction. This layer invokes the machine learning model to analyze historical career data and generate predictions. It also handles data processing tasks like normalizing inputs, executing feature engineering steps, and applying trained models. The Service layer ensures that the predictions align with predefined business rules, such as career path guidelines set by the government agency.

3. **Repository Layer:** The Repository layer manages the data access required for predictions. It retrieves the necessary data (e.g., employee history, job descriptions, and organizational structures) from the database and other sources. The data is preprocessed and transformed before being fed into the machine learning models. The repository also manages connections to the database to retrieve prediction results, enabling the system to maintain accurate and up-to-date information.

   Additionally, Models are an essential part of this architecture:

4. **Entity:** Entities represent the core business objects and data structures within the application. They define the attributes and relationships of the data stored in the database.

5. **DTO (Data Transfer Objects):** DTOs are used to transfer data between layers, particularly between the controller and service layers. They encapsulate the data that needs to be sent or received, ensuring a clear separation from the internal data structures.

This architecture not only organizes the system in a modular and maintainable way but also directly supports the specific needs of the career path prediction system by ensuring that data flow is streamlined, predictions are generated accurately, and the system remains scalable and easy to maintain as new features or data are introduced.

Chapter 6

Implementation

This chapter details the implementation phase of the career path prediction framework for civil servants in Southeast Asia, following the CRISP-DM methodology from data preparation to deployment. It begins with data collection and preprocessing from the NCSA database to ensure data integrity and accuracy. The modeling phase then develops and fine-tunes four machine learning algorithms—Decision Tree (DT), Random Forest (RF), XGBoost, and Multilayer Perceptron (MLP)—leveraging their strengths to enhance prediction accuracy and robustness. Evaluation of these models is conducted using performance metrics such as accuracy, precision, recall, and F1-score, with feature importance analyzed through SHAP values. The deployment section outlines the creation of a functional prototype using the Spring Boot framework, detailing the system architecture and model integration into a user-accessible application. Finally, the chapter addresses the validation of the deployed model, ensuring its reliability and effectiveness in providing valuable career path predictions for civil servants.

6.1 Data Preparation

Data preparation for this research began with collecting relevant data from the NCSA database, the most comprehensive source for civil servant data. Although the project relied on a single database, multiple tables within the NCSA database were selected to gather pertinent information. These tables included data on employee demographics, job history, rank history, and educational background. Initially, performance and competency history were considered due to their potential relevance. However, upon review, it was determined that the data in these tables was insufficient for robust analysis, leading to their exclusion from the final dataset.

Once the relevant data was selected, data cleaning was performed to ensure data integrity and accuracy. Missing or null values were systematically excluded, ensuring that only complete records were used for modeling to avoid potential biases or inaccuracies. Additionally, duplicate records were identified and removed to ensure each entry represented a unique instance of a civil servant's career progression, preventing skewed analysis and model results.

Following the data cleaning, the process of mapping from `position_id` to a new column, which is a combination of `position_type`, `position_group`, and `position_class`, is carried out. However, for now, these positions have not yet been grouped. Thus, before the mapping process is carried out, a new column called `position_group` is created first, and these positions are grouped into the combination of 8 position fields, namely: Management, ICT, Law, Economics, Politics, Psychology, Social, and Linguistics. This mapping process results in 34 managerial positions and 68 functional positions. The mapping for the position follows the circular letter from the Head of the National Civil Service Agency, Number 4 of 2023, regarding the Management of Talent within the National Civil Service Agency.

Grouping educational majors is also necessary. Each major will be grouped into educational fields, namely: Civil, Dental nursing, Dentistry, Economics, Education, ICT, Humanities, Law, Linguistics, Management, Medical, Nursing, Pharmacy, Politics, Psychology, Social, Theology, and 'Other'. This education mapping is based on the regulation of the Minister of Education and Culture of the Republic of Indonesia, number 154 of 2014. The mapping process for education can simplify the number of unique majors from 537 to 18 educational fields, each with their respective education levels.

After completing the mapping step, new columns in the position history table were created and populated with the position history and education history groups. These columns indicate whether the civil servants had ever held each position or obtained each degree in their respective fields before attaining that position.

Data integration then took place, combining information from the selected tables within the NCSA database to create a comprehensive dataset. Queries were crafted to build a data frame for each position type, resulting in 2,919 records for managerial positions and 4,396 records for functional positions. The outcomes of these queries were datasets without null values, eliminating the need for further imputation steps. Features such as `cs_id`, `position_history_id`, `rank_history_id`, and `education_history_id`, which do not substantially contribute to the next position prediction process, were excluded to improve the model's focus on pertinent information. Features such as `position_month_service` and `month_service` were calculated using `position_start_date` of the current and next positions and `pcs_start_date`. Furthermore, the `next_position` feature contains combinations of position type, position group, and position class. The completed features in the data frame are listed in Table C.1 in the Appendix C.

The next step was to exclude rare classes that appear fewer than three times to avoid issues during the 5-fold cross-validation of the dataset. Following this, dummy variables were created using one-hot encoding. This step is crucial since most machine learning algorithms require numerical input, and categorical variables such as `current_position` must be converted to a numerical format. Additionally, this avoids interpreting the feature as having any natural order. Further steps involved scaling the features using `StandardScaler` to ensure that each feature contributes equally to the distance calculation and model prediction. Feature scaling is also important to achieve optimal model convergence during training. Concurrently, `LabelEncoder` was applied to the target variable, `next_position`, to ensure compatibility with the machine learning algorithm and prevent potential runtime errors.

6.2 Modeling

In order to effectively predict career paths, a diverse set of machine learning models is proposed and thoroughly developed. This variety in models allows for a more comprehensive analysis, leveraging the strengths of different algorithms to improve overall prediction accuracy and robustness. Based on the analysis of the literature review, specifically in Section 2.3.4 and 2.3.5, this project employs three tree-based algorithms—Decision Tree, Random Forest, and XGBoost—along with a neural network algorithm, Multilayer Perceptron. Despite the value of Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs) in various applications, they are not utilized in this project. This decision is due to the nature of the dataset, which is neither sequential nor image-based. LSTM networks are specifically designed for sequential data, such as time series or natural language processing tasks, where the order of the data points is crucial. On the other hand, CNNs excel at processing grid-like data, such as images, where spatial hierarchies and patterns are essential. Given that the dataset in this project does not fit these criteria, using LSTM or CNN models would not be appropriate or beneficial. Instead, the chosen algorithms—tree-based methods and Multilayer Perceptrons—are better suited to handle the structured, non-sequential data used for predicting career paths.

The primary motivation for employing a variety of regression algorithms is to leverage the unique strengths of each approach. A decision tree model was implemented as a baseline due to its simplicity and interpretability. Decision trees are easy to understand and visualize, making them a good starting point for model development. Building on the decision tree, a random forest model was employed to enhance predictive performance and reduce overfitting. Random forests leverage the power of multiple weak learners by creating an ensemble of decision trees, which improves robustness and accuracy compared to a single decision tree. XGBoost, an advanced gradient-boosting technique, was also implemented. XGBoost is particularly powerful because of its ability to handle missing data and its effectiveness in preventing overfitting through its boosting approach. This algorithm is known for its high performance and efficiency in predictive tasks. Finally, a Multilayer Perceptron, a type of artificial neural network algorithm, was utilized. Artificial neural networks, particularly Multilayer Perceptrons, are capable of capturing complex, non-linear relationships within the data. This makes them highly suitable for modeling intricate patterns and interactions in large datasets. Each of these algorithms was selected to capitalize on their specific advantages, thereby ensuring a comprehensive and robust predictive modeling approach for career path prediction.

At the beginning of this phase, each model is initialized with its specific algorithm and a predefined set of hyperparameters. The initial hyperparameters for each model can be found in Appendix D. To ensure that each model achieved its optimal performance, balancing complexity, and predictive accuracy, each model's hyperparameters were fine-tuned using `RandomizedSearch`. For the Decision Tree model, parameters such as the maximum depth of the tree and the minimum number of samples required to split a node were adjusted. In the case of the Random Forest model, the number of trees in the forest, the maximum depth of each tree, and the number of features considered for splitting were optimized. XGBoost, known for its efficiency and predictive power, required tuning of its learning rate, maximum depth, and the number of boosting rounds. Finally, for the Multilayer Perceptron, the architecture was fine-tuned by adjusting the number of layers, the number of neurons in each layer, the learning rate, and the activation functions. Randomized search was chosen for its efficiency in exploring a wide range of hyperparameter values, enabling the identification of the best configurations without exhaustive computation.

To ensure that the models developed are robust and reliable, 5-fold cross-validation is employed during the model training process. The process involves partitioning the available data into five equally sized subsets using `StratifiedKFold` to ensure that each fold has approximately the same proportion of each class as the entire dataset. In each of the five iterations, four of these subsets are used as training data, and the remaining subset is used as validation data. This ensures that every data point is used for both training and validation, providing a comprehensive evaluation of the model's performance. By averaging the results across the five folds, 5-fold cross-validation helps mitigate the risk of overfitting and provides a more generalized measure of the model's accuracy. This method balances the need for reliable performance estimation with computational efficiency, making it a popular choice for model evaluation in scenarios where data is limited.

6.3 Evaluation

6.3.1 Performance Metrics

To demonstrate the effectiveness of the novel model implemented in this career path prediction, a comparative analysis is conducted. The evaluation process involves the calculation of four metrics, including accuracy, precision, recall, and f1-score [46].

Before calculating the result of the performance metrics, decoding the predicted labels back to their original labels is crucial not only to ensure that the results are interpretable and useful for decision-making but also to find the equivalence of the result in managerial position prediction. The same `LabelEncoder` instance used for encoding the original labels is utilized for decoding, ensuring consistency and correctness in mapping the encoded labels back to their original form.

Additionally, unlike functional positions, managerial positions are often mapped to more than one position field. Therefore, defining the equivalences of every position field is essential. Notably, the equivalence must share the same position type and position class. For example, positions mapped to `'1_LAW-MANAGEMENT-POLITIC-SOCIAL_15'` are equivalent to positions mapped to `'1_LAW-MANAGEMENT-POLITIC_15'`, since individuals who can hold positions requiring expertise in law, management, politics, and social fields should also be suitable for positions requiring expertise in law, management, and politics.

Table 6.1 presents a detailed performance evaluation of four machine learning models: Decision Tree (DT), Random Forest (RF), XGBoost, and Multilayer Perceptron (MLP), across two types of positions: Managerial and Functional.

Table 6.1: Performance Evaluation of Test Data

ML Model	**Managerial**				**Functional**			
-	Accuracy	Recall	Precision	F1-Score	Accuracy	Recall	Precision	F1-Score
DT	62.25	68.58	62.42	64.26	77.25	77.25	76.25	76.46
RF	**65.61**	**72.44**	64.62	63.72	83.33	83.33	82.43	82.58
XgBoost	63.14	69.67	65.03	66.55	**83.54**	**83.54**	**83.16**	**82.99**
MLP	65.51	70.67	**65.83**	**67.54**	72.08	72.08	75.87	71.84

Table 6.1 shows that for managerial positions, the RF classifier demonstrates the highest accuracy at 65.61%, indicating it predicts managerial roles more accurately than the other models. This classifier also achieves the highest recall (72.44%), suggesting it is particularly effective at identifying actual managerial positions within the dataset. However, its precision (64.62%) is slightly lower compared to MLP, which implies that while it identifies a high number of actual managerial positions, it also has a higher rate of false positives. XGBoost, with an accuracy of 63.14%, also performs well and has a balanced F1-Score of 66.55%. The MLP model, despite a slightly lower accuracy (65.51%), shows solid performance with a recall of 70.67% and the highest precision (65.83%). Its F1-Score (67.54%) is the highest among all models, indicating a balanced approach similar to XGBoost in handling managerial positions. The DT classifier has the lowest performance among the four models for managerial positions, with an accuracy of 62.25% and the lowest precision (62.42%). However, its recall (68.58%) is comparable to the other classifiers, indicating it is reasonably good at identifying actual managerial positions but tends to predict more false positives.

Table 6.1 also suggests that for functional positions, XGBoost stands out as the top performer with the highest accuracy (83.54%), recall (83.54%), precision (83.16%), and F1-Score (82.99%). This indicates that XGBoost is highly effective and reliable across all metrics for predicting functional positions, making it the best model for this task. The RF classifier closely follows XGBoost, with an accuracy of 83.33% and similar recall (83.33%) and precision (82.43%). Its F1-Score (82.58%) is also very close to that of XGBoost, suggesting that RF is nearly as effective as XGBoost in predicting functional positions. The DT classifier, while not as strong as XGBoost and RF, still performs reasonably well with an accuracy of 77.25% and balanced recall (77.25%) and precision (76.25%). Its F1-Score (76.46%) indicates a solid performance, though not at the level of the ensemble methods. The MLP classifier shows the lowest performance for functional positions, with an accuracy of 72.08%. However, it maintains balanced recall (72.08%) and precision (75.87%). Its F1-Score (71.84%) is the lowest among the four classifiers, indicating that MLP may struggle more with the complexity of functional position predictions compared to the other models.

The lower performance metrics observed for managerial positions compared to functional positions can be attributed to several factors. Managerial roles often require a broader range of skills and competencies, making them inherently more complex to predict accurately. Moreover, decisions regarding managerial positions are frequently influenced by subjective criteria and organizational politics, which are not easily captured by algorithmic models. This political aspect of managerial decisions can lead to inconsistencies and biases in the training data, further complicating the prediction process. As a result, even advanced models may struggle to achieve the same level of precision and recall for managerial positions as they do for functional ones.

Overall, ensemble methods like Random Forest and XGBoost exhibit superior performance, particularly for functional positions, due to their ability to handle complex patterns and reduce overfitting. The detailed evaluation underscores the importance of selecting appropriate models based on the specific characteristics of the prediction task.

6.3.2 Feature Importance

Feature importance in the selected machine learning algorithms is verified through the SHAP (SHapley Additive exPlanations) method [41]. This method provides a unified measure of feature importance by distributing the prediction among the features in a fair way, offering insights into how each feature contributes to the model's output.

Figures 6.1 and 6.2 illustrate the top feature importances for the MLP and XGBoost models in predicting managerial and functional positions, respectively. The SHAP values indicate the magnitude of impact each feature has on the model's predictions.

Figure 6.1 shows that in predicting managerial positions using the MLP classifier, "current_position" stands out as the most influential feature, followed by "position_history" and "education". These features significantly impact the model's ability to predict managerial positions, highlighting the importance of a candidate's current role, previous positions, and educational background. Other notable features include "rank_start_date", "rank", and "class", which further contribute to the model's predictions.

Figure 6.2 shows that in predicting functional positions using the XGBoost classifier, "position_history" is the most important feature, followed by "education" and "rank". This indicates that for functional positions, an individual's past roles, educational background, and rank are critical predictors. Other significant features include "current_position", "month_service", and "position_month_service", which highlight the relevance of a candidate's current role and their duration in service.

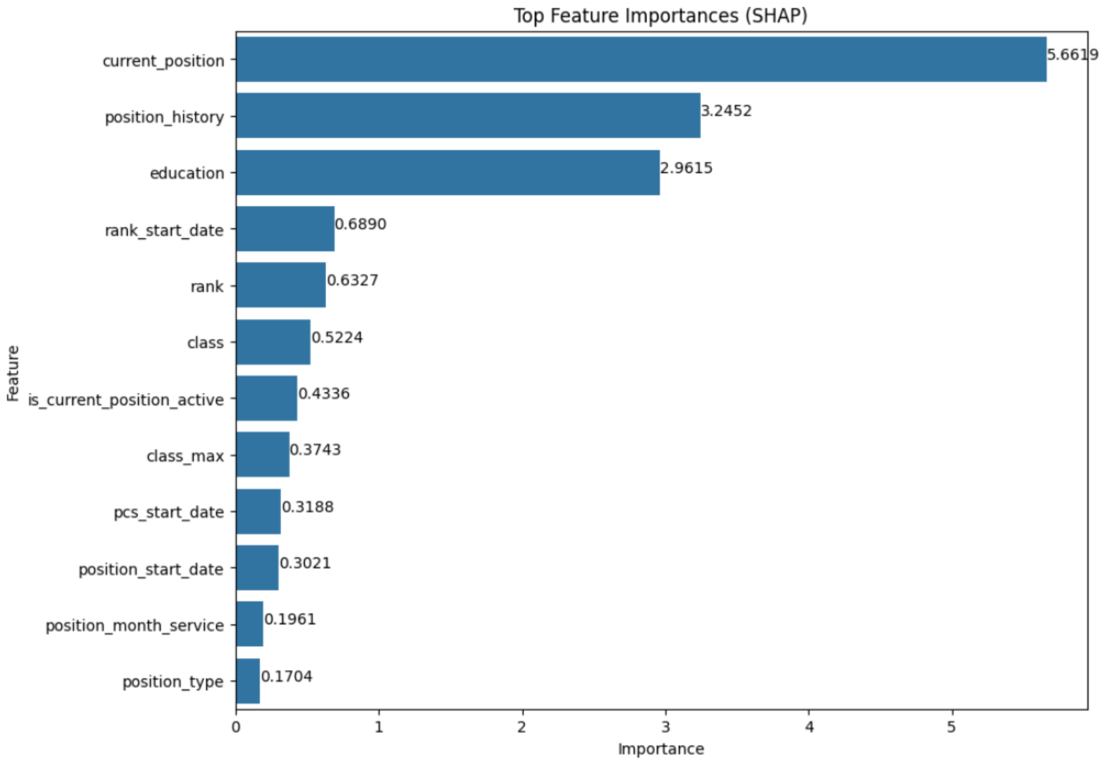


FIGURE 6.1: Top Feature Importances (SHAP) for MLP Model Predicting Managerial Positions

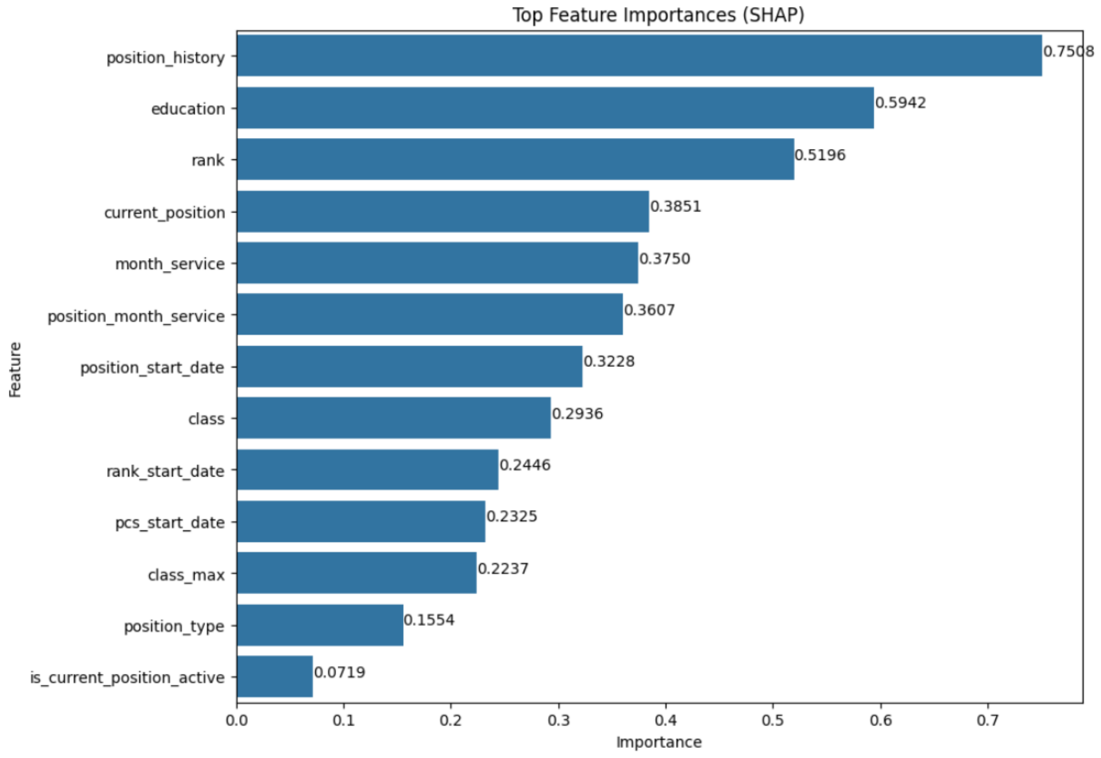


FIGURE 6.2: Top Feature Importances (SHAP) for XgBoost Model Predicting Functional Positions

6.4 Deployment

It is important to note that the scope of this project does not include deploying the final model in a real-world operational setting. Instead, the focus is on developing a comprehensive prototype that demonstrates the potential capabilities of the career path prediction framework.

The deployment phase of the career path prediction framework centers on creating a prototype that integrates the predictive model into a functional application accessible to users. This prototype leverages the Spring Boot 2.6.3 framework, a widely-used tool in Java for building stand-alone, production-grade applications. Java 17, chosen for its performance enhancements and long-term support, underpins the development environment.

The implementation of this prototype ensures that the predictive model can be effectively utilized in a simulated environment, offering valuable career path predictions for civil servants. By serving as a proof of concept, the prototype demonstrates how the predictive model can be integrated into a functional system, highlighting its potential utility in providing actionable career path predictions.

6.4.1 Technology Stack

Several key technologies are included in the project to support its functionality:

1. **Java Development Kit (JDK)**: Provides the Java Runtime Environment (JRE), an interpreter/loader (Java), a compiler (javac), an archiver (jar), a documentation generator (Javadoc), and other tools needed for Java development, ensuring compatibility with Spring Boot and other Java-based frameworks.

2. **Maven**: Manages project dependencies and automates the build process, ensuring consistency and efficiency across development and deployment stages.

3. **Spring Web**: Provides the necessary components for building web applications, including RESTful services.

4. **PostgreSQL JDBC Driver**: Used to connect to the PostgreSQL database.

5. **Spring Data JPA (Java Persistence API)**: Facilitates the management of relational data in Java applications. JPA is used for mapping Java objects to database tables, ensuring seamless integration with the PostgreSQL database.

6. **Swagger**: Integrated for API documentation and testing. Swagger provides a user-friendly interface for accessing and testing the API endpoints, making it easier for developers and users to interact with the application.

6.4.2 API Overview

The prototype includes a set of three API endpoints that allow users to interact with the career path prediction system. These APIs are designed to facilitate access to key functionalities, enabling both civil servants and HR professionals to retrieve and utilize career-related data effectively. The system's primary API is the View Prediction API, which serves the core functionality of predicting the next possible career positions for an employee after a specified number of months into the future. The prediction is generated based on the employee's profile and historical data, allowing users to gain insights into both functional and managerial positions they are likely to attain.

In addition to the View Prediction API, the system also features the Position Information API and Career Path Recommendation API. The Position Information API retrieves detailed data about a specific job position, including qualifications, responsibilities, and common career trajectories, helping employees and HR managers better understand what each role entails. Meanwhile, the Career Path Recommendation API builds on the predictions by offering personalized recommendations for career paths based on projected months and position data, helping guide employees in aligning their career goals with organizational needs.

The application exposes its functionalities through a RESTful API, accessible via Swagger. Swagger's interface allows users to test the endpoints directly from the browser, providing a convenient way to understand and utilize the APIs. This not only helps in validating inputs and outputs but also ensures that both technical and non-technical users can effectively engage with the system's features. The APIs are structured with clear endpoints and JSON data formatting, making them easy to integrate into existing systems or workflows within the organization.

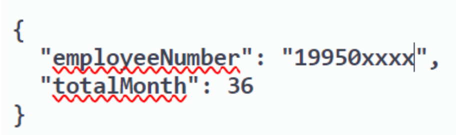
View Prediction API

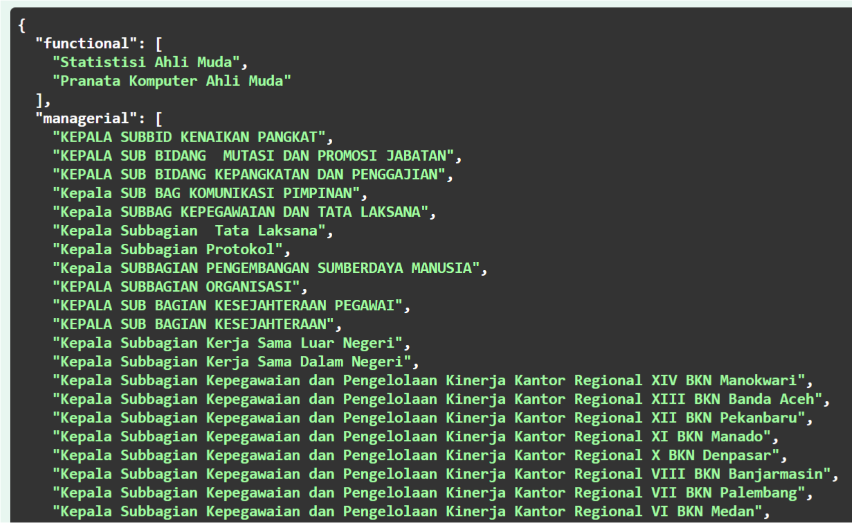
The View Prediction API provides a list of potential career paths categorized into functional and managerial job positions. Users can input an employee's details, such as their employee number and a projected number of months into the future, and the API returns the predicted positions based on the machine learning model's analysis. This feature is crucial for both employees looking to plan their future roles and HR professionals aiming to strategically manage talent and workforce distribution.

Table 6.2 provides a detailed breakdown of the View Prediction API, including its endpoint, required request parameters, and the expected response structure. The table outlines how users can interact with the API to generate predictions, making it easy for them to incorporate this functionality into daily HR operations or career planning initiatives.

TABLE 6.2: Career Path Prediction API specification

Endpoint	**POST /api/predict**
Description	Given employee number and total months, returns lists of functional and managerial positions
Request Headers	`Content-Type: application/json`

Request Body	employeeNumber: String. Required.
	totalMonth: Integer. Required.
	Example:
	
	{
	"employeeNumber": "19950xxxx",
	"totalMonth": 36
	}

Response Body	The JSON response consists of two arrays, each containing position names categorized by type: functional and managerial. `functional`: Array of strings. This array contains the names of functional job positions. `managerial`: Array of strings. This array contains the names of managerial job positions. Example: 

```
{
  "functional": [
    "Statistisi Ahli Muda",
    "Pranata Komputer Ahli Muda"
  ],
  "managerial": [
    "KEPALA SUBBID KENAIKAN PANGKAT",
    "KEPALA SUB BIDANG  MUTASI DAN PROMOSI JABATAN",
    "KEPALA SUB BIDANG KEPANGKATAN DAN PENGGAJIAN",
    "Kepala SUB BAG KOMUNIKASI PIMPINAN",
    "Kepala SUBBAG KEPEGAWAIAN DAN TATA LAKSANA",
    "Kepala Subbagian  Tata Laksana",
    "Kepala Subbagian Protokol",
    "Kepala SUBBAGIAN PENGEMBANGAN SUMBERDAYA MANUSIA",
    "KEPALA SUBBAGIAN ORGANISASI",
    "KEPALA SUB BAGIAN KESEJAHTERAAN PEGAWAI",
    "KEPALA SUB BAGIAN KESEJAHTERAAN",
    "Kepala Subbagian Kerja Sama Luar Negeri",
    "Kepala Subbagian Kerja Sama Dalam Negeri",
    "Kepala Subbagian Kepegawaian dan Pengelolaan Kinerja Kantor Regional XIV BKN Manokwari",
    "Kepala Subbagian Kepegawaian dan Pengelolaan Kinerja Kantor Regional XIII BKN Banda Aceh",
    "Kepala Subbagian Kepegawaian dan Pengelolaan Kinerja Kantor Regional XII BKN Pekanbaru",
    "Kepala Subbagian Kepegawaian dan Pengelolaan Kinerja Kantor Regional XI BKN Manado",
    "Kepala Subbagian Kepegawaian dan Pengelolaan Kinerja Kantor Regional X BKN Denpasar",
    "Kepala Subbagian Kepegawaian dan Pengelolaan Kinerja Kantor Regional VIII BKN Banjarmasin",
    "Kepala Subbagian Kepegawaian dan Pengelolaan Kinerja Kantor Regional VII BKN Palembang",
    "Kepala Subbagian Kepegawaian dan Pengelolaan Kinerja Kantor Regional VI BKN Medan",
```

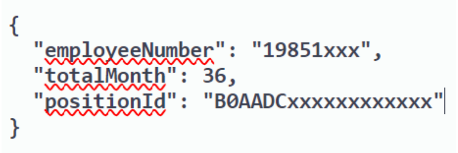
View Recommendation API

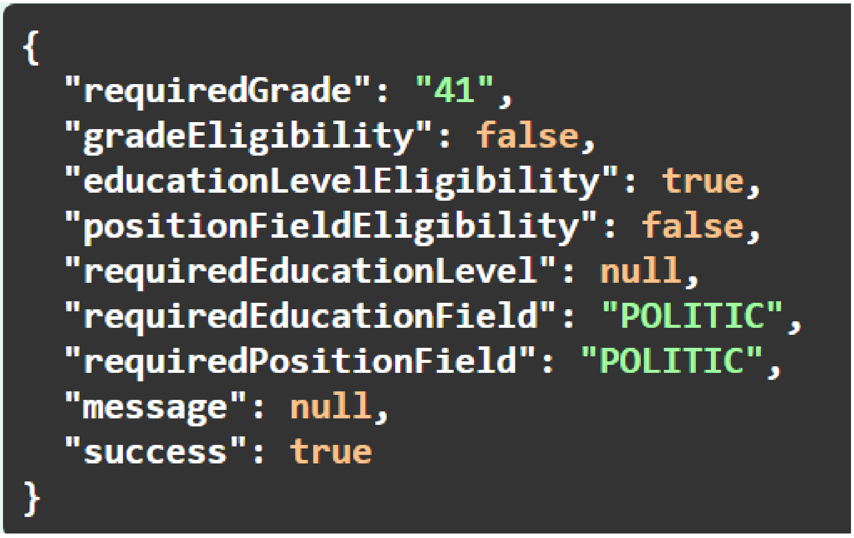
While the core feature of the system is the View Prediction API, which forecasts future career moves based on historical data, the View Recommendation API complements it by offering personalized career guidance tailored to an employee's profile and projected service duration. This API provides actionable recommendations by evaluating the predicted position alongside factors such as grade, education, and field alignment.

The recommendations generated by this API are designed to help employees and HR professionals focus on realistic career paths, providing insights into what steps can be taken to align future roles with organizational goals. By integrating the predicted position ID, the system ensures that the recommendations are relevant and targeted. Table 6.3 details the Career Path Recommendation API, covering the endpoint, input parameters (like position ID), and response structure, offering a clear guide for utilizing this functionality effectively.

TABLE 6.3: Career Path Recommendation API specification

Endpoint	**POST /api/recommendation**
Description	Given employee number total months, and positionId, returns personalized career path recommendations.
Request Headers	`Content-Type: application/json`

Request Body	employeeNumber: String. Required.
	totalMonth: Integer. Required.
	positionId: String. Required.
	Example:
	
	{
	  "employeeNumber": "19851xxx",
	  "totalMonth": 36,
	  "positionId": "B0AADCxxxxxxxxxxxx"
	}

Response Body	The JSON response consists of various eligibility fields and required attributes for the recommended position, such as the required grade, eligibility criteria, and required fields of education and position.
	Example:
	
	{
	  "requiredGrade": "41",
	  "gradeEligibility": false,
	  "educationLevelEligibility": true,
	  "positionFieldEligibility": false,
	  "requiredEducationLevel": null,
	  "requiredEducationField": "POLITIC",
	  "requiredPositionField": "POLITIC",
	  "message": null,
	  "success": true
	}

View Position Information API

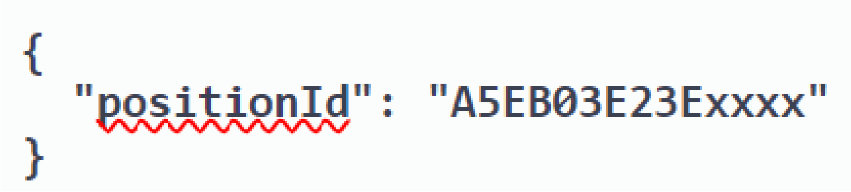
The Position Information API provides detailed information about specific job positions within the organization. This API is essential for both HR professionals and employees seeking to understand the qualifications, responsibilities, and typical career paths associated with a given position. By offering comprehensive data on the requirements and characteristics of various roles, the API enables users to make informed career decisions and align their career development strategies with organizational expectations.

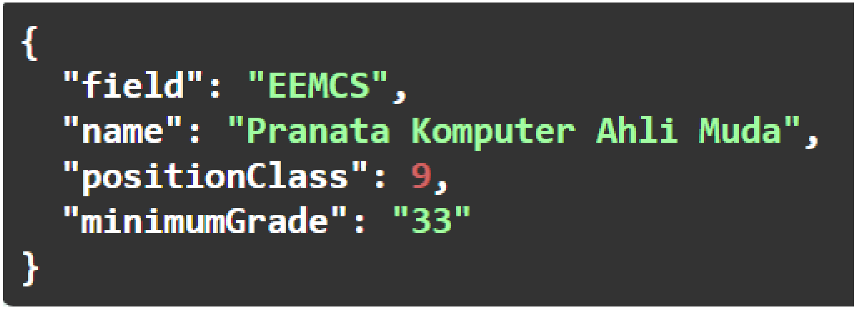
This API plays a crucial role in the system by allowing users to retrieve critical details such as the field of work, position name, class, and minimum grade. Whether used by HR professionals for strategic planning or by employees exploring potential career moves, the information delivered by this API helps users better understand the landscape of available positions and how they align with their career goals. The specifications of the Position Information API are presented in Table 6.4, which outlines the endpoint, required inputs, and the format of the response.

Table 6.4: Position Information API specification

Endpoint	**GET /api/position-info**
Description	retrieve detailed information about various job positions, including qualifications, responsibilities, and typical career trajectories.

Request Headers	`Content-Type: application/json`

Request Body	positionId: String. Required.
	Example:
	
	`{`
	`  "positionId": "A5EB03E23Exxxx"`
	`}`

Response Body	The JSON response consists of various fields detailing the job position, such as the field of work, position name, position class, and minimum grade.
	Example:
	
	`{`
	`  "field": "EEMCS",`
	`  "name": "Pranata Komputer Ahli Muda",`
	`  "positionClass": 9,`
	`  "minimumGrade": "33"`
	`}`

Chapter 7

Validation

In this chapter, we present the validation of the proposed career path prediction system designed to assist civil servants and HR professionals in government organizations. The primary aim of this validation is to assess the system's potential effectiveness, predictive accuracy, usability, and integration within existing HR workflows. Given that the system is currently in the prototype stage, the validation process focuses on gathering qualitative and quantitative feedback from potential users and stakeholders through various methods. To evaluate the potential acceptance and usability of the proposed career path prediction system, this study utilizes a combination of the Technology Acceptance Model (TAM) [14] and the Unified Theory of Acceptance and Use of Technology (UTAUT) [64]. TAM is primarily used to assess users' perceptions of the system's usefulness and ease of use, while UTAUT provides additional insights into the influence of social factors and facilitating conditions on the system's adoption. This hybrid approach allows for a comprehensive evaluation of the system from both individual and organizational perspectives.

7.1 Participant

To ensure that the validation of the career path prediction system effectively meets its intended objectives and provides valuable insights into its potential utility and integration within government organizations, three key participants with extensive experience in relevant fields were invited to take part in the validation process. These participants were selected based on their roles, expertise, and long tenure within the organization, ensuring a comprehensive evaluation from both technical and practical perspectives. Table 7.1 below provides details on the composition of the participating panel. The panel includes a Human Resources Analyst, a Software Engineer, and a Civil Servant who also has experience as a software engineer. While the Human Resources Analyst and Software Engineer provide expert insights into the system's technical and functional aspects, the Civil Servant represents the end-user perspective, ensuring that the system's usability and relevance to daily operations are thoroughly assessed.

Table 7.1: Participating Panel Composition

ID	Role	Years of Experience
P1	Human Resources Analyst	13 years
P2	Software Engineer	15 years
P3	Civil Servant	16 years

This carefully selected panel was chosen to capture a wide range of perspectives on the system's effectiveness, usability, and potential for integration into existing workflows. The inclusion of a participant who represents the end-user ensures that the system is not only technically robust but also meets the practical needs of its intended users, making it both effective and user-friendly.

7.2 Validation Process

The validation process began with a comprehensive explanation of the project, including an overview of the objectives and scope of the career path prediction system, a detailed description of the methodology used to develop the system, and an introduction to the validation methods selected for this study. These methods included a combined survey utilizing both the Technology Acceptance Model (TAM) and the Unified Theory of Acceptance and Use of Technology (UTAUT). The goal was to gather in-depth feedback on the system's potential usefulness, ease of use, social influence, and facilitating conditions, as well as its integration into existing HR workflows.

Following the project explanation, a presentation and demonstration of the prototype were conducted. This step involved showcasing the main features and functionalities of the prediction system, including how the system processes data and generates career path predictions. To facilitate a thorough evaluation, participants were presented with realistic scenarios that demonstrated the system's application in career planning. These scenarios were designed to elicit feedback on the system's predictive accuracy, ease of use, alignment with organizational goals, and overall relevance to users' needs.

The demonstration also included discussions on how the system could be integrated into current HR systems, with a particular emphasis on gathering insights into the practicality of implementation and potential barriers. Participants were encouraged to provide feedback on the system's usability, trustworthiness, social influence, and likelihood of adoption based on their professional experiences and roles within the organization. This feedback was essential in assessing the system's readiness for deployment and identifying areas for further refinement.

7.3 Validation Criteria and Questions

After the demonstration, participants were asked to provide their feedback through structured surveys. The surveys consisted of questions specifically designed to evaluate the career path prediction system, focusing on key aspects relevant to its practical implementation and utility. These aspects include the system's perceived usefulness in career planning, predictive accuracy, ease of understanding, trustworthiness, and potential for integration into existing HR workflows.

Participants were asked to respond to each question using a scale from 1 to 5, where 1 indicates strong disagreement or a very negative evaluation, and 5 indicates strong agreement or a very positive evaluation. The validation criteria and corresponding questions used to evaluate the career path prediction system are listed in Table 7.2. Furthermore, two open-ended questions were posed to gather feedback aimed at refining the model:

1. What aspects of the framework do you believe are its strongest?
2. Do you have any further comments or suggestions for enhancing the framework?

Table 7.2: Validation Criteria and Questions

No.	Question	Criterion	Model
1	How useful do you find the prediction system for supporting career planning initiatives?	**Perceived Usefulness (PU)**	TAM
2	How well do you think the system can predict the next career position for employees?		
3	How effective do you find the system in aligning employees' career paths with organizational goals?		
4	How easy is it to use the system for generating and interpreting career path predictions?	**Perceived Ease of Use (PEOU)**	
5	How likely are you to recommend this system to other HR colleagues?	**Behavioral Intention to Use (BI)**	
6	How much do you think employees will trust and rely on the predictions made by the system?		
7	How well do you think the system can be integrated into existing HR systems?	**Facilitating Conditions**	UTAUT
8	How confident are you in the availability of technical support for using this system?		

7.4 Validation Result and Feedback Analysis

After conducting the validation round with all participants, this section summarizes the feedback and opinions regarding the assessment and validation of the proposed career path prediction system. Table 7.3 below details the assessments provided by each participant, focusing on the system's perceived usefulness, predictive accuracy, ease of use, trustworthiness, and potential for integration into existing HR workflows. The table includes scores from each participant, with an average score per question and the overall average score also calculated. Overall, the system received a generally positive reception, with an overall average score of 4.42, indicating that participants found the system to be effective and useful in supporting their needs related to career planning and management within the organization.

One of the key strengths highlighted by the participants was the system's potential usefulness in career planning, where it received a perfect score of 5.00 from all participants. This suggests that the participants see strong potential in the system's ability to assist employees in strategically planning their career paths, which aligns with one of the primary objectives of the prototype. Additionally, the concept of the system being easy to use, particularly in generating and interpreting career path predictions, was rated highly, also achieving a perfect score of 5.00. This high rating reflects participants' belief that, once fully developed, the system could be user-friendly and accessible, although this is based on their understanding of the prototype rather than direct interaction with a final product. Trust and reliability in the system's predictions were also positively noted, with the prototype earning an average score of 4.67. This indicates that participants are optimistic about the system's ability to provide reliable predictions, which is crucial for its future success.

However, the analysis also pointed to areas where the prototype could be refined before full development. The system's ability to accurately predict the next career position for employees received a lower average score of 3.67, the lowest in the survey. This suggests that participants have some reservations about the prototype's current predictive capabilities, which may require further development and testing. Enhancing the predictive algorithms or incorporating additional data points could help address these concerns as the system moves beyond the prototype stage. Additionally, confidence in the availability of technical support for the system prototype was relatively lower, with an average score of 4.00. While this is still a positive score, it indicates that participants foresee potential challenges in technical support, which could impact the system's adoption and long-term use once fully implemented. Addressing this concern by planning for robust support structures or providing more comprehensive training could be beneficial as the system is developed further.

Table 7.3: Questionnaire Results

No.	**Questions**	**P1**	**P2**	**P3**	**AVG**
1	How useful do you find the prediction system for supporting career planning initiatives?	5	5	5	5.00
2	How well do you think the system can predict the next career position for employees?	3	4	4	3.67
3	How effective do you find the system in aligning employees' career paths with organizational goals?	4	4	5	4.33
4	How easy is it to use the system for generating and interpreting career path predictions?	5	5	5	5.00
5	How likely are you to recommend this system to other HR colleagues?	4	4	5	4.33
6	How much do you think employees will trust and rely on the predictions made by the system?	5	5	4	4.67
7	How well do you think the system can be integrated into existing HR systems?	4	4	5	4.33
8	How confident are you in the availability of technical support for using this system?	3	4	5	4
AVG					**4.42**

In addition to the structured survey questions, participants offered valuable insights through open-ended feedback, which shed light on both the strengths of the system and potential areas for improvement. Participants recognized the system's ability to effectively predict employees' career paths, which they believe empowers individuals to strategically plan their careers and manage associated risks, thereby enhancing their ability to achieve professional objectives within the organization.

Regarding areas for improvement, participants recommended expanding the system's factors beyond the current indicators of education, grade, and job history. They suggested incorporating employee aspirations and adding a technical training component to the features. These enhancements, they believe, would further refine the system's predictive accuracy and ensure that it aligns more closely with the nuanced career development needs of employees.

Chapter 8

Discussion of Findings

This chapter provides a comprehensive discussion of the findings from the study, focusing on addressing the key research questions explored throughout the research. Section 8.1 specifically answers sub-research questions 5, 6, 7, 8, and 9, drawing conclusions from the development and validation of the proposed framework. Sub-research questions 1 through 4 are not discussed in this chapter because they were thoroughly addressed in Chapter 2, where the theoretical foundation and contextual background of the study were established. Section 8.2 examines the study's limitations and offers directions for future research, providing insights into areas where further exploration is needed.

8.1 Answers to the Research Questions

8.1.1 What functional and non-functional requirements need to be satisfied by the proposed framework? (RQ-5)

We identified and defined the essential functional and non-functional requirements needed to ensure the system's effectiveness and reliability. These requirements were primarily shaped by insights gained from qualitative interviews with stakeholders in a real-world government organization, ensuring they align with practical needs and expectations. The broader context and understanding of career path prediction, derived from the Systematic Literature Review, informed the overall design and helped ensure that the requirements were grounded in current research trends and best practices.

The functional requirements, such as data collection and preprocessing (FR1), integration of machine learning algorithms (FR2), model assessment and validation (FR3), real-time career path prediction (FR4), system interoperability (FR5), and endpoints for various predictions (FR6), were derived from a thorough understanding of the business needs and objectives. These requirements focus on ensuring that the framework can handle the entire data pipeline—from collecting and cleaning data to applying machine learning models and generating real-time predictions. Achieving these functional requirements involves integrating robust data preprocessing techniques, selecting suitable machine learning algorithms, implementing validation metrics, and ensuring compatibility with existing IT infrastructure.

On the non-functional side, the framework emphasizes performance (NFR1), scalability (NFR2), interpretability (NFR3), ease of maintenance (NFR4), high availability and reliability (NFR5), compliance with OAuth 2.0 for authorization (NFR6), adherence to data privacy and security regulations (NFR7), and non-discrimination (NFR8). These non-functional requirements ensure the system is not only efficient and capable of handling increasing data volumes but also user-friendly, secure, and equitable. While the proposed design and preliminary models suggest these requirements can be met, their true effectiveness will be assessed post-implementation. This forward-looking approach allows for iterative improvements based on real-world performance and feedback.

However, achieving all these requirements fully can only be validated after the framework's complete implementation and testing phases. Since this research is focused on the development of the framework itself, with the prototype being one of its outputs, some requirements, for example, those related to real-time performance and system interoperability, are based on anticipated capabilities and established best practices within the field.

8.1.2 What framework is suitable for implementing the career path prediction of employees in a public organization in Southeast Asia? (RQ-6)

The central contribution of this research is the framework for career path prediction implemented in Chapters 5 and 6. Our empirical evaluation indicated that this novel framework is well-suited for the context of employees in a public organization in Indonesia. The tailored approach of this framework, from data preparation to deployment, ensures that it addresses the unique requirements and complexities of career progression within Indonesian public sector organizations. This makes the framework highly suitable for this context, as it is not only technically robust but also contextually relevant to the specific needs of civil servants in Indonesia.

A unique aspect of this framework is the detailed mapping of positions and educational backgrounds to their respective fields. This step ensures that the data accurately reflects the diverse competencies and qualifications relevant to various roles. By categorizing and aligning the positions and educational backgrounds, the framework can better capture the nuances required for accurate predictions. This mapping allows the model to incorporate a wide range of attributes and skills, thereby enhancing the predictive power and relevance of the results.

Another novel feature is the determination of position equivalences. For managerial positions, the framework checks for equivalences to ensure that individuals qualified for one role are appropriately considered for similar roles that require similar competencies. This approach addresses the complexity and diversity of managerial positions, ensuring a more holistic and accurate prediction. By doing so, the framework recognizes the overlapping skill sets required for different managerial roles and ensures that potential candidates are not overlooked due to rigid role definitions.

Leveraging machine learning to predict career paths proves highly effective in handling complex patterns and relationships within the data. However, while the framework demonstrates high potential for aiding in the strategic management and development of civil servants' career paths, the performance, especially for managerial positions, is not as strong as desired. The models show better performance for functional positions, with higher accuracy, precision, recall, and F1 scores, indicating their robustness and reliability in these contexts. For managerial roles, decisions regarding these positions are often shaped by subjective judgments and organizational dynamics, which are challenging to encapsulate within algorithmic models. These political and personal factors introduce variability and biases into the training data, making the prediction process more complex and less consistent.

8.1.3 Which machine learning algorithm demonstrates the best predictive performance in determining the next career path of employees in a public organization in Southeast Asia? (RQ-7)

A comprehensive evaluation of various models was conducted to identify the best machine learning algorithm for predicting the next career path. The models were assessed based on key metrics, including accuracy, recall, precision, and F1-score for both functional and managerial positions (Section 6.3). For functional positions, the XGBoost algorithm emerged as the best performer. XGBoost demonstrated superior predictive accuracy, the highest recall, precision, and F1 score. These results indicate that XGBoost is particularly effective at identifying the correct career path for functional positions while maintaining a balance between false positives and false negatives. The algorithm's advanced boosting techniques set it apart, making it the most reliable choice for functional position predictions in this context.

In contrast, the evaluation for managerial positions yielded different results. While the Random Forest algorithm achieved the highest accuracy and the highest recall, it was the Multilayer Perceptron (MLP) that showed superior performance in terms of precision and F1 score. MLP slightly outperformed Random Forest's precision and F1-score. This indicates that MLP is more effective in minimizing false positives and providing a balanced predictive performance for managerial positions despite Random Forest's higher accuracy. The discrepancy in performance between functional and managerial positions highlights the importance of context-specific model selection. For functional positions, XGBoost's advanced boosting mechanism provides a distinct advantage. Conversely, for managerial positions, the MLP's ability to capture intricate patterns through its neural network structure proves to be more beneficial in achieving higher precision and a balanced F1 score.

8.1.4 Which features contribute the most to the predicting results? (RQ-8)

The comparative analysis of feature importance (Subsection 6.3.2) reveals that while both managerial and functional positions value career history and education, there are notable differences in other key predictors. For managerial positions, the current role of a candidate is the most influential factor, suggesting that individuals currently in significant roles are more likely to be suitable for managerial positions. In contrast, for functional positions, hierarchical rank and specific technical expertise are more critical, reflecting the importance of depth of knowledge and specialized skills.

Furthermore, the tenure in specific ranks (e.g., "rank_start_date") is more influential for managerial positions, highlighting the value of prolonged experience and stability in leadership roles. For functional positions, the duration of service in relevant roles (e.g., "month_service" and "position_month_service") is emphasized, indicating the importance of sustained performance and experience in technical areas. Education plays a significant role in both managerial and functional positions, although the type of education may differ. Managerial positions might value broader management or leadership-focused education, while functional positions might prioritize specialized technical or domain-specific education.

8.1.5 To what extent is the proposed framework considered useful and usable by practitioners in the field? (RQ-9)

The validation results demonstrate that the proposed framework is generally perceived as highly useful by the practitioners. The proposed framework has been recognized by practitioners as highly valuable, particularly for its potential to significantly enhance career planning and decision-making processes within organizations. The system's prototype has garnered positive feedback regarding its anticipated usability, with practitioners expressing confidence that, once fully developed, the framework will be both user-friendly and accessible. This positive outlook reflects the system's intuitive design and its potential to streamline career-related strategies, making it a promising tool for workforce management.

However, the findings also underscore the importance of continued development, particularly in refining the predictive accuracy and ensuring robust technical support. While the concept is strong, these areas need further attention to fully realize the framework's potential in practical settings. Addressing these critical aspects will be essential to align the framework with the needs and expectations of its users, thereby ensuring its successful implementation and adoption in the field. Enhancements in these areas will not only bolster the system's effectiveness but also foster greater trust and reliance among its users.

8.2 Study Limitations and Direction for Future Research

8.2.1 Study Limitations

Despite the promising results, this study has several limitations [19, 69] that must be acknowledged. Firstly, the data used for model training and validation was sourced from civil servants within a single organization in the Southeast Asian public sector, most specifically in Indonesia. This limitation may affect the generalizability of the findings to other organizations or different public sector environments, and the dataset size was limited, which could impact the robustness of the model.

Secondly, the dataset lacked comprehensive information on employee performance and competencies, which are crucial factors for accurately predicting career paths, especially for structural or managerial positions. The absence of this data could have constrained the models' ability to fully capture the nuances of career progression.

Third, the study was limited to predicting the next career move rather than mapping out the entire career path to the highest positions. This focus on the immediate next step does not account for the long-term career trajectories and potential career milestones an employee might achieve.

Furthermore, we acknowledge the limitation inherent to the use of qualitative interviews as a research method [32]. In Chapter 4, we report the requirements that came out of interviewing six stakeholders. The central question concerning qualitative interviews is whether the findings would be the same if we had interviewed different individuals in the same organization. Of course, as methodologies suggest, in qualitative studies it is hard to claim the generalizability of the findings. However, it might be theoretically possible to observe the transferability of our findings to similar but different contexts: following Ghaisas et al [19], if we had interviewed different individuals in similar roles, capacities, and jobs, working in similar departments, to those where our interviewees were, it might be possible to obtain similar observations as those of our participants. This is because the similarity of participants' contexts is likely to create similar organizational mechanisms and circumstances as those of our participants [19]. Furthermore, we chose our participants for their typicality and in an effort to ensure a diversity of perspectives. Therefore, as per the argumentation of research methodologists, we think that the requirements that we elicited based on the interviews might be relevant also for organizations beyond the context in which these emerged. For example, the functional and non-functional requirements might well be relevant for any organization in Southeast Asia that collects career-related data in a systematic way, that wishes to make their career path knowledge transparent for civil servants and that is interested in implementing data-mining-based solutions for career path prediction and is committed to leverage the outcome of career prediction for their improved human resource management and management of talent.

8.2.2 Direction for Future Research

Based on the identified limitations, several recommendations can be made to enhance the predictive capabilities and applicability of the career path prediction framework. Firstly, future research should aim to incorporate data from multiple organizations within the Southeast Asian public sector and beyond. This will be instrumental in improving the generalizability of the results in this thesis. One line for future research is to expand the dataset to include a larger and more diverse sample of civil servants will improve the generalizability of the findings and increase the robustness of the models. This broader dataset will help in understanding different organizational contexts and improve the overall accuracy of career path predictions. Another line of research to improve generalizability is to carry out case study research in other organizations and their stakeholders in order to possibly refine the framework based on a more elaborate set of requirements.

Secondly, it is essential to include comprehensive information on employee performance and competencies in the datasets. Future studies should focus on collecting and integrating detailed performance metrics and competency profiles, as these factors are crucial for accurately predicting career paths, especially for structural or managerial positions. Incorporating this data will allow the models to better capture the complexities and nuances of career progression, leading to more precise and insightful predictions.

Finally, future research should extend beyond predicting the next immediate career move to mapping out entire career paths up to the highest positions. Developing models capable of forecasting long-term career trajectories and identifying potential career milestones will provide more valuable insights for career planning and development. This approach can help organizations better understand and support the long-term growth of their employees, ensuring more strategic and effective human resource management. By addressing these recommendations, future research can build on the foundation laid by this study, advancing the field of career path prediction and providing more comprehensive and actionable insights for the public sector.

Chapter 9

Conclusion

9.1 Conclusion

This thesis offers a comprehensive approach to predicting career advancements within Southeast Asia's public sector through the use of machine learning. The research was initiated by conducting a thorough examination of the current state-of-the-art methodologies in career path prediction. This review identified the latest tools and techniques used in the field, highlighting existing gaps and opportunities for improvement.

The exploratory methodology employed in this thesis followed the CRISP-DM framework, ensuring a structured and systematic approach to data mining and analysis. Several machine learning algorithms were meticulously assessed based on predefined criteria, ensuring the selection of the most suitable algorithms for predicting career paths, taking into account the unique dynamics and complexities of career progression within the public sector.

The functional and non-functional requirements of the proposed framework were carefully defined to ensure both high performance and practical usability. These requirements were essential in developing a robust and scalable system that can be effectively implemented in real-world settings.

The chosen machine learning models were implemented through the development of a prototype system. This practical application highlights the model's utility in providing clear guidance on the steps required for career advancement and the qualifications needed. The insights gained from the model can significantly enhance job satisfaction by aligning employees' career aspirations with organizational goals, reducing turnover, and fostering a more motivated and engaged workforce.

A crucial aspect of this research involved the identification and analysis of feature importance using SHAP values. The feature importance analysis revealed that factors such as position history, education, rank, and current position are the most influential in predicting managerial positions. Understanding these key drivers provides valuable insights into the critical factors contributing to career advancement, allowing for more targeted and effective career development strategies within the public sector. The detailed analysis of feature importance not only aids in model interpretability but also helps stakeholders in making informed decisions to support employee growth and organizational success.

Lastly, the usability of the proposed framework was rigorously assessed through expert validation involving key stakeholders, including HR professionals, software engineers, and civil servants. Feedback from participants highlighted the system's strong potential for supporting career planning and aligning employee aspirations with organizational goals, while also identifying areas for further enhancement, such as expanding predictive criteria and incorporating technical training components. Overall, this validation confirmed that the framework offers a practical and scalable tool for career path prediction within the public sector, with the potential to significantly enhance career development strategies and contribute meaningfully to organizational success.

9.2 Study Contributions

This thesis makes several important contributions. Firstly, it fills a critical gap identified in the state-of-the-art scientific research on the topic of career path prediction. Unlike prior publications that exploit publicly available data from social network platforms, this thesis focused on the application of predictive models to real-world cases of civil servants. This is a domain that has been relatively under-explored compared to career path predictions based on data from online platforms like LinkedIn or student datasets. By developing tailored predictive models for civil service careers, this research provides insights that are directly applicable to the unique challenges and dynamics of public sector careers. This focus helps optimize workforce planning, enhance talent management strategies, and inform policy decisions, thereby contributing to the overall efficiency and effectiveness of public service organizations.

Moreover, this thesis contributes to the global conversation in the community of researchers in Software Engineering and Information Systems on career path prediction by providing a focused analysis within the Southeast Asian context. While existing literature indicates a diverse range of countries contributing to this field, there has been a noticeable gap in region-specific studies that consider the cultural and contextual nuances influencing career trajectories. By incorporating data from a Southeast Asian public sector organization, this research enriches the collective knowledge base with perspectives and methodologies from an underrepresented region, fostering a more comprehensive understanding of career trajectory prediction.

This thesis also makes a contribution to the field of Requirements Engineering (RE) for ML-based systems by describing the application of exploratory and experimental RE to a real-world case. The thesis responded to the call of RE researchers for more empirical work in RE for ML-based systems [43]. Moreover, this research addresses a business application (i.e., a system for career path prediction), a type of application that has been relatively underexplored in RE. As indicated in Chapter 1, most published cases in the literature focus on cyber-physical systems rather than business information systems. From this perspective, the thesis provides a framework and its exemplified application in a specific context (i.e., government and career path prediction), contributing new knowledge to the field.

Finally, this thesis introduces a new framework for the implementation of career path prediction in public sector organizations. The prototype developed based on this framework demonstrates its technical feasibility and suitability for the Indonesian context, exemplifying a Southeast Asian country. This framework is particularly noteworthy as most cases in the literature have focused on cyber-physical systems, not business information systems. Researchers could use the framework as a theoretical model in future empirical evaluation studies. It could serve as a starting point for experimenting with its refinement or extension, generating empirical evidence that is critical for understanding and improving its generalizability.

9.3 Practical Implications

In practical terms, this thesis demonstrates the feasibility and benefits of integrating machine learning into human resource management systems within the public sector. The prototype developed as part of this research serves as a proof of concept, illustrating how data-driven approaches can be used to inform decision-making processes.

By forecasting the next career moves of civil servants, the framework aids employers in optimizing workforce planning and allocation. This can lead to more strategic deployment of human resources, ensuring that employees are placed in roles where they can be most effective, thereby enhancing overall organizational efficiency.

From the civil servants' perspective, the model also helps employees understand the steps required for advancement and the qualifications needed. This clarity can significantly enhance job satisfaction by aligning employees' career aspirations with organizational goals, reducing turnover, and fostering a more motivated and engaged workforce. By addressing these practical needs from both the employer's and employee's perspectives, this thesis contributes to a more efficient and effective public sector workforce management.

Bibliography

[1] K. Ahmad, M. Abdelrazek, C. Arora, M. Bano, and J. Grundy. Requirements engineering for artificial intelligence systems: A systematic mapping study. Information and Software Technology. 180, 2023. doi:10.48550/arXiv.2212.10693.

[2] I. F. Alexander. A Better Fit - Characterising the Stakeholders. 2004.

[3] Data Science Process Alliance. *Evaluating CRISP-DM for data science*. Data Science Process Alliance, 2021.

[4] S. Amershi, A. Begel, C. Bird, R. DeLine, H. Gall, E. Kamar, N. Nagappan, B. Nushi, and T. Zimmermann. Software engineering for machine learning: A case study. In *IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pages 291–300, 2019. doi:10.1109/ICSE-SEIP.2019.00042.

[5] R. P. Archana, S. M. Anzar, and N. P. Subheesh. Career Recommender System Using Decision Trees. In *2023 IEEE Global Engineering Education Conference (EDUCON)*, pages 1–5, 2023. doi:10.1109/EDUCON54358.2023.10125277.

[6] A. P. S. A;ves, M. Kalinowski, G. Giray, D. Mendez, N. Lavesson, K. Azevedo, H. Villamizar, T. Escovedo, H. Lopes, and S. Biffl. Status Quo and Problems of Requirements Engineering for Machine Learning: Results from an International Survey. In *International Conference on Product-Focused Software Process Improvement*, pages 159–174, 2023. doi:10.48550/arXiv.2310.06726.

[7] L. Breiman. Bagging Predictors. pages 123–140, 1996. doi:10.1023/A:1018054314350.

[8] L. Breiman. Random Forests. pages 5–32, 2001. doi:10.1023/A:1010933404324.

[9] G. Casu, M. G. Mariani, R. Chiesa, D. Guglielmi, and P. Gremigni. The Role of Organizational Citizenship Behavior and Gender between Job Satisfaction and Task Performance. 2021. doi:10.3390/ijerph18189499.

[10] J. Chakraborty, S. Majumder, and T. Menzies. Bias in Machine Learning Software: Why? How? What to do? pages 429–440, 2021. doi:10.1145/3468264.3468537.

[11] T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. pages 785–794, 2016. doi:10.1145/2939672.2939785.

[12] M. Daneva and R. J. Wieringa. Requirements engineering for cross-organizational ERP implementation: Undocumented assumptions and potential mismatches. In *Proceedings of the 13th IEEE International Conference on Requirements Engineering, RE 2005*, pages 63–74, 2005. doi:10.1109/RE.2005.59.

[13] M. Daneva and R. J. Wieringa. Requirements Engineering for Enterprise Systems: What We Know and What We Don't Know. In *Intentional Perspectives on Information Systems Engineering*, pages 115–136, 2010. doi:10.1007/978-3-642-12544-7_7.

[14] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw. User Acceptance of Computer Technology: A Comparison of Two Theoretical Models. 35:982–1003, 1989. doi:10.2307/249008.

[15] J.-J. Decorte, J. Van Hautte, J. Deleu, C. Develder, and T. Demeester. Career Path Prediction using Resume Representation Learning and Skill-based Matching. In *CEUR Workshop Proceedings*, 2023.

[16] S. Farooqui and A. Nagendra. The Impact of Person Organization Fit on Job Satisfaction and Performance of the Employees. 11:122—129, 2014. doi:10.1016/S2212-5671(14)00182-8.

[17] J. H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. pages 1189–1232, 2001. doi:10.1214/aos/1013203451.

[18] F. A. Gers and E. Schmidhubers. LSTM Recurrent Networks Learn Simple Context-Free and Context-Sensitive Languages. 12:1333–1340, 2001. doi:10.1109/72.963769.

[19] S. Ghaisas, P. Rose, M. Daneva, K. Sikkel, and R. J. Wieringa. Generalizing by Similarity: Lessons Learnt from Industrial Case Studies. pages 37–42, 2013. doi:10.1109/CESI.2013.6618468.

[20] A. Ghosh, B. Woolf, S. Zilberstein, and A. Lan. Skill-based Career Path Modeling and Recommendation. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1156–1165, 2020. doi:10.1109/BigData50022.2020.9377992.

[21] G. Giray. A software engineering perspective on engineering machine learning systems: State of the art and challenges. 2021. doi:10.1016/j.jss.2021.111031.

[22] A. Gjorgjevikj, K. Mishev, L. Antovski, and T. Trajanov. Requirements Engineering in Machine Learning Projects. *IEEE Access*, 11:72186–72208, 2023. doi:10.1109/ACCESS.2023.3294840.

[23] I. Guyon and A. Elisseeff. *An Introduction to Variable and Feature Selection*. Journal of Machine Learning Research, 2003.

[24] B. Harsha, N. Sravanthi, N. Sankeerthana, and M. Suneetha. Career Choice Using Machine Learning Algorithms. In *5th International Conference on Inventive Computation Technologies, ICICT 2022 - Proceedings*, pages 171–176, 2022. doi:10.1109/ICICT54344.2022.9850697.

[25] M. He, D. Shen, Y. Zhu, R. He, T. Wang, and Z. Zhang. Career Trajectory Prediction based on CNN. In *Proceedings - IEEE International Conference on Service Operations and Logistics, and Informatics 2019, SOLI 2019*, pages 22–26, 2019. doi:10.1109/SOLI48380.2019.8955009.

[26] M. He, X. Zhan, D. Shen, Y. Zhu, H. Zhao, and R. He. What about Your Next Job? Predicting Professional Career Trajectory Using Neural Networks. In *ACM International Conference Proceeding Series*, pages 184–189, 2021. doi:10.1145/3490725.3490753.

[27] A. José-García, A. Sneyd, A. Melro, A. Ollagnier, and G. Tarling. C3-IoC: A Career Guidance System for Assessing Student Skills using Machine Learning and Network Visualisation. In *International Journal of Artificial Intelligence in Education*, 2022. doi:10.1007/s40593-022-00317-y.

[28] C. Kaestner. The world and the machine and responsible machine learning. https://ckaestne.medium.com/the-world-and-the-machine-and-responsible-machine-learning-1ae72353c5ae, 2020. Online; Accessed: 2024-08-02.

[29] A. Kamal, B. Naushad, H. Rafiq, and S. Tahzeeb. Smart Career Guidance System. In *Proceedings - 2021 IEEE 4th International Conference on Computing and Information Sciences, ICCIS 2021*, 2021. doi:10.1109/ICCIS54243.2021.9676408.

[30] A. M. Karunia and A. Ika. 35 persen asn di ri kinerjanya rendah, bkn: Seperti "kayu mati" karena malas. https://money.kompas.com/read/2022/07/21/142000926/35-persen-asn-di-ri-kinerjanya-rendah-bkn--seperti-kayu-mati-karena-malas#google_vignette. Online; Accessed: 2023-12-06.

[31] S. Kim. Individual-level Factors and Organizational Performance in Government Organizations. 15:245–261, 2005. doi:10.1093/jopart/mui013.

[32] N. King and C. Horrocks. *Interview in Qualitative Research*. SAGE Publications, 2010.

[33] B. Kitchenham and S. Charters. *Guidelines for performing Systematic Literature Reviews in Software Engineering. 2*. 2007.

[34] H. Kolhe, R. Chaturvedi, S. Chandore, G. Sakarkar, and G. Sharma. Career Path Prediction System Using Supervised Learning Based on Users' Profile. In *Computational Intelligence*, pages 583–595, 2023. doi:10.1007/978-981-19-7346-8_50.

[35] V. R. Kumbhar, M. M. Maddel, and Y. Raut. Smart Model For Career Guidance Using Hybrid Deep Learning Technique. In *2023 1st International Conference on Innovations in High Speed Communication and Signal Processing (IHCSP)*, pages 327–331, 2023. doi:10.1109/IHCSP56702.2023.10127152.

[36] S. Lausen. *Software Requirements: Styles and Techniques*. 2001.

[37] Y. Lecun, Y. Bengio, and G. Hinton. Deep Learning. pages 436–444, 2015. doi:10.1038/nature14539.

[38] H. Li, Y. Ge, H. Zhu, H. Xiong, and H. Zhao. Prospecting the career development of talents: A survival analysis perspective. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 917–925, 2017. doi:10.1145/3097983.3098107.

[39] L. Li, J. Yang, H. Jing, Q. He, H. Tong, and B-C. Chen. NEMO: Next career move prediction with contextual embedding. In *26th International World Wide Web Conference 2017, WWW 2017 Companion*, pages 505–513, 2017. doi:10.1145/3041021.3054200.

[40] Y. Liu, L. Zhang, L. Nie, Y. Yan, and D. S. Rosenblum. Fortune teller: predicting your career path. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 201—207, 2016. doi:10.1609/aaai.v30i1.9969.

[41] S. Lundberg and S-I. Lee. A Unified Approach to Interpreting Model Predictions. 2010. doi:10.48550/ARXIV.1705.07874.

[42] L. E. Lwakatare, A. Raj, I. Crnkovic, J. Bosch, and H. H. Olsson. Large-scale machine learning systems in real-world industrial settings: A review of challenges and solutions. 2020. doi:10.1016/j.infsof.2020.106368.

[43] A. Martinez-Fernandez, J. Bogner, X. Franch, M. Oriol, J. Siebert, A. Trendowicz, A. M. Vollmer, and S. Wagner. Software engineering for AI-based systems: a survey. *ACM Transactions on Software Engineering and Methodology (TOSEM)* , 31:1–59, 2022. doi:10.1145/3487043.

[44] F. Martínez-Plumed, L. Contreras-Ochando, C. Ferri, J. Hernández-Orallo, M. Kull, N. Lachiche, M. J. Ramírez-Quintana, and P. Flach. Crisp-dm twenty years later: From data mining processes to data science trajectories. In *IEEE Transactions on Knowledge and Data Engineering 33*, pages 3048–3061, 2021. doi:10.1109/TKDE.2019.2962680.

[45] Q. Meng, H. Zhu, K. Xiao, L. Zhang, and H. Xiong. A Hierarchical Career-Path-Aware Neural Network for Job Mobility Prediction. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 14–24, 2019. doi:10.1145/3292500.3330969.

[46] G. Naidu, T. Zuva, E. M. Sibanda, R. Silhavy, and P. Silhavy. A Review of Evaluation Metrics in Machine Learning Algorithms. pages 5–25, 2023. doi:10.1007/978-3-031-35314-7_2.

[47] Y. Pan, X. Peng, T. Hu, and J. Luo. Understanding what affects career progression using linkedin and twitter data. In *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, pages 2047–2055, 2017. doi:10.1109/BigData.2017.8258151.

[48] J. R. Quinlan. Induction of Decision Tree. *Machine Learning*, pages 81–106, 1986. doi:10.1007/BF00116251.

[49] M. Rane, S. Kalal, J. Chandegara, T. Kakkad, T. Jain, and S. Jagtap. Career Prediction Website using Machine Learning. In *2023 3rd International Conference on Intelligent Technologies, CONIT 2023*, pages 583–595, 2023. doi:10.1109/CONIT59222.2023.10205747.

[50] T. R. Razak, M. A. Hashim, N. M. Noor, I. H. A. Halim, and N. F. F. Shamsul. Career path recommendation system for UiTM Perlis students using fuzzy logic. In *2014 5th International Conference on Intelligent and Advanced Systems (ICIAS)*, pages 1–5, 2014. doi:10.1109/ICIAS.2014.6869553.

[51] N. Roulin and A. Bangerter. Students' use of extra-curricular activities for positional advantage in competitive job markets. 26:21–47, 2013. doi:10.1080/13639080.2011.623122.

[52] M. Roy, A. K. Bhoi, and K. Sharma. Multimodal Machine Learning approaches for Career Prediction. In *2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)*, pages 1–5, 2022. doi:10.1109/ASSIC55218.2022.10088305.

[53] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986. doi:10.1038/323533a0.

[54] L. Rutkowski, M. Jaworski, L. Pietruczuk, and P. Duda. The CART decision tree for mining data streams. pages 1–15, 2014. doi:10.1016/j.ins.2013.12.060.

[55] J. M. Saldana. *The Coding Manual for Qualitative Researchers*. SAGE Publications, 2015.

[56] A. Shankhdhar, A. Gupta, A. K. Singh, and R. Pradhan. A Vocational Career Advisory Application Built Using Unsupervised Machine Learning Frameworks. 836:471–480, 2022. doi:10.1007/978-981-16-8542-2_38.

[57] D. Sobnath, T. Kaduk, I. U. Rehman, and O. Isiaq. Feature Selection for UK Disabled Students' Engagement Post Higher Education: A Machine Learning Approach for a Predictive Employment Model. pages 159530–159541, 2020. doi:10.1109/ACCESS.2020.3018663.

[58] M. Sodanil, S. Chotirat, L. Poomhiran, and K. Viriyapant. Guideline for Academic Support of Student Career Path Using Mining Algorithm. In *Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval*, pages 133—-137, 2019. doi:10.1145/3342827.3342841.

[59] Z. Soliman, P. Langlais, and L. Bourg. Learning Career Progression by Mining Social Media Profiles. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 446–452, 2019. doi:10.1007/978-3-030-18305-9_43.

[60] A. Thomas, A. K. Varghese, P. L. Alex, B. J. Mathews, and L. K. Dhanya. Analysis of Machine Learning Algorithms for Predicting the Suitable Career After High School. In *Proceedings of Third International Conference on Communication, Computing and Electronics Systems*, pages 89–1054, 2022. doi:10.1007/978-981-16-8862-1_7.

[61] TimPublikasiPanselnas. Ditutup 11 oktober, pelamar seleksi casn 2023 mencapai 2.409.882. https://www.bkn.go.id/ditutup-11-oktober-pelamar-seleksi-casn-2023-mencapai-2-409-882/. Online; Accessed: 2023-12-06.

[62] TimPublikasiPanselnas. Mulai 09 november, 1.853.617 pelamar casn 2023 berkompetisi pada skd dan seleksi kompetensi. https://www.bkn.go.id/mulai-09-november-1-853-617-pelamar-casn-2023-berkompetisi-pada-skd-dan-seleksi-kompetensi-2/. Online; Accessed: 2023-12-06.

[63] T. D. Truong, N. Q. H. Ton, C. N. Truong, and S. P. Nguyen. ECPASurv: Exploring Career Path with Attention Mechanism and Survival Analysis. In *2022 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 136–139, 2022. doi:10.1109/RIVF55975.2022.10013830.

[64] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis. User Acceptance of Information Technology: Toward a Unified View. 27:425–478, 2003. doi:10.2307/30036540.

[65] P. Verma, S. K. Sood, and S. Kalra. Student career path recommendation in engineering stream based on three-dimensional model. In *Computer Applications in Engineering Education*, pages 578–593, 2017. doi:10.1002/cae.21822.

[66] S. Vignesh, C. Shivani Priyanka, H. Shree Manju, and K. Mythili. An Intelligent Career Guidance System using Machine Learning. In *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 987–990, 2021. doi:10.1109/ICACCS51430.2021.9441978.

[67] H. Villamizar, T. Escovedo, and M. Kalinowski. Requirements engineering for machine learning: A systematic mapping study. In *2021 47th Euromicro Conference on Software Engineering and Advanced Applications*, pages 29–36, 2021. doi:10.48550/arXiv.2212.10693.

[68] A. Vogelsang and M. Borg. Requirements engineering for machine learning: Perspectives from data scientists. In *IEEE 27th International Requirements Engineering Conference Workshops (REW)*, pages 245–251, 2019. doi:10.1109/REW.2019.00050.

[69] C. Wohlin. Experimentation in Software Engineering. 2000. doi:10.1007/978-1-4615-4625-2.

[70] S. Wood, M. van Veldhoven, M. Croon, and L.M. de Menezes. Enriched job design, high involvement management and organizational performance: The mediating roles of job satisfaction and well-being. 65:419–445, 2016. doi:10.1177/0018726711432476.

[71] A. K. Yadav, A. Dixit, A. Tripathi, S. K. Chowdhary, and V. Jangra. Career Prediction System using ANN MLP Classifier. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7, 2023. doi:10.1109/ICCCNT56998.2023.10307057.

[72] J. Ćulibrk, M. Delić, S. Mitrović, and D. Ćulibrk. Job Satisfaction, Organizational Commitment and Job Involvement: The Mediating Role of Job Involvement. 2018. doi:10.3389/fpsyg.2018.00132.

Appendix A

Quantitative and Qualitative Analysis of the Systematic Literature Review

TABLE A.1: Quantitative Analysis of the Literature

Literature	**Machine Learning**	**Source of Data and Features**	**Data Type**	**Implementation**
R. P. Archana, S. M. Anzar, and N. P. Subheesh (2023) [5]	G-CAPS, Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM) and linear regression	The **Handwriting Features** datasets, explaining **Psychological Behavior** and **Holland Code Personality Traits**, were collected from engineering students and working professionals holding engineering degrees.	Unstructured	Fig.1 shows the G-CAPS System Architecture
A. Ghosh, B. Woolf, S. Zilberstein, and A. Lan. (2020) [20]	Monotonic Nonlinear State-Space (MNSS), RNN, Job2Vec, HRM, NEMO	The datasets consist of **educational and professional experiences**, along with their **skill sets**, collected from LinkedIn and Indeed. Table 1 lists the number of user profiles, experiences, unique skills, **companies**, and **job titles** for both datasets.	Semi-structured	Fig. 1 depicts the structure of the MNSS model

B. Harsha, N. Sravanthi, N. Sankeerthana, and M. Suneetha (2022) [24]	DT, RF, SVM	Data were obtained from numerous educational institutions, including **marks in SSC, percentages in mathematics, physics, and chemistry, intermediate marks, capabilities such as reading and writing scores, gender, and category**.	Structured	Fig 1. illustrates the architecture of the proposed method
M. He, D. Shen, Y. Zhu, R. He, T. Wang, and Z. Zhang (2019) [25]	Convolutional Neural Network (CNN)	The dataset was downloaded from a competition game of **DataCastle** in China. The base information includes **ID number, gender, age, major, and degree**. The employment history comprises multiple pieces of **work experience**, each detailing **start time, end time, industry, company scale, salary, position name, department, and job type**.	Structured	The CNN model's architecture is indicated in Fig. 5
M. He, X. Zhan, D. Shen, Y. Zhu, H. Zhao, and R. He (2021) [26]	CNN and Long Short-Term Memory (LSTM) network, Multi-Layer Perceptron (MLP). Logistical Regression (LR), RF, Gradient Boosting Decision Tree(GBDT), Extreme Gradient Boosting(XGBoost).	The dataset was provided by **DataCastle** and consists of Job seekers' **personal information** and **employment history**.	Structured	Figure 1 shows the architecture of the proposed model based on CNN and LSTM

A. José-García, A. Sneyd, A. Melro, A. Ollagnier, G. Tarling, H. Zhang, M. Stevenson, R. Everson, R. Arthur (2022) [27]	C3-IoC	Users can upload their CVs in PDF format and a list of **technical skills** is extracted from it. For the **non-technical skills**, users need to fill out a questionnaire with 24 questions. Authors also used datasets from O*NET and Find a Job	Semi-structured	Fig. 2 depicts the C3-IoC system architecture, comprising three main modules. Fig. 3 shows the screenshot of the C3-IoC system
A. Kamal, B. Naushad, H. Rafiq, S. Tahzeeb (2021) [29]	XGBoost and RF	The dataset is derived from the answers provided by questionnaire respondents to 80 questions. It consists of **personality traits, domains of interest, academic information, personal details such as name, age, gender, qualifications, favorite subjects during school years, and details about their careers.**	Structured	Figure 1 shows the Flow Chart of Smart Career Guidance System
H. Kolhe, R. Chaturvedi, S. Chandore, G. Sakarkar, and G. Sharma (2023) [34]	XGBoost, DT, RF, and SVM.	The dataset was manually created and includes academic and non-academic areas which are numeric (e.g., **logical quotient rating, hackathons, coding skills rating, and public speaking points**) and categorical columns (e.g., **extra-courses did and certifications**)	Structured	Fig. 2 illustrates the flowchart of the whole process, Fig. 4-8 shows the GUI snapshot of the system
V. R. Kumbhar, M. M. Maddel, and Y. Raut. (2023) [35]	Recurrent Neural Network (RNN), LSTM, DT, RF, NLP	A dataset of questions and possible answers of student's **interests** used in career counseling	Semi-structured	Fig. 1 illustrates the approach used for Career Guidance Using Hybrid Deep Learning Techniques

H. Li, Y. Ge, H. Zhu, H. Xiong, and H. Zhao (2017) [38]	COX, Log-logistic, Log-gaussian, Weibull, Exponential, M-LASSO, M-L2,1, MTLSAV2, MTLSA, DF, CDT+SE, and CDT+DE	The dataset is a set of anonymized employee career records consisting of static profile information (**gender, age, year of start date, month of start date, initial level, and initial subordinate number**) and dynamic information, such as **performance rating, number of changed superiors, and number of changed subordinates**, from a high-tech company across a time span of 48 months from January 1st, 2011, to December 31st, 2014.	Structured	-
L. Li, J. Yang, H. Jing, Q. He, H. Tong, and B-C. Chen (2017) [39]	LSTM	The data is **skill, company, title, school, and location** of LinkedIn members.	Semi-structured	Figure 2. depicts the framework of NEMO
Y. Liu, L. Zhang, L. Nie, Y. Yan, and D. S. Rosenblum (2016) [40]	Multi-source Learning Framework with a Fused Lasso penalty MSLFL, the regularized Multi-view Multi-task learning model (regMVMT), SVM, Regularized Least Square (RLS), Multi-task Learning (MTL), and Fused Lasso (FL)	The dataset, comprising demographic characteristics (e.g., **education and participation**), **LIWC features**, and **user topic features**, was created by crawling data from About.me and obtaining data from social media platforms such as LinkedIn, Facebook, and Twitter	Unstructured	-

Q. Meng, H. Zhu, K. Xiao, L. Zhang, and H. Xiong (2019) [45]	Non-sequential models (e.g., LR, RF, and DT), sequential models (e.g., Conditional Random Field (CRF), Continuous Time Markov Chain (CTMC)) and the stochastic time series models (e.g., Poisson Process (PP) Multi-variable Hawkes Process (MHP)), Hierarchical Career-Path-aware Neural Network (HCPNN), and two modified versions of HCPNN (HCPOP and HCPOS)	The data were collected from a famous online professional social platform and consist of personal-specific features (static and includes freely structured **self-description** texts and the **number of social connections**), company-specific features (e.g., **job duration at the company, company personnel flow, company description, company name, type, size, location, age**), and position-specific features (e.g., **position type, service duration**, etc.).	Unstructured	The graphical representation of the HCPNN model is depicted in Figure 4. Figure 5. demonstrates the process of predicting the next employer and Figure 6. illustrates the process of predicting job duration
Y. Pan, X. Peng, T. Hu, and J. Luo. (2017) [47]	Support Vector Regression (SVR) and Ensemble Tree Learner	The dataset utilized in this paper is a valid dataset provided by the author of [3], who collected this data from about.me. It consists of measurements for **Personal Traits (LIWC)**, partitioning of Career Stages (**job level, job title**), and formulation of Career Progression (**duration**).	Structured	-
M. Rane, S. Kalal, J. Chandegara, T. Kakkad, T. Jain, and S. Jagtap (2023) [49]	DT and K-Nearest Neighbors (KNN)	The dataset for the project is in the form of answers to a standard questionnaire developed with various technical and psychological factors in mind. It contains multiple choices from various people regarding their **field of interest.**	Structured	-

T. R. Razak, M. A. Hashim, N. M. Noor, I. H. A. Halim, and N. F. F. Shamsul (2014) [50]	Fuzzy Logic (FL)	The data are obtained by conducting a series of interviews with the domain expert and consist of **personality and skills**.	Structured	Figure 1 illustrates the entire process of implementing the Career Path Selection Recommendation System (CPSRS).
M. Roy, A. K. Bhoi, and K. Sharma (2022) [52]	ADABOOST, SVM, RF, and DT	Data is collected using a standard questionnaire (Table 1) from the students of a technical institute. It consists of **gender, place of birth and current residence, parents' literacy status, family financial status, percentage in PCM in the 12th board examination, medium of primary education (English/Hindi/Bengali), community affiliation, literacy rate or number of graduates in the family, participation in extracurricular activities, and interest in technical or management subjects**.	Structured	-
A. Shankhdhar, A. Gupta, A. K. Singh, and R. Pradhan (2022) [56]	SVM and DT	Data is collected from delegates in various affiliations, through the LinkedIn API, made subjectively, and from school alumni database. It includes **academic scores**, **specializations**, **programming and analytical skills**, **memory**, and **personal details** like relationships, interests, sports, hackathons, workshops, certifications, and favorite books.	Semi-structured	-

D. Sobnath, T. Kaduk, I. U. Rehman, and O. Isiaq (2020) [57]	Logistic Regression, Linear Discriminant Decision Tree, and Gaussian Naive Bayes. Feature selection techniques: Random Forest and Extra Tree Classifier	The data was requested from the Higher Education Statistics Agency (HESA). It consists of **age, HE institution, level of DLHE qualification, class of the first degree, disability type, highest qualification on entry, and JACS code**.	Structured	-
M. Sodanil, S. Chotirat, L. Poomhiran, and K. Viriyapant (2019) [58]	Apriori algorithm	Data were collected from the registration and assessment system and the education service division. It consists of students' **grades** from 25 main courses in the field of information technology.	Structured	-
Z. Soliman, P. Langlais, and L. Bourg. (2019) [59]	PreLa baseline, Naïve Bayes, N-gram model, RNN-based model	The **job history** dataset was collected from LinkedIn.	Semi-structured	-
A. Thomas, A. K. Varghese, P. L. Alex, B. J. Mathews, and L. K. Dhanya (2022) [60]	Naïve Bayes, DT, Logistic Regression, ANN, KNN, AdaBoost, MLP, SVM with RBF kernel	The dataset was collected using a Google form. Some data came from working individuals across various organizations, some from a platform called SurveyTandem, and some were generated randomly or synthetically. Features include **IQ, high school academic scores, EQ using O.C.E.A.N test, memory, career satisfaction, interests, certifications, leadership, self-learning ability, teamwork skills, socioeconomic factors, work-life balance, and eligibility for certifications/exams**	Structured	-

T. D. Truong, N. Q. H. Ton, C. N. Truong, and S. P. Nguyen (2022) [63]	Exploring Career Path with Attention Mechanism and Survival Analysis (ECPASurv), RF, LR, LSTM, NEMO	More than 10,000 employee profiles in the IT domain were collected from a well-known online professional social platform to build the experimental dataset with the support of Neurond Tech JSC. The dataset includes **employee experience and skill sets**.	Semi-structured	Fig. 2 shows the overview of the proposed model
P. Verma, S. K. Sood, and S. Kalra (2017) [65]	Analysis Hierarchical Process (AHP), FL	The data were obtained from students' career-related skill questionnaires and consist of **personality and skill, parental advice, and other factors (career, income, education), family pressure, pursuing the same career as seniors, mentorship, and friends' approval/advice**.	Structured	Figure 1 depicts the proposed career path recommendation model.
S. Vignesh, C. Shivani Priyanka, H. Shree Manju, and K. Mythil (2021) [66]	KNN, SVM, Naïve Bayes	The dataset was developed manually and consists of **core skills** and **sub-skills**, such as analytical skills, logical reasoning skills, mathematical skills, problem-solving skills, programming skills, creativity skills, and hardware skills.	Structured	Figure 1 illustrated architecture diagram of the recommender system
A. K. Yadav, A. Dixit, A. Tripathi, S. K. Chowdhary, and V. Jangra (2023) [71]	KNN, Naive Bayes, RF, SVM, Gradient Boost, MLP, LR, ANN MLP Classifier	The dataset was compiled from multiple papers and datasets available on Kaggle and consists of information on different **jobs and the required skills** for each	Structured	Figure 1 depicts the flow chart of the System

Table A.2: Qualitative Analysis of the Literature

Literature	**Main Purpose**	**Outcomes**	**Other Remarks**
R. P. Archana, S. M. Anzar, and N. P. Subheesh (2023) [5]	“This paper proposes a novel method titled Graphology-based Career Analysis and Prediction System (GCAPS) that can help a student choose the right career that suits his personality. It uses the vocational personality traits according to the well-known Holland theory.”	The G-CAPS model holds the potential to generate relevant applications in the field of engineering education. However, being a proposal at the moment, the model is unable to provide results regarding the analysis and performance of G-CAPS.	The computational cost is higher in G-CAPS due to the use of independent CNNs in model development. Similarly, the classification processes cannot be performed simultaneously, as all the features from a single handwritten image cannot be extracted together. This issue could potentially be addressed in future works by implementing suitable Delay Flip-Flops (DFF).
A. Ghosh, B. Woolf, S. Zilberstein, and A. Lan. (2020) [20]	“In this paper, we propose a novel and interpretable model, the monotonic nonlinear state-space (MNSS) model, to analyze online user professional profiles and provide i) actionable feedback to users on skills they need to acquire and ii) recommendations on their future career path.”	The results show that MNSS and NSS outperform all other baselines in most cases.	The dataset does not contain counterfactual information, such as job offers that users turned down or jobs for which they did not qualify. Therefore, the only aspect that can be analyzed is the observed career decisions made by each user, and it cannot take into account the real-life constraints they faced or study user qualifications. Moreover, the career path recommendations are generated from historical user career path data and may be biased.

B. Harsha, N. Sravanthi, N. Sankeerthana, and M. Suneetha (2022) [24]	“To assist the students in making a right career choice, this research study investigates the feasibility of predicting the right career path using survey data in a scientific and methodical manner by considering all the elements. This research study proposes a machine learning based career path prediction to assist the students in selecting the suitable vocation for their bright future.”	The results show that the accuracy of Random Forest is 88.33%, while Decision Tree exhibits an accuracy of 86.53%. SVM achieved the highest accuracy at 90.3%, and as a result, it was selected for all subsequent data predictions.	-
M. He, D. Shen, Y. Zhu, R. He, T. Wang, and Z. Zhang (2019) [25]	“The major contributions of this paper can be summarized as follows. We study the problem of career trajectory prediction for talents, with a focus on predicting the talents’ next job’s position name, salary, and company scale. We propose a novel CNN architecture for modeling the career trajectory of talents. We conduct extensive evaluations with real-world talent data to demonstrate the effectiveness of our career trajectory model in terms of predicting position name, salary, and company scale.”	Table 1 presents the optimal outcomes for each task achieved with the CNN model employed in this study. Notably, the accuracy for positions surpasses that reported in [17], registering at 0.488, whereas the best micro precision in [17] is noted as 0.472, and the macro precision is 0.424. However, the precisions for salary and company scale in this paper’s results are lower..	-

M. He, X. Zhan, D. Shen, Y. Zhu, H. Zhao, and R. He (2021) [26]	“This article makes the main contributions as follows: We design effective data visualization to show some interesting insights about the job transitions obtained from a large real-world dataset. We propose a novel model for professional career trajectory prediction by feature fusion. We construct extensive experiments on a large real-world dataset, and the results have demonstrated the effectiveness of our proposed model.”	Both the CNN and LSTM-based models proposed for these three tasks exhibit superior performance compared to all baseline models by a considerable margin. It appears that the specifically designed local neural network components contribute to enhanced performance, leading to more accurate estimations for the next job.	-
A. José-García, A. Sneyd, A. Melro, A. Ollagnier, G. Tarling, H. Zhang, M. Stevenson, R. Everson, R. Arthur (2022) [27]	“In this paper, we introduce an AI-based solution named C3-IoC (https://c3-ioc.co.uk), which intends to help students explore career paths in IT according to their level of education, skills and prior experience. The C3-IoC presents a novel similarity metric method for relating existing job roles to a range of technical and non-technical skills. This also allows the visualization of a job role network, placing the student within communities of job roles.”	The results indicate that the proposed tool consistently provided users with the anticipated job role and career path, often exceeding their expectations, and fewer instances were reported of inaccurate or incorrect results.	The effectiveness of the outcomes (job roles) produced by C3-IoC will primarily rely on the caliber of the input information supplied by the user, encompassing both soft and technical skills.

A. Kamal, B. Naushad, H. Rafiq, S. Tahzeeb (2021) [29]	"Our smart career guidance system would aid the youth in their endeavor of choosing the most appropriate career path for themselves."	The project has successfully attained its predetermined objectives. Furthermore, it was noted that the Random Forest Classifier demonstrated superior performance compared to the XGBoost classifier, particularly when dealing with more than two classes.	The career guidance system undergoes training using data gathered from students in particular fields. As a result, the system's capability to aid students in choosing suitable courses is confined to similar areas and fields.; Due to limited access to the dataset, the resulting model might not always be accurate.; Given that it is a multi-class classification problem, some classes are not equally represented (i.e., an imbalanced dataset), leading to inaccurate predictions for minority classes.

H. Kolhe, R. Chaturvedi, S. Chandore, G. Sakarkar, and G. Sharma (2023) [34]	“This research attempted to develop a model for the user which predicts the career path in a precise manner and gives actionable feedback and career recommendations to encourage them to make significant career judgments.”	The career recommendation system has been developed. The model prompts users with questions related to their current profession, activities, education, age, gender, academic and non-academic performance, and more. The system requires information about the user’s current profession, the desired future profession, and their education details. Subsequently, the system provides a prediction for the most favorable career decision. This process assists users in making informed decisions, ensuring they choose a career path that aligns with their preferences and avoids potential mistakes.	-
V. R. Kumbhar, M. M. Maddel, and Y. Raut. (2023) [35]	"Considering the user’s skills, interests, and preferences, the system will suggest different career paths. The results of this assignment can help students zero in on a subject area that fits their knowledge and interests the best."	This study suggests that when students have a more comprehensive grasp of career concepts, they are better equipped to make informed decisions regarding their career paths.	-

H. Li, Y. Ge, H. Zhu, H. Xiong, and H. Zhao (2017) [38]	"In this paper, we proposed a novel survival analysis approach for modeling the career paths of employees, which is based on multitask learning with ranking constraint formulation."	The results indicate that CDT+SE and CDT+DE outperform all baseline methods. Notably, CDT+DE exhibits a slightly better performance than CDT+SE. Additionally, methods incorporating dynamic features demonstrate significantly better performance compared to those without dynamic features. Modeling censored data leads to more accurate turnover behavior prediction, impacting the performance comparison of M-LASSO*, M-L2,1*, and other multitask learning-based methods. Multitask learning-based methods show a slight advantage over most parametric survival models. Importantly, the results on Data 1 surpass those on Data 2. This discrepancy may be attributed to the fact that most employees in Data 1 have historical observations in the training set, enhancing the accuracy of predictions on their career paths observed in the testing set.	-

L. Li, J. Yang, H. Jing, Q. He, H. Tong, and B-C. Chen (2017) [39]	"In this paper, we study the problem of Next Career Move Prediction to predict an employee's next career move. We propose a contextual LSTM model named NEMO that integrates the profile context as well as career path dynamics. The proposed model follows the encoder-decoder architecture and we show significant improvements over strong baselines."	The proposed model follows the encoder-decoder architecture and exhibits significant enhancements compared to strong baseline models.	In this research, the model assumes that attributes (e.g., skills) are static for simplicity, which may not hold true in practical scenarios.
Y. Liu, L. Zhang, L. Nie, Y. Yan, and D. S. Rosenblum (2016) [40]	"The objective of this work is to predict the future career stages of a given user, the so-called career path modeling, which can provide potential benefits for employees, employers, and headhunters. For employees, they can get information about their current career stages, the time point for their next job-hopping, as well as the whole picture of their own career paths. For employers, they will be informed of the career progressions of their employees and decide what would be the best time to promote their employees or increase their salaries. When it comes to headhunters, they could be advised of the appropriate time to talk to their target customers as well as the proper job positions for their customers."	The last four multi-task learning methods consistently outperform the initial two mono-task learning methods. This observation confirms that the tasks are interdependent, and simultaneous learning enhances overall performance.; MSLFL and regMVMT achieve higher accuracies compared to MTL.; FL performs better than MTL and regMVMT.; The MSLFL model notably outperforms regMVMT in sales and marketing, showcasing its superiority in software engineering and consultancy. This emphasizes the intricate nature of career progressions across diverse paths.	-

Q. Meng, H. Zhu, K. Xiao, L. Zhang, and H. Xiong (2019) [45]	"In this paper, we propose to address the problem of job mobility prediction by answering two specific questions: (1) 'Who is your next employer?' and (2) 'How long will you work for your next employer?' The first question is regarding the position prediction, and the second one tells the eventual duration of your new job."	HCPNN provided much better accuracy for both prediction tasks.	It is necessary to address the dynamic hierarchical nature of career paths for employees, encompassing internal and external job mobilities. Additionally, considering the influence between environmental factors and individual historical patterns is important.
Y. Pan, X. Peng, T. Hu, and J. Luo. (2017) [47]	"Our main objective is to determine how factors such as personality, industry, and education background impact one's career path, and the highest career stage one could reach. In this study, we bring a novel methodology to determine a career stage based on the job title and company information, so that a career path that consists of several stages could represent occupational growth."	The ensemble learners demonstrate superior performance over SVR across all industries. In the four most prominent sectors (IT, Marketing, Media, PR), the forecasting accuracies show enhancements of 1.35%, 1.75%, 0.45%, and 0.85%, respectively, in comparison to the 61.85% baseline method. While the improvements in accuracy are modest individually, the collective consideration of correlated personalities across various industries highlights the potential to advance our prediction objectives.	Many users neglect to complete the education section on LinkedIn, while some only provide a school name without specifying their degrees and majors. Consequently, gathering comprehensive data for analyzing other aspects of educational background, including degrees, programs, and majors, becomes challenging. Additionally, our predictions may apply to individuals in common industries who wish to post on social media. However, employees from traditional industries may not frequently leave traces on social media.
M. Rane, S. Kalal, J. Chandegara, T. Kakkad, T. Jain, and S. Jagtap (2023) [49]	"The motive of this research was to design and develop a website for career prediction which predicts fitting options for a candidate in choosing a suitable field."	The proposed system successfully generates relevant career recommendations using machine learning models based on the user's input provided in the form of answers to a questionnaire.	-

T. R. Razak, M. A. Hashim, N. M. Noor, I. H. A. Halim, and N. F. F. Shamsul (2014) [50]	"The factor that may cause students not successful in their career is due to wrongly choose a job that suits with them. It requires a decision-making process at an early stage. So, this study proposed a system that gives recommendations for student about their career based on academic result and their abilities by using a fuzzy logic approach."	The findings from this study will assist UiTM students in simplifying their career selection process, enhancing its convenience, adaptability, and efficiency. This is due to the ability to engage in self-testing without the need for extensive guidance from counselors.	-
M. Roy, A. K. Bhoi, and K. Sharma (2022) [52]	"In this research paper, the prediction of the student's career is made using ADABOOST, Support Vector Machine (SVM), Random Forest (RF), and Decision Tree (DT) approaches."	According to the results, SVM produces a superior outcome with an accuracy value of 98 percent, followed by ADABOOST, which achieves an accuracy of 88 percent.	-
A. Shankhdhar, A. Gupta, A. K. Singh, and R. Pradhan (2022) [56]	"This paper centers essentially around the prediction of the software engineering space of applicants in the vocation field. Most appropriate space as per interests and capabilities of the applicant is anticipated and reasonable employment opportunities pertinent to the area are shown."	Based on the findings, SVM exhibited greater precision at 84.12 percent, followed by the decision tree with an accuracy of 81.2 percent.	-

D. Sobnath, T. Kaduk, I. U. Rehman, and O. Isiaq (2020) [57]	"This study aims to identify the potential predictive features, which will improve the chances of engaging disabled school leavers in employment about 6 months after graduation."	Features such as age, institution, and disability type, among others, were recognized as significant predictors. The feature selection algorithms also emphasized specific learning difficulties such as dyslexia, dyspraxia, or ADHD as crucial features for a predictive model. The Decision Tree Classifier and Logistic Regression models both delivered optimal outcomes in forecasting the Standard Occupation Classification (SOC) for disabled school leavers in the UK, achieving an accuracy of 96%.	-

M. Sodanil, S. Chotirat, L. Poomhiran, and K. Viriyapant (2019) [58]	“This research aims to analyze the relationships between courses that are likely to produce a future career for students using the Apriori algorithm”	The connections among the 25 courses showed a notable influence on students’ future careers. The assessment of rules demonstrated an efficiency of 92.56% in accuracy when 14 rules were applied during curriculum meetings and revisions. These rules not only served as guidelines for students in securing internships aligned with their abilities but also increased their likelihood of being hired. This creates an avenue for students to obtain jobs directly within their field of study post-graduation, a metric that higher education institutions can employ to gauge the success of graduate production	-
Z. Soliman, P. Langlais, and L. Bourg. (2019) [59]	“In this work, we focus on exploring methods by which we can model the career trajectory of a given candidate, with the help of data mining techniques applied to professional social media data”	The accuracy at rank 1 is approximately 35% on the dataset with the CNN-LSTM model for exact job title matches, and it increases to 67% at rank 10. When examining the mean percentile rank, the CNN-LSTM model consistently identifies the target within the 6 best-scoring prediction labels on average.	The variation in job titles for similar responsibilities posed a significant challenge in this research.

A. Thomas, A. K. Varghese, P. L. Alex, B. J. Mathews, and L. K. Dhanya (2022) [60]	“This paper investigates the various machine learning algorithms and their applications and analyses if they are suitable for the task. The aim of this paper is to compare the performance of the algorithms and decide which algorithm is better suited for the task”	Naive Bayes, decision tree, and SVM consistently delivered commendable performance as the dataset size increased. In terms of accuracy, SVM with RBF kernel, decision tree, and multilayer perceptron (MLP) stood out as the top performers. When considering execution time as the determining factor, Naive Bayes emerged as the most efficient.	-
T. D. Truong, N. Q. H. Ton, C. N. Truong, and S. P. Nguyen (2022) [63]	“In this paper, we propose a new model called Exploring Career Path with Attention Mechanism and Survival Analysis (ECPASurv) to incorporate information about an employee to solve two problems, with the first one being the prediction of the risk score of an employee and the second one related to the forecast of the next job title that an employee may want to apply for.”	The proposed model exhibited improved accuracy for both prediction tasks and demonstrated competitiveness with other state-of-the-art methods.	-
P. Verma, S. K. Sood, and S. Kalra (2017) [65]	“This paper proposes a comprehensive student-centric recommendation system based on research analytics framework for choosing the best career path.”	The outcomes from user-based evaluation indicate that the proposed method produces more satisfactory recommendations for career paths in comparison to other baseline approaches.	-

S. Vignesh, C. Shivani Priyanka, H. Shree Manju, and K. Mythil (2021) [66]	"We came up with an idea of providing an objective assessment of one's skill set and caliber that recommend a right stream to choose and hence we picked this as our problem statement and started thinking through how we can help the students in addressing this question."	The recommendations provided by the proposed system are more accurate than those of the existing career guidance system.	-
A. K. Yadav, A. Dixit, A. Tripathi, S. K. Chowdhary, and V. Jangra (2023) [71]	"This research paper proposes a career guidance system based on the ANN MLP classifier that uses objective data to provide accurate predictions for students and users exploring career options."	Compared to traditional career guidance methods, the proposed system boasts a higher accuracy rate of 95.334%	-

Appendix B

Interview Transcript

Deputy of the ASN Information System

1. Do you think it is important to predict the career path of civil servants? Why?

   Ans: It is very important. An employee needs to know what he or she is going to be in the future. It is impossible for someone to not know what his or her targets are. The constitution also regulates the career path of civil servants. Someone should not be stuck in one position because everybody wants to develop and progress in the future.

2. Can you provide an overview of the current career path structure for civil servants in Indonesia?

   Ans: It depends on the person's job and function when they enroll in the institution. However, this does not mean that they will stay on the initial path. It is possible for someone with an IT background to work in other units unrelated to IT. It depends on their self-development in the future. The more competencies they have, the more possibilities will open up for them. This is also in line with the concept of a merit system that utilizes potential (qualification and competence) and performance. If their potential and performance are high, they can be placed in any position and be considered capable of doing the job. Civil servants are also able to pursue different majors for their bachelor's and master's degrees. The more qualifications they have, the higher the level they can attain, which will also give them greater possibilities.

3. Can you describe any specific patterns or trends in career progression among civil servants that you have observed?

   Ans: Someone will not always follow the same pattern. Indeed, there are some people who stay in the same pattern, but there are more people who will be in different patterns compared to their initial position. This is also in line with the merit system, where people are expected to experience different fields if they want to be promoted to a higher level. It is possible for them to go back to their initial field.

4. Are there any existing methods or tools used for predicting career paths or performance within the civil service?

   Ans: Based on the existing law, there is already talent management, but it is still in the trial phase. There are three variables: competence, qualification, and performance. Two variables, potential and performance, will predict what position someone will be in. For example, someone might be lazy but competent, or a workhorse who only works but does not want to develop themselves. Or someone might have high performance but lack sufficient competence (only capable in their specific field). In the context of competency development, these two variables determine what position someone will be in.

5. Are you trusting the result?

   Ans: Yes, but it will depend greatly on the validity of the data produced, especially performance data. Often, there are human issues when assessing performance, where sometimes there is discomfort in giving a poor rating. Ideally, the assessment should reflect the actual conditions. If the data is obtained correctly, it will show whether the person is capable of performing tasks in multiple fields or just one field. The issue is the quality of the data. The data produced must be of high quality so that it can be used for policymaking. For competencies, we have already assessed them using technology (CACT) with the hope that the competency assessment process will become more transparent. If both sets of data are valid, we will be able to determine the best position for this person and their likelihood of success in that role will be high. However, if the data is not of good quality, the individual will fail in the position because the data does not reflect the actual conditions. BKN itself is conducting trials on how to use this data to develop the careers of civil servants (especially for echelons 3 and 4, and now even for Echelon 2). Since this is still new, we are currently examining whether the data accurately reflects the actual conditions, which should ensure the individual's success in the position.

6. What are the key factors that influence career progression within the civil service?

   Ans: Qualification, competence, and performance. However, not everyone with a higher education is successful in their career. Competence: we want to see if someone is competent in their field. Ideally, if their qualifications and competence are good, their performance will also be good. Except if the person is lazy or lacks passion in that field. Passion will greatly determine whether the person can have good performance.

7. How is data related to civil servants collected, stored, and managed within the information system?

   Ans: Data is collected from the information services provided to all civil servants (SIASN, SSCASN, MySIASN). This data will serve as the single basis for providing services. Changes or updates to the data will also be based on the services provided. At certain points, there will also be assessments, and thus, the data will be updated accordingly. When someone undergoes competency development, their data will also be updated. SIASN is not only for providing services but also for updating data.

8. What challenges or limitations exist in accurately predicting career paths using traditional methods?

   Ans: There is subjectivity from leaders, including nepotism, resulting in many leadership positions being filled not based on actual conditions (corruption, collusion, and nepotism are very high). Moreover, with an open political system, corruption, collusion, and nepotism are increasing. This is what we aim to fix with the system. How can the system minimize corruption, collusion, and nepotism so that the appointed individuals are those who truly have the qualifications in that field? In the regions, people with good potential are dismissed because their political aspirations do not align with the local leaders.

9. How do you envision machine learning being applied to predict the career paths of civil servants?

   Ans: One method that can capture comprehensive information about a person is by obtaining as much data as possible. The more complete the data that can be obtained from someone, the better it will be able to depict their behavior. Machine learning is capable of capturing this, for example, by analyzing their views towards the government. This makes the data presented for determining someone's promotion richer. Big data analysis shows how this data meets statistical data requirements in Indonesia and helps leaders make decisions. It also allows intelligence agencies to predict an individual's behavior within the government. Relying solely on statistical data will heavily depend on the quality of that data.

10. Are there any ethical or privacy considerations associated with using machine learning for predicting career paths?

    Ans: Yes, we have the Personal Data Protection Law, but its effectiveness depends on how well it can be enforced. Of course, there are sanctions and measures that the government must take to protect personal data. However, for example, if someone speaks in the public domain but is conveyed by the individual privately, is it still considered confidential?

11. How do you foresee the implementation of machine learning impacting decision-making processes related to civil servant career development?

    Ans: It will have a significant impact. When we can combine structured and unstructured data to ensure that the person is in the right place.

12. What potential benefits or outcomes do you anticipate from leveraging machine learning in this domain?

    Ans: The policies adopted will be more targeted. With more comprehensive information, it seems less likely to make mistakes.

Director of Deployment and Enhancement Civil Servant Information System

1. Do you think it is important to predict the career path of civil servants? Why? Ans: It is important because it serves as guidance or a reference for civil servants, so they know how far they can advance to the highest positions. It provides motivation, enabling them to deliver their best performance to achieve those top positions.

2. Can you provide an overview of the current career path structure for civil servants in Indonesia?

   Ans: Career paths are determined by the organization, providing guidance for its civil servants. Civil servants follow this path and cannot set their own targets, which can be ineffective for the organization as it does not achieve maximum performance. With career path predictions, their performance can be maximized.

3. Can you describe any specific patterns or trends in career progression among civil servants that you have observed?

   Ans: Regularly, a civil servant can attain a career in echelon 2 at the age of 50. However, with career path predictions, this can be accelerated by demonstrating performance achievements, allowing them to receive an Extraordinary Promotion (KPLB).

4. Are there any existing methods or tools used for predicting career paths or performance within the civil service?

   Ans: There is no method yet that can predict this because it is left to the individual civil servants. If they want to advance their careers quickly, they do so in their own way. However, with guidance, these civil servants will become more competitive, allowing the organization to obtain the best people.

5. Are you trusting the result?

   Ans: -

6. What are the key factors that influence career progression within the civil service?

   Ans: Having performance above expectations, possessing strategic thinking, and making breakthroughs that the organization can leverage to achieve its performance goals. There is alignment between individual and organizational needs. In terms of data, it is performance data. It can be seen in the portfolio. Whether the experience supports them in reaching the pinnacle of their career. Assessment data.

7. How is data related to civil servants collected, stored, and managed within the information system?

   Ans: Integration with the API is already in place. Data is collected from data producers that flow into BKN/assessment. When there is a data update, it will also be updated in SIASN.

8. What challenges or limitations exist in accurately predicting career paths using traditional methods?

   Ans: Traditional methods are not based on data and are assumptive. By using regulation (UU 20 of 2023), a civil servant can reach high career levels through talent management. This can encourage civil servants to compete in demonstrating their best performance.

9. How do you envision machine learning being applied to predict the career paths of civil servants?

   Ans: In conducting a data-driven analysis process, especially with large amounts of data (4.3 million) and varying characteristics such as job levels, education, and gender, it would be difficult to use a concept that does not read data. Machine learning will help identify which features are useful for decision-making.

10. Are there any ethical or privacy considerations associated with using machine learning for predicting career paths?

    Ans: In the context of the Personal Data Protection Law, there is a concept of consent. When civil servants provide their data to BKN, BKN has the authority to process that data to make policies that have a positive impact on the civil servants. Therefore, there is no issue.

11. How do you foresee the implementation of machine learning impacting decision-making processes related to civil servant career development?

    Ans: Using machine learning certainly makes it easier to interpret big data. With the available data, its application can be generalized, not only for individual civil servants but also on a national scale.

12. What potential benefits or outcomes do you anticipate from leveraging machine learning in this domain?

    Ans: To obtain accurate results that can be used for decision-making.

Senior Software Engineer in the Directorate of Deployment and Enhancement Civil Servant Information System A

1. Do you think it is important to predict the career path of civil servants? Why?

   Ans: Personally, I feel it's not that important because I work to do the task at hand (the best) and make an impact. When our performance is good, we will likely be sought after and considered by the leadership. Work as well and as focused as possible. When the leadership appreciates us, we will be promoted. But in general, having an ideal path means that the development of civil servants should align with their passion (qualifications and competence).

2. Can you provide an overview of the current career path structure for civil servants in Indonesia?

   Ans: In the current situation, there are many disruptions from various factors: politics (especially at the regional level). Regional civil servants have become hopeless about becoming officials (JPT) because, in their mindset, those who will be appointed are the people closest to the leaders.

3. Can you describe any specific patterns or trends in career progression among civil servants that you have observed?

   Ans: Civil servants now have diverse backgrounds. Although they may have IT qualifications, they might not have a passion for it, which is actually permissible according to regulations. Education might not be the only requirement; perhaps a portfolio (training or experience) is more important.

4. Are there any existing methods or tools used for predicting career paths or performance within the civil service?

   Ans: We have regulations designed to establish career paths, but they have not yet been implemented properly.

5. Are you trusting the result?

   Ans: Yes, but adjustments are needed for the current conditions.

6. What are the key factors that influence career progression within the civil service?

   Ans: Motivation, loyalty, and integrity are crucial. Many intelligent people do not progress in their careers due to a lack of loyalty and integrity. We also have Ber-AKHLAK/ASN values and must consider involvement in legal issues/violations (poor reputation).

7. How is data related to civil servants collected, stored, and managed within the information system?

   Ans: Data on job history, training, rank, and education are stored. However, data on competencies, social and cultural awareness, digital literacy, and emerging skills are still being sought through the CACT test. This data is obtained from the assessment center, but now a rapid assessment is being planned. 8. What challenges or limitations exist in accurately predicting career paths using traditional methods?

Ans: Previously, the agency determined what we had to do, even though what was designed by the agency might not align with the employee's own passion. Currently, it is encouraged for individuals to design their own career paths (self-nomination).

8. How do you envision machine learning being applied to predict the career paths of civil servants?

   Ans: Machine Learning (ML) is a technology with two sides. It can have an extraordinary impact on matters related to politics and provocation. However, with minimal negative aspects, it can be very helpful (similar to Decision Support Systems).

9. Are there any ethical or privacy considerations associated with using machine learning for predicting career paths?

   Ans: We can choose the data that we want to process.

10. How do you foresee the implementation of machine learning impacting decision-making processes related to civil servant career development?

    Ans: It will have a positive impact because it can assist in decision-making when we are unsure. The more data used, the better the results will be.

11. What potential benefits or outcomes do you anticipate from leveraging machine learning in this domain?

    Ans: To be able to provide recommendations and strengthen the decisions we want to achieve. Sometimes we are unaware of our potential, but ML can provide us with data-based decisions.

Senior Software Engineer in the Directorate of Deployment and Enhancement Civil Servant Information System B

1. Do you think it is important to predict the career path of civil servants? Why?

   Ans: It is important to eliminate the mindset that becoming a civil servant is only about being a 'civil servant', but rather about understanding what one can become. This helps in gaining better work motivation because not everyone knows the purpose of working or building a career as a civil servant. This way, they can bring out their best potential.

2. If there is an application/feature to predict the next career path, who will be the primary users of the predictive system?

   Ans: HR, so that personnel managers can understand and make decisions. However, it is expected that the results can also be shown to each civil servant so they can see their potential career paths.

3. Can you provide an overview of the current career path structure for civil servants in Indonesia? Ans: Currently, with the implementation of bureaucratic reform, we have seen the simplification of the bureaucracy. The career ladder, which originally included echelon 5, now only includes echelons 3 and 4, and the number of these positions is limited. To reach echelon 3, certain requirements must be met, such as years of service, rank, and others. Since the simplification of the bureaucracy, functional positions have become quite attractive. Why? All intermediate levels have an age limit of 60 years, the same as directors, although the income is not the same. Anyone can fill these roles as long as they pass the competency exams and

there are available positions. Looking at it, one directorate only has one echelon 2 position, while intermediate functional positions can number around 5-10 people, thus offering a greater chance compared to becoming a director.

4. Can you describe any specific patterns or trends in career progression among civil servants that you have observed?

   Ans: The pattern is not based on education but rather on the institution. For example, in BKN, which is a personnel agency, individuals usually become personnel analysts. In the Attorney General's Office, civil servants with a law degree typically have a career as prosecutors.

5. Are there any existing methods or tools used for predicting career paths or performance within the civil service?

   Ans: Not Yet

6. Are you trusting the result?

   Ans: -

7. What challenges or limitations exist in accurately predicting career paths using traditional methods?

   Ans: -

8. How do you envision machine learning being applied to predict the career paths of civil servants?

   Ans: This is something worth trying. So far, we have never predicted the career paths of civil servants. We don't know where our careers are headed, nor do we know if we can choose A, B, C, or D for our career paths. With technology, there should be a clearer picture of suitable career directions, especially for those who are newly entering the service.

9. Are there any ethical or privacy considerations associated with using machine learning for predicting career paths? Are there any legal or regulatory requirements the system must comply with?

   Ans: Predictions have not yet been regulated; currently, the focus is on managing whether individuals are eligible for promotion as per the Ministerial Regulation (Permenpan) No. 3 of 2019.

10. How do you foresee the implementation of machine learning impacting decision-making processes related to civil servant career development?

    Ans: The first stage can serve as a recommendation for training and other preparatory steps.

11. What potential benefits or outcomes do you anticipate from leveraging machine learning in this domain?

    Ans: Civil servants will become a profession that is not looked down upon.

12. What types of data are currently collected about civil servants?

    Ans: Personal data, job history, disciplinary actions, training, education, performance, competency, and potential data, integrity, and morality.

13. How is this data stored and managed?

    Ans: The data will be stored in a database managed by the Directorate of Personnel Data Management and Information Presentation. The mechanism for obtaining the data will involve web services provided by the respective institutions. Data on integrity and morality will adhere to the regulations set by PPATK. Since the data cannot be used indefinitely, regular updates will be necessary.

14. What are the data sources, and how often is the data updated? Ans: The data will be obtained from the respective institutions. Competency data will be updated every 2-3 years, integrity and morality data annually, and personal data can be provided by the civil servants themselves or their respective institutions.

15. How should the data be accessed and integrated into the API?

    Ans: The data will be obtained directly from the database by submitting a request specifying the required data, and it is usually provided only to one institution at a time.

16. What specific endpoints are needed for the API?

    Ans: Endpoints for predictions.

17. What should be the format of the requests and responses?

    Ans: JSON

18. Are there any specific API standards or protocols that need to be followed?

    Ans: Oauth 2.0

Senior Software Engineer in the Directorate of Deployment and Enhancement Civil Servant Information System C

1. Do you think it is important to predict the career path of civil servants? Why?

   Ans: It is important so we can have a goal because we shouldn't remain at the entry-level forever. Even if we can't reach the head of the organization, we can know (with our education) what options we have.

2. If there is an application/feature to predict the next career path, who will be the primary users of the predictive system?

   Ans: HR and the civil servants

3. Can you provide an overview of the current career path structure for civil servants in Indonesia? Ans: Educational background and level of education are important. Master's degree graduates (S2) have the opportunity to reach Echelon 2 positions, but those with lower education levels will find it difficult to advance to other positions.

4. Can you describe any specific patterns or trends in career progression among civil servants that you have observed?

   Ans: A person's career path has many possibilities as long as the position aimed for aligns with their skills and educational background. However, another influencing factor is the individual's interest, so it is possible for someone to hold a position that appears unrelated to their education.

5. Are there any existing methods or tools used for predicting career paths or performance within the civil service?

   Ans: Not Yet

6. Are you trusting the result?

   Ans: -

7. What challenges or limitations exist in accurately predicting career paths using traditional methods?

   Ans: -

8. How do you envision machine learning being applied to predict the career paths of civil servants?

   Ans: I hope that in the future, someone can see the entire path and know what needs to be fulfilled to reach that point. It should be limited by age.

9. Are there any ethical or privacy considerations associated with using machine learning for predicting career paths? Are there any legal or regulatory requirements the system must comply with?

   Ans: It must comply with civil service regulations. Because, although there are currently no regulations specifically regarding the use of machine learning to predict career paths, there are regulations about career paths in general.

10. How do you foresee the implementation of machine learning impacting decision-making processes related to civil servant career development?

    Ans: Currently, the predictions can only be used as recommendations. In reality, we still see external factors influencing decision-making. With the implementation of machine learning technology, it is hoped that this can be avoided.

11. What potential benefits or outcomes do you anticipate from leveraging machine learning in this domain?

    Ans: Employees can be aware of what they can become, so if they are given an unsuitable position or someone else is given a position that should be suitable for them, they can raise questions. In other words, there will be transparency.

12. What types of data are currently collected about civil servants?

    Ans: Education, rank, date of assuming a position for calculating tenure, date of birth for age calculation, performance.

13. How is this data stored and managed?

    Ans: We have a database managed by the Directorate of Data Processing and Civil Service Information Presentation.

14. What are the data sources, and how often is the data updated? Ans: The data comes from government organizations or individual civil servants and is updated whenever there is a change in the data.

15. How should the data be accessed and integrated into the API?

    Ans: Ideally, it should be accessed through an API, but since one does not exist currently, it can be accessed directly from the database.

16. What specific endpoints are needed for the API?

    Ans: Endpoints for predictions and endpoints for position information (requirements to fulfill those positions).

17. What should be the format of the requests and responses?

    Ans: JSON

18. Are there any specific API standards or protocols that need to be followed?

    Ans: REST API, OAuth 2.0, and the programming language we use is Java with the Spring Boot framework.

19. What are any other requirements for the system?

    Ans: For security, the system should use two-factor authentication. The system should also have high availability and reliability, meaning it can be accessed from anywhere at any time, and any downtime should be well-explained and justified. For responsiveness, the access time should be less than 5 seconds. The machine learning models should also be trained monthly.

Intermediate Data Scientist in Directorate of Data Management & Information Presentation

1. Do you think it is important to predict the career path of civil servants? Why?

   Ans: From an employment perspective: it is important so that we can see what a civil servant can become in the future. Especially for those who have potential, and it is evident from the beginning, allowing us to predict their future. From the civil servant's perspective: it is important because it can provide motivation.

2. Can you provide an overview of the current career path structure for civil servants in Indonesia?

   Ans: In general, we are not yet very linear between education and positions. For example, someone with a bachelor's degree in IT may not necessarily hold a position in the IT field. Perhaps in the future, with career path prediction, this alignment can be achieved.

3. Can you describe any specific patterns or trends in career progression among civil servants that you have observed?

   Ans: In practice, it doesn't have to align with formal education because not all passions come from formal education, especially now with many courses and training available. However, recruitment into certain positions must be linear with formal education.

4. Are there any existing methods or tools used for predicting career paths or performance within the civil service?

   Ans: -

5. Are you trusting the result?

   Ans: -

6. What are the key factors that influence career progression within the civil service?

   Ans: The potential, competence, and motivation of the civil servant. Education cannot necessarily be considered a key factor.

7. How is data related to civil servants collected, stored, and managed within the information system?

   Ans: The data is stored in a database and processed, one of which is through the self-updating data process. Regarding potential and competence, we request institutions to update their civil servants' data through assessments conducted by the Competency Assessment Center.

8. What challenges or limitations exist in accurately predicting career paths using traditional methods?

   Ans: Limited data, because past data is incomplete.

9. How do you envision machine learning being applied to predict the career paths of civil servants?

   Ans: We need ML and have not yet implemented it. ML also comes with performance measurement. We can use predictive analytics in ML to measure the accuracy of the ML model in predicting the career paths of civil servants. However, we still need accurate data.

10. Are there any ethical or privacy considerations associated with using machine learning for predicting career paths?

    Ans: Yes. We must consider the scope of the data being used, whether NCSA is acting as the personnel management agency or as an institution, because NCSA can view data on all civil servants across Indonesia. Access should be restricted to executives and the individuals concerned.

11. How do you foresee the implementation of machine learning impacting decision-making processes related to civil servant career development?

    Ans: ML will be very useful. We are currently implementing the ASN career system. ML can be applied there. We can use it to view succession planning for vacant positions.

12. What potential benefits or outcomes do you anticipate from leveraging machine learning in this domain?

    Ans: Someone can see their predicted career path.

Appendix C

Features

Table C.1: Features

Feature	Description
next_position	The identifier of next position
current_position	The identifier of current position
is_position_active	Active/inactive status of current position
position_month_service	The total of time the civil servant being employed in current position (in months)
month_service	The total of time the civil servant being employed (in months)
is_current_position_active	Active/inactive status of the current position
management_position	Records indicating whether a civil servant has ever served in a position within the management field/group.
ict_position	Records indicating whether a civil servant has ever served in a position within the ICT field/group.
law_position	Records indicating whether a civil servant has ever served in a position within the law field/group.
economics_position	Records indicating whether a civil servant has ever served in a position within the economics field/group.
politic_position	Records indicating whether a civil servant has ever served in a position within the political field/group.
psychology_position	Records indicating whether a civil servant has ever served in a position within the psychology field/group.
social_position	Records indicating whether a civil servant has ever served in a position within the social field/group.
civil_position	Records indicating whether a civil servant has ever served in a position within the civil field/group.
linguistic_position	Records indicating whether a civil servant has ever served in a position within the linguistic field/group.
pcs_start_date	The starting date of the person becoming prospective civil servant
rank_start_date	The starting date of the person in the new rank
position_start_date	The starting date of the current position
position_type	Type of the position. 1 = managerial; 2 = functional; 4 = administration

class_max	Maximum Position class
class	Current Position class
rank	Civil servant current rank
first_rank	Civil servant rank when first employed
dental_nurse_degree	Records of the civil servants holding a degree in dental nursing
dentistry_degree	Records indicating whether a civil servant holds a degree in dentistry
economic_degree	Records indicating whether a civil servant holds a degree in economics
education_degree	Records indicating whether a civil servant holds a degree in economics
ict_degree	Records indicating whether a civil servant holds a degree in ICT
humaniora_degree	Records indicating whether a civil servant holds a degree in humaniora
law_degree	Records indicating whether a civil servant holds a degree in law
linguistic_degree	Records indicating whether a civil servant holds a degree in linguistics
management_degree	Records indicating whether a civil servant holds a degree in management
medical_degree	Records indicating whether a civil servant holds a degree in medical
nursery_degree	Records indicating whether a civil servant holds a degree in nursery
pharmacy_degree	Records indicating whether a civil servant holds a degree in pharmacy
politic_degree	Records indicating whether a civil servant holds a degree in politics
psychology_degree	Records indicating whether a civil servant holds a degree in psychology
social_degree	Records indicating whether a civil servant holds a degree in social
other_degree	Records indicating whether a civil servant holds a degree in other majors

Appendix D

Initial Set of Hyperparameters for the Machine Learning Models

Table D.1: Initial Set of Hyperparameters for the Machine Learning Models

ML Model	**Hyperparameter**	**Description**	**Initial Value**
Decision Tree	criterion	The function to measure the quality of a split	gini, entropy
	max_depth	Maximum depth of the tree	None, 3-10 in steps of 1
	min_samples_split	Minimum number of samples required to split an internal node	2-10 in steps of 1
	min_samples_leaf	Minimum number of samples required to be at a leaf node	1-10 in steps of 1
	max_features	Maximum number of features considered for splitting	None, auto, sqrt, log2
Random Forest	n_estimators	Number of trees in the forest	100-1000 in steps of 100
	max_depth	Maximum depth of each tree	None, 3-10 in steps of 1
	min_samples_split	Minimum number of samples required to split an internal node	2-10 in steps of 1
	max_features	Maximum number of features to consider for splitting	None, auto, sqrt, log2
XGBoost	n_estimators	Number of boosting rounds	100-1000 in steps of 100
	max_depth	Maximum depth of each tree	3-10 in steps of 1
	learning_rate	Step size shrinkage to prevent overfitting	0.001, 0.01, 0.1, 0.5, 0.9, 1.0

	subsample	Fraction of samples used for fitting the trees	0.5-1 in steps of 0.05
	colsample_bytree	Fraction of features used for fitting the trees	0.5-1 in steps of 0.05
Multilayer Perceptron	hidden_layer_sizes	Number of neurons in each hidden layer	(100,), (50,50), (100,50,25)
	activation	Activation function for the hidden layer	identity, logistic, tanh, relu
	solver	The solver for weight optimization	lbfgs, sgd, adam
	alpha	L2 penalty (regularization term)	0.0001, 0.001, 0.01, 0.1
	learning_rate	Learning rate schedule for weight updates	constant, invscaling, adaptive

Appendix E

Confirmation of Data Usage and Access Restrictions

NATIONAL CIVIL SERVICE AGENCY
Jalan Mayor Jenderal Sutoyo Nomor 12 Cililitan, Kramat Jati, Jakarta Timur 13640
Phone +62-21 8093008; Facsimile +62-218090421
Web: www.bkn.go.id | Email: humas@bkn.go.id

Ref : 5445/B-HM.04.02/SD/A/2024 13 August 2024

Dr. Maya Daneva
Associate Professor
School of Computer Science - University of Twente

Re – Confirmation on Data Usage and Access Restrictions
in the Master Thesis of Ms. Juwita P. Pasaribu

Dear Dr. Maya Daneva,

I am writing in response to your recent request in a letter dated 16 July 2024 regarding the use of data for Ms Juwita P. Pasaribu's thesis. I am pleased to inform you that we grant approval for the use of the requested data, subject to certain conditions outlined below.

Please note that while we approve the use of the data, some categories, particularly personal data, have restricted access due to their sensitive nature, i.e., confidentiality and privacy concerns. These restrictions are in place to protect the integrity and privacy of the individuals involved. We trust that these limitations will be respected in accordance with our data protection policies, ensuring that access to sensitive information is carefully controlled.

We believe that the data provided will not only be instrumental in the successful completion of the thesis but also have a significant impact on the advancement of civil servant data management in Indonesia.

We greatly appreciate your understanding and cooperation in this matter.

Sincerely Yours,
For Chairman National Civil Service Agency
Prime Secretary

~

- UU ITE No 11 Tahun 2008 Pasal 5 Ayat 1
"Informasi Elektronik dan/atau Dokumen Elektronik dan/atau hasil cetaknya merupakan alat bukti hukum yang sah."
- Dokumen ini telah ditandatangani secara elektronik menggunakan sertifikat elektronik yang diterbitkan BSrE

Balai Sertifikasi Elektronik