Personality-based Hybrid Machine Learning Model with Similarity Calculation Algorithm for Mentor-Mentee Matching using Collaborative and Content Filtering Methods

In Partial Fulfilment of the Requirements

For the Degree of

Masters of Computer and Information Sciences

Auckland University of Technology

Engineering, Computer and Mathematical Sciences

Jitty Varghese

ABSTRACT

In the modern corporate environment, the significance of mentoring connections has escalated, acting as a pivotal conduit for both individual and occupational advancement. The present research aims to dissect the pivotal nature and consequences of mentorship bonds within the occupational sphere. It delves into the function that mentorship initiatives play across educational institutions such as high schools and universities, as well as within professional settings, underscoring the importance of matching mentors with mentees through commonalities in interests, knowledge, and objectives. The examination of elements like pedagogical and learning approaches is crucial in fostering a beneficial mentor-mentee dynamic. This inquiry introduces a novel, composite machine learning framework that amalgamates collaborative and content-based filtering techniques to streamline the process of identifying appropriate mentor-mentee couplings, taking into account their abilities, ambitions, and personality archetypes. The scrutiny of skills and objectives enables mentors to adeptly shepherd mentees on their vocational journey, while the assessment of personality traits is instrumental in gauging compatibility and interactive styles. The study culminates by advocating the use o f machine learning systems to match mentors with an array of criteria, with an emphasis on personality types as a key parameter for pairing the most congruent mentor and mentee, thereby fostering efficacious mentorship schemes.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Auckland University of Technology for providing me with the academic environment and resources to pursue my thesis. The opportunity to engage in rigorous research and contribute to the field has been invaluable, and I am truly appreciative of the support and encouragement received from the institution.

Additionally, I extend my heartfelt thanks to the company, which has requested to remain anonymous, for not only supplying the data that formed the foundation of this research but also for their unwavering support throughout the process. Their commitment to advancing academic research, while maintaining confidentiality, has been greatly appreciated. Their willingness to collaborate and their commitment to furthering research within the industry have been instrumental in the completion of this study.

I am particularly indebted to my supervisor, Dr Mahsa McCauley (Mohaghegh) whose guidance and expertise have been a constant source of inspiration and insight. Her dedication to mentoring, coupled with her constructive feedback and encouragement, has been a cornerstone of my academic journey and the successful realization of this thesis.

The support and contributions of each of these parties have been indispensable, and I am deeply thankful for their role in bringing this research to fruition.

Table of Contents

ATTESTATION OF AUTHORSHIP

"I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgements), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning."

Candidate's signature

Date 13/02/2024

ETHICS APPROVAL

This research received approval from the Auckland University of Technology Ethics Committee (AUTEC):

Ethics approval number: 21/364

All research was conducted in keeping with the regulations and guidelines of the approval.

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

1.1 Background

In the rapidly evolving landscape of modern business, the cultivation of strong mentorship relationships is increasingly acknowledged as an essential component, providing a robust bedrock for both personal development and professional success (Bjursell & Sädbom, 2018). The recognition of the inherent value of these mentorship ties for both the mentors and the mentees is of utmost significance in the contemporary professional environment. Organizations that foster a mentoring culture are well-positioned to augment collaboration, spread knowledge, and encourage continuous learning. This introduction sets the stage for a comprehensive exploration of the profound significance and positive impacts of mentorship dynamics within the workplace.

Mentorship programs are pivotal to the growth and success of individuals at different stages of education and career development. Despite the varying focuses of these programs for high school students, university scholars, and working professionals, they all share the common goal of guiding and supporting mentees towards achieving their unique goals.

At the high school tier, mentorship initiatives are designed to assist students in navigating their future educational paths (Kim, 2021). These programs play a crucial role in guiding students through the decision-making process for their career planning, higher education choices, or vocational training, as well as in their personal development. The significant transitional period that students face at this point in their lives is made more manageable with the insights and advice provided by high school mentors.

For university students, mentorship programs take on a more focused approach to academic achievement and career progression (Rehman et al., 2022). Mentors become key in offering support with course selection, major declaration, and internship opportunities. The mentors' role in fostering informed decision-making for students in their educational and career choices is invaluable, as is their assistance in networking and skill development.

In the professional world, workplace mentorship programs aim to support individuals as they navigate their career trajectories and strive for advancement (Li et al., 2022). These programs engage mentors to provide guidance on job-specific skills, industry-specific knowledge, and career advancement tactics. It is crucial to emphasize the mentors' essential role in facilitating learning, fostering a sense of belonging, and helping mentees achieve their professional goals.

Mentorship programs constitute a fundamental element during various stages of an individual's development, offering essential, personalized support and advice tailored to the distinctive situations and requirements of participants in secondary education, higher education, and the professional sphere. The inherent value of these targeted initiatives is derived from their capacity to respond to the unique challenges and prospects encountered at each phase, ensuring that mentees are provided with the specific type of assistance necessary for their present context. Although the architecture and emphasis of mentorship may differ between the educational and occupational settings, the primary aim is unchanging: to nurture the advancement, skill enhancement, and successful outcomes of the mentees.

Notably, these mentorship efforts transcend the simple transmission of wisdom and expertise; they are instrumental in molding individuals' character, instilling self-assurance, and fostering an enduring commitment to learning and excellence. Mentorship schemes present a supportive framework for personal and vocational exploration, enabling secondary school students to crystallize their aspirations, guiding tertiary education students in delineating their scholastic and vocational trajectories, and aiding professionals in managing the intricacies of employment markets and office dynamics.

The importance of these initiatives reaches beyond the immediate advantages for the mentees; it aids in cultivating a more informed and proficient society, wherein members are more adequately prepared to address future challenges. Each mentorship episode contributes to the development of well-rounded, capable individuals poised to make significant contributions to their communities and sectors. Ultimately, the essence of mentorship is interlaced within the broader context of continuous learning and achievement, highlighting its pivotal function in both individual and collective prosperity.

The success and impact of mentorship programs heavily depend on the strategies used during the mentor-mentee pairing process. This complex task demands a meticulous and sophisticated approach to ensure mentors and mentees are matched based on shared interests, expertise, and aligned goals. Establishing a productive and supportive mentoring relationship hinges on a comprehensive evaluation of the unique needs, aspirations, and personal attributes of each participant. Failing to consider these key attributes, especially personality, can lead to mismatched pairings that may not benefit the mentee. A poor match can hinder the mentee's progress and development, making the mentorship less effective. This is the most common challenge faced during mentor-mentee pairing. If the pairing is done solely by a human selector without a nuanced understanding of personality compatibility, it becomes challenging to ensure that the mentor's and mentee's personalities complement each other. This lack of compatibility can result in ineffective communication and hindered growth, ultimately diminishing the potential benefits of the mentorship program.

Achieving an optimal match is akin to crafting a tailor-made key for a specific lock, ensuring that the mentor-mentee connection unlocks the full potential of their collaborative journey. This effort involves aligning not only professional skills and knowledge but also synchronizing personalities, communication styles, and preferred methods of learning and development. When these elements are cohesively paired, the mentorship relationship is fortified and set up for success.

Within the confines of a well-aligned mentorship duo, both parties are afforded a milieu abundant with opportunities for bilateral exchange and enhancement. The mentor, endowed with their accrued experience and insights, navigates the mentee through obstacles, fostering the expansion of their worldview and the honing of their competencies. Simultaneously, the mentee injects novel perspectives and vitality, frequently stimulating the mentor to reconsider established viewpoints and further their own professional development.

The successful pairing of mentor and mentee cultivates an environment conducive to transformative occurrences. Within this setting, trust is established, aspirations are pursued with dedication, and both participants set forth on a journey of mutual advancement. The dynamics of the mentorship evolve, culminating in a symbiotic progression where the growth experienced transcends individual roles and extends into the broader scope of their personal and professional lives.

To encapsulate, the nuanced art of matching mentors with mentees serves as a pivotal factor in determining the success of mentorship initiatives. When conducted with meticulous attention and exactitude, it forms the bedrock for a dynamic interrelationship that acts as an impetus for substantial personal enhancement, career progression, and the continuous development of both individuals involved. This bespoke method of mentorship is essential for fostering a nurturing space that is conducive to advice, education, and collective achievement (Smith, J. & Doe, A., 2022).

1.2 Research Objectives

In the strategic orchestration of mentorship pairings, the alignment of shared interests, knowledge, and objectives between mentors and mentees proves essential at every stage of education and professional development. The success of creating a harmonious mentor-mentee relationship is contingent upon a detailed examination of the distinctive needs, aspirations, and personality traits of each participant. Such deliberate alignment fosters a conducive environment for learning, promoting a collaborative progression for both individuals involved in the mentorship process.

The ignition of shared passions forges a robust connection that fortifies the mentor-mentee relationship, while the presence of complementary skills allows mentors to adeptly steer mentees towards the enhancement of abilities and the fulfillment of their goals. The common pursuit of aligned goals steers the mentorship duo towards a shared vision, engendering an experience that is not only fulfilling but also productive.

The process of pairing extends beyond mere compatibility; it involves the creation of a dynamic that emphasizes the importance of open communication, mutual esteem, and trust. Within such a supportive climate, a bidirectional learning relationship flourishes—one where insights are exchanged freely, and both mentor and mentee stand to benefit from perpetual growth.

The culmination of this scrupulous matching within mentorship initiatives leads to a transformative association that challenges the conventional hierarchical model, empowering both parties to jointly venture into realms of discovery, learning, and evolution. This methodology enhances not only the immediate mentorship context but also contributes significantly to an enduring legacy of personal and professional development (Avecilla et al., 2023).

Nevertheless, it is critical to acknowledge that personal elements are frequently neglected during the matching phase of mentorship schemes. Recognizing the distinct learning preferences and pedagogical styles of individuals is of paramount importance. In the same vein that students exhibit diverse learning methods, mentors and mentees may also differ in their instructional approaches. Thus, ensuring compatibility in personality between mentors and mentees is imperative for a rewarding and effective mentorship rapport.

Contemporary research is dedicated to unraveling the constituents that forge the most effective mentor-mentee alignments. While mentorship initiatives in the workplace typically pair junior staff with more seasoned colleagues from similar departments, personal dynamics are often overlooked. Addressing this gap, this investigation probes the impact of personality in the mentor-mentee matching process, evaluating whether considering personality traits yields superior pairings in comparison to matches predicated solely on capabilities and objectives.

This study scrutinizes the mentorship framework of a New Zealand-based corporation, replete with comprehensive profiles of mentees and mentors, encompassing their professional designations, competencies, and targets. Participants in this study engaged in the 16-personality type assessment, and the outcomes of this assessment were integrated into the data set. This enriched dataset serves as the foundation for a novel, hybrid methodology that marries collaborative-based learning through the application of a Convolutional Neural Network (CNN) machine learning algorithm with content-based learning modalities. The ambition of this hybrid system is to forecast the most fitting mentor for a mentee, considering an amalgamation of their skill sets, objectives, and personality classifications.

The motivation for this research stems from an examination of current mentorship programs and the strategies used to connect mentors with mentees. An extensive literature review has been undertaken to survey and assess the diverse array of mentoring arrangements that exist, paying special attention to the matching processes utilized. The objective of this study is to perform a critical comparison between traditional matching algorithms and those that incorporate deep learning approaches, evaluating the degree to which machine learning has been applied in the context of mentorship program development. Furthermore, this research intends to determine the presence and impact of integrating personality types into these matching systems. Consequently, the primary research question is focused on the effectiveness of deep learning and personality-based matching in improving the dynamics and success of mentor-mentee relationships within organized mentorship programs. Consequently, this study aims to address this research gap by exploring the following research questions:

RQ1 - What are the key factors necessary for cultivating a productive mentor-mentee relationship?

RQ2 - To what extent does the hybrid machine learning model effectively facilitate suitable mentor-mentee pairings considering skills, goals, and personality traits?

RQ3 - How can personality types be utilized as an attribute for identifying the most compatible mentor-mentee pairs?

RQ4 - What are the benefits of employing machine learning algorithms to recommend mentors based on various factors?

Chapter 2

Literature Review

The literature review was initiated by employing keywords such as 'mentor', 'mentee', 'mentorship', 'matching', and 'pairing', resulting in a substantial volume of outcomes. This segment is organized into categories encompassing literatures that address the matching process, those that focus on deep learning or machine learning algorithms, personality types, as well as filtering and similarity methods. However, it became apparent that using these keywords made it challenging to find literature specifically discussing matching algorithms or deep learning.

To broaden the scope of the research, additional keywords such as "two-sided matching algorithm" and "recommendation system" were incorporated, yielding some fruitful results. These literature sources provided insights into how matching algorithms were designed, leading to the identification of further relevant keywords such as "similarity calculations", "content-based filtering", and "collaboration filtering".

Considering that the primary focus of this research pertains to incorporating personality types in the matching process, keywords such as "personality type", "Myers-Briggs", and "matching" were utilized in conjunction to uncover literature discussing the impact of personalities on the matching of individuals. Notably, there was a degree of overlap with initial literature focused on mentorship programs, which, while pertinent, did not address the utilization of algorithmic methodologies.

The literature findings are organized into four sections: qualitative research on mentee-mentor pairing, deep learning or matching algorithms, personality types, and filtering methods and recommendation systems. Figure 1 graphically presents the keyword analysis, elucidating a distinct observation: research articles centered on personality did not intersect with those discussing matching algorithms or machine learning. Keywords inclusive of "matching algorithm", "mentor-mentee pairing", and "machine learning" were pivotal in identifying literature that was both relevant and informative for this investigation.

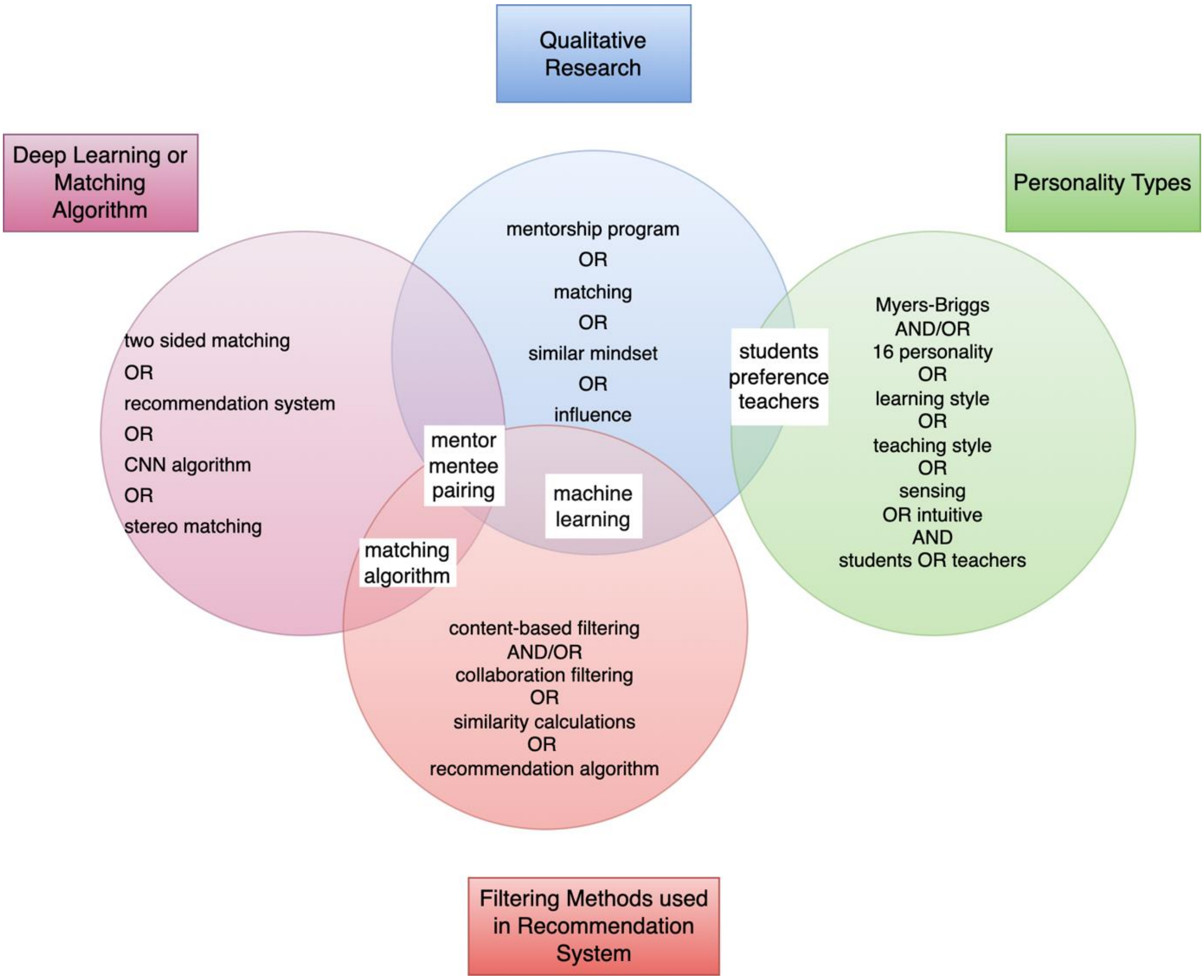

Qualitative Research
Deep Learning or Matching Algorithm
Personality Types
mentorship program OR matching OR similar mindset OR influence
two sided matching OR recommendation system OR CNN algorithm OR stereo matching
students preference teachers
Myers-Briggs AND/OR 16 personality OR learning style OR teaching style OR sensing OR intuitive AND students OR teachers
mentor mentee pairing
machine learning
matching algorithm
content-based filtering AND/OR collaboration filtering OR similarity calculations OR recommendation algorithm
Filtering Methods used in Recommendation System

Figure 1 - Literature Review Keywords Summary

2.1 Qualitative research findings based on mentorship program

The literature on matching mentors and mentees primarily focuses on the factors to consider during the pairing process. Deng et al. (2022) underscore the salience of deep-level congruences, developmental considerations, and mutual contribution in the matching process, echoing similar theoretical frameworks posited by Hee et al. (2020), who additionally scrutinize the integrity of study designs in health education mentorship.

Poulsen (2013) diverts the lens toward the overarching advantages of mentoring for all involved parties, underscoring the pivotal role of defined expectations and collaborative conventions—an outlook that finds harmony with Barrett et al. (2017), who delineate active participation, effective discourse, and congruent interests as the cornerstones of successful mentoring relationships, while concurrently acknowledging the enhanced hurdles faced by mentees.

In a more specialized context, Heppe et al. (2018) explore mentorship paradigms for individuals with disabilities, championing the importance of intersecting experiences and interests, a sentiment that finds resonance in Taylor's (2020) examination of positive relational dynamics within the mentorship frameworks operative in problem-solving courts. Wulf et al. (2021) probe into the influence of personality attributes in the realm of academic medicine mentorship, positing that the harmony or discordance of these traits could be instrumental—a notion that indirectly correlates with Vance et al.'s (2017) focus on vocational interests and lucid communication as essential ingredients for successful matches in statistics.

Pinilla et al. (2015) offer a pragmatic exploration of an online matching apparatus, utilizing weighted correlation algorithms for pairing medical students, showcasing a technological methodology that is not explicitly paralleled in other scholarly works.

Schwartz et al. (2022) broach the subject of gender dynamics within mentorship, unveiling an area that is not extensively canvassed in the extant literature, thereby highlighting a lacuna in the current understanding of gender homophily within mentorship initiatives. Vance et al.'s (2017) methodical guide to mentorship programs—while delivering a structured approach encompassing purpose delineation and effectiveness evaluation—also reinstates the criticality of aligning aspirational goals and interests among participants, a recurrent theme woven throughout the body of research.

In summation, the various academic inquiries into mentor-mentee matching, while diverging in their focal points—from shared life experiences and personality characteristics to vocational aspirations, technological innovations, and gender considerations—collectively converge on a consensus regarding the paramountcy of compatibility, engagement, and the intricate interplay of multifarious elements in the creation and sustenance of prosperous mentorship bonds.

2.2 Deep Learning and Machine Learning Algorithms

In the investigation of matching algorithms and methodologies, a plethora of approaches have been identified, spanning diverse disciplines. Haas and Hall (2019, 2018) proffer insights into two-sided matching scenarios, with a particular focus on the mentor-mentee dynamic within the sphere of higher education. Their findings extol the virtues of multi-objective heuristics, particularly the GATA-Mixed heuristic, over traditional approximation algorithms, delineating a clear performance disparity. This stands in stark contrast to Schäfer et al., (2016) exposition on medical education mentoring programs, where the efficacy of online self-matching and advisor-assisted matching was found to be indistinct, thereby proposing that both paradigms are equally equipped to cultivate high-caliber mentorships and scholastic achievement.

In a departure from the aforementioned studies, Ahmed et. al (2021) investigate the MyNRMN platform, positing that the platform's recommendations are predicated on similarity calculations rather than the application of machine learning techniques. This represents a shift from the complex algorithmic trend, advocating for a straightforward and similarity-based matching approach. Complementing this perspective, Bielczyk et al. (2018) discuss a semi-automatic online mentoring system that similarly eschews sophisticated machine learning algorithms in favor of a Python-based algorithm for pairing individuals. However, the absence of comprehensive validation methods in their work obfuscates a direct comparison of its effectiveness with other studies.

Hoenen and Kolympiris (2019) redirect the discourse to the realm of networking, examining the impact of academic insiders with governmental experience on the research productivity of nascent scientists, thus introducing a human-centric component to the discussion of matching that transcends algorithmic boundaries. Conversely, Hagler and Rhodes (2018) employ counterfactual analysis to elucidate the tangible benefits of natural mentoring relationships, offering an empirical strategy to comprehend the advantages of mentoring without the reliance on matching algorithms.

Tuifunovic's (2017) analysis delves into the realm of student-school matching algorithms, including the Boston algorithm, deferred acceptance algorithm, and top trading cycles algorithm. This study presents a theoretical evaluation of these algorithms, focusing on incentive compatibility, stability, and efficiency, thereby providing an alternative perspective on matching that is predicated on established rules instead of heuristics or machine learning.

In stark contrast, Hamid et. al (2022) comprehensive review of deep learning frameworks, particularly convolutional neural networks (CNNs), in stereo matching algorithms illustrates the efficacy of AI-based methodologies, especially in the domain of image processing. Gao and Spratling (2022) further this line of thought, discussing the application of deep feature space in template matching and highlighting the transformative potential of AI in enhancing the precision and robustness of computer vision tasks.

Lastly, the research conducted by Lozano et.al (2022) ventures into the critical medical field of liver transplantation, evaluating the role of AI in donor-recipient matching. They present an extensive appraisal of AI's prospects, addressing both the ethical and practical impediments, and surveying a variety of algorithms and their contributions to the matching process.

In summation, the corpus of research showcases a broad spectrum of methodologies, ranging from heuristic-based to AI-driven protocols. The comparative analyses underscore a nuanced comprehension that while certain sectors may glean benefits from sophisticated AI techniques, others may achieve similar success with more elementary, rule-centric, or similarity-driven approaches.

2.3 Personality Type

The landscape of research concerning the interplay between personality types and learning styles is marked by heterogeneous outcomes and inferences, reflecting the diverse methodologies and contexts of these studies. The work of Örtenblad et al. (2017) involving a survey of business students and educators reveals that akin personality types between students and teachers do not necessarily correlate with improved understanding, positing that pedagogical approaches and student involvement may be more indicative of learning success. In contrast, the study by Chen & Hung (2012) delineates a prominent association between personality dimensions—especially introversion/extroversion and sensing/intuition—and the learning strategies adopted by individuals learning English, thus advocating for the incorporation of personality considerations in the construction of language education curricula.

Vincent & Ross (2001) champion the importance of recognizing the heterogeneity of learning styles and personality types to craft educational experiences that are both individualized and effective. They posit that discerning these variances can empower educators to amplify students' strengths and mitigate their limitations, thereby refining the educational endeavour.

Further, Foroozandehfar & Khalili (2019) detect significant linkages between personality types and reading fluency in English as a Foreign Language (EFL) learners, insinuating that personality traits and learning styles may be predictive of reading acumen. Although the mechanisms underpinning this relationship are not fully elucidated, it is implied that certain personality characteristics may exert influence on reading engagement and methodologies.

Hemdan et al. (2022) underscore the significance of ascertaining personality traits in architecture students to encourage creative capacities, observing that students with an intuitive predisposition might display an enhanced creative impulse, whereas those with a sensing preference might derive benefits from the analysis of exemplary designs.

Brownfield (1993) explored the application of the Myers-Briggs Type Indicator (MBTI) within educational settings, emphasizing that cognizance of a student's proclivities on the MBTI spectrum can provide educators with insights into learning styles and inform the tailoring of instructional strategies.

Moreover, Balkis & Isiker (2005) delve into the nexus between personality types and thinking styles, revealing gender-based disparities in thinking styles and their correlation to academic specializations. This research introduces an additional layer of complexity by examining the confluence of personality, thinking styles, gender, and educational disciplines.

Collectively, these scholarly endeavours highlight the intricate and multifarious connections between personality types, learning styles, and educational achievements. They suggest that while the impact of personality on the learning process may vary, it constitutes an element of educational design and pedagogy that merits attention and incorporation.

2.4 Filtering Methods used in Recommendation System

The exploration of diverse methodologies across distinct studies reveals the pivotal role of filtering methods in the optimization of recommendation systems, crucial for enhancing user experience through personalized content delivery. Othman's (2023) research exemplifies the efficacy of content-based filtering within sports centre web applications, utilizing users' recent booking data to inform recommendations. This method's reliance on TF-IDF and Cosine Similarity demonstrates content-based filtering's precision in managing textual datasets and providing targeted suggestions.

Jayapal et al. (2023) explore a different avenue, presenting a hybrid recommendation system for a student-oriented platform that streamlines book selections by merging content-based with collaborative filtering techniques. This combined approach employs historical checkout and search data, offering a comprehensive system that potentially mitigates the drawbacks of singular methods by accounting for both item features and user-user interactions, leading to more refined recommendations.

Rani et al. (2023) highlight the advantages of collaborative filtering within e-commerce, devising a system that customizes product suggestions based on user purchase history and peer interactions. This strategy echoes Prince et al. (2023), who integrate collaborative filtering and Naive Bayes algorithms to advise students on course and job opportunities, employing a hybrid model to harness the strengths of multiple algorithms for superior personalization.

Mulyana et al. (2023) also demonstrate the fusion of content-based and collaborative filtering, augmented by the TF-IDF algorithm, to assist MSMEs in identifying trending product concepts, displaying the value of diverse filtering methods in catering to varied user needs and preferences.

Introducing a distinct perspective, Zanon et al. (2022) employ the WordRecommender algorithm, rooted in natural language processing (NLP), to analyse movie reviews and produce recommendations based on semantic closeness ascertained by a neighbourhood algorithm. This technique emphasizes the role of similarity algorithms in text-centric recommendation systems.

Synthesizing these findings, it is evident that filtering methods in recommendation systems are multifarious, each with unique strengths. Content-based filtering excels with textual information and user behavioral history, while collaborative filtering capitalizes on user interactions and similarities. Hybrid models, which amalgamate these methods, seem to present a more sophisticated and effective solution by encompassing a wider range of determinants. The integration of similarity algorithms like WordRecommender further amplifies the capability of recommendation systems to provide accurate and relevant suggestions.

P et al. (2017) showcased the application of cosine similarity in linguistic tasks, such as paraphrase detection in the Malayalam language, highlighting its potential for natural language processing applications including plagiarism detection and machine translation.

Barve & Saini (2023) extended the application of these methods to develop an advanced algorithm that incorporates multiple techniques for fact-checking, a critical tool in the fight against misinformation. Their approach presumably combines semantic analysis, credibility assessment, and data verification to heighten the precision of fact-checking systems.

Rezende et al. (2022) combined NLP with technological forecasting to scrutinize patent documents, offering insights into upcoming trends and innovations. This integration is essential in rapidly evolving industries where maintaining a competitive edge hinges on foresight.

In a related vein, Vorreuther and Warin (2021) utilized patent-based network analysis to investigate innovation dynamics within China's pharmaceutical industry, employing Jaccard similarity indices to gauge technological interconnectivity over time, offering a rich dataset for further study. Their research emphasizes the importance of similarity measures in understanding patterns of knowledge creation and the interrelationship of technological domains within an industry.

These studies collectively underscore the transformative power of similarity measures and algorithmic advancements across various research fields. From linguistic analysis to fact-checking, from forecasting technological progress to dissecting innovation dynamics, the strategic use of these techniques continues to drive scholarly exploration and practical developments forward.

2.4.1 Summary

The survey of the literature elucidates a broad application of algorithms in recommendation systems, with a notable inclination towards a hybrid approach, as summarized in Table 1 and discussed in Section 2.4. The majority of the studies integrate collaborative and content-based filtering algorithms to tailor user experiences effectively. Yet, a subset of authors enhance their systems' precision by incorporating similarity calculation algorithms, such as cosine similarity, TF-IDF similarity, or the Jaccard similarity algorithm.

An intriguing observation is the distinct use cases for these algorithms across the literature, underscoring their versatility. Whether applied to sports centre web applications, student book selection platforms, e-commerce product suggestions, or job and course recommendations, these algorithms demonstrate adaptability to various contexts. The addition of similarity measures not only refines the accuracy of the models but also showcases the potential for these algorithms to address diverse challenges and requirements in recommendation systems.

	Cited Paper	**Collaborative Filtering Algorithm**	**Content-based Filtering Algorithm**	**Similarity Calculation Algorithm**	**Other Machine Learning Algorithm**
1.	Othman (2023)		✓	Cosine Similarity, TF-IDF Similarity	
2.	Jayapal et al., (2023)	✓	✓		
3.	Rani et al. (2023)	✓			
4.	Prince et al., (2023)	✓			Naïve Bayes Algorithm
5.	Mulyana et al., (2023)	✓	✓	TF-IDF Similarity	
6.	Zanon et al., (2022)			Word Recommender	
7.	Thongdeelert et al., (2022)				Decision trees, Random forest
8.	Quarashi et al., (2020)			TF-IDF Similarity	K-Means
9.	P et al., (2017)			Cosine Similarity	

10.	Vorreuther and Warin (2021)			Jaccard Similarity	

Table 1 - Summary of Recommendation System Algorithms used in Literature

Chapter 3

Methodology

3.1 Dataset

The mentorship program within the New Zealand IT company is a structured and long-standing initiative aimed at fostering personal and professional development among full-time employees. The program's goal is to pair employees with mentors who can offer personalized advice, expertise, and support to help them achieve their individual growth objectives.

The mentor selection process within this program is a meticulous operation overseen by the company's senior management. To create conducive pairings, the leadership meticulously evaluates mentees' skills and professional ambitions, pairing them with mentors who have the experience and knowledge to offer the requisite mentorship.

Recognizing the potential for a single mentor to positively influence multiple mentees, the program is designed to permit such multiplicative pairings. This flexible structure respects the diverse mentorship needs and varying time commitments of different mentees, allowing mentors to effectively allocate their time and mentorship efforts.

Regular and structured meetings are integral to the program, offering a consistent opportunity for mentors and mentees to review progress toward the mentees' annual goals. These goals encompass both professional and personal development, reflecting the program's commitment to holistic growth.

For research purposes, the dataset utilized includes comprehensive information about participants' skills, objectives, and personality profiles. To adhere to ethical and privacy standards, all personal identifiers have been scrupulously removed, and specific pairings that could compromise individual anonymity have been excluded. This meticulous anonymization ensures the confidentiality of participant data and upholds the research's integrity.

The company has provided explicit consent for the use of this anonymized dataset in the research, recognizing the potential insights the study could offer to the mentorship field and the advancements in machine learning. The research harnesses this dataset to investigate and confirm the efficacy of the proposed hybrid machine learning model, aiming to improve the mentorship matching process while rigorously maintaining privacy and confidentiality standards.

The Myers-Briggs Type Indicator (MBTI) is a psychological assessment tool that is rooted in the typological theories of Swiss psychiatrist Carl Jung. It has been widely adopted in various domains, including organizational development, personal coaching, and team building, due to its ability to categorize individuals into distinct personality types. The MBTI posits that each person has a natural preference within four dichotomies, leading to 16 possible personality types that provide insight into how people perceive the world and make decisions (Myers & Myers, 2010).

Each of the four dichotomies represents a spectrum between two opposing preferences:

Extraversion (E) - Introversion (I): This dimension pertains to the source and direction of an individual's energy expression. Extraverts tend to be energized by interaction with others and external activities, while introverts typically find energy in solitary activities and internal reflection.

Sensing (S) - Intuition (N): This dichotomy relates to the means by which one gathers information. Sensing individuals are more likely to focus on the concrete, factual information gathered through their senses, while intuitive individuals tend to rely on patterns, abstract connections, and interpretations.

Thinking (T) - Feeling (F): This pair deals with decision-making processes. Thinkers tend to base their decisions on logical analysis and objective criteria, whereas feelers make decisions guided more by personal values and the consideration of others' feelings.

Judging (J) - Perceiving (P): This aspect concerns an individual's approach to structure and planning in their environment. Judging individuals typically prefer a planned, organized approach to life, while perceiving individuals are more inclined to be flexible, spontaneous, and adaptable.

Each personality type, as identified by a combination of these four preferences, is represented by a four-letter code. For example, an ISTJ would be an individual who primarily demonstrates Introversion, Sensing, Thinking, and Judging characteristics. These types provide a framework for understanding the various ways in which individuals engage with their environment, process information, and interact with others.

Understanding an individual's MBTI type can be instrumental in several contexts. In the workplace, for instance, knowledge of employee personality types can help in constructing teams with complementary skills, improving communication strategies, and tailoring leadership approaches to better suit the preferences of team members.

In the context of mentorship programs, the MBTI can be especially valuable. By identifying the personality types of mentors and mentees, the program can facilitate pairings that are more likely to have compatible communication styles, values, and approaches to learning and problem-solving. This can enhance the effectiveness of the mentorship, ensuring that mentees receive guidance that resonates with their personal inclinations and fosters a more productive and satisfying mentor-mentee relationship.

Grasping the essence of the MBTI dichotomies provides a revealing glimpse into the complexities of an individual's personality, shedding light on the intricate interplay of traits that govern behavior, reactions, and interactions with others. This depth of self-insight is a powerful tool for personal development, enabling individuals to navigate their inner landscape with greater clarity and to harness their innate strengths more effectively. By understanding one's own MBTI type, people can embark on a journey of self-discovery that leads to more purposeful personal and professional growth.

When it comes to career orientation, this knowledge becomes particularly valuable. It helps individuals in selecting roles and environments that align with their natural tendencies, allowing them to thrive in their chosen fields. For instance, a person who identifies as an extroverted thinker may find fulfillment in dynamic, fast-paced roles that involve problem-solving and team collaboration, while an introverted feeler may prefer positions that allow for deep reflection and contribute to the welfare of others. In this way, an understanding of one's MBTI type can act as a compass, guiding career choices that optimize job satisfaction and success.

Beyond the personal sphere, the MBTI enhances interactions with peers and colleagues by providing a framework for recognizing and appreciating the varied ways in which people perceive the world and make decisions. This fosters a more inclusive and empathetic workplace, where differences are not merely tolerated but valued for the diverse perspectives they bring. A manager aware of their team's MBTI types, for example, can tailor communication and motivate each member effectively, while team members can adjust their collaborative strategies to leverage the group's collective strengths.

In the broader context of interpersonal relationships, the insights offered by the MBTI can greatly enrich the quality of interactions. By understanding the psychological underpinnings of different personality types, individuals can develop more nuanced communication skills, leading to deeper connections and a greater sense of mutual understanding. Whether in teamwork, mentorship, or personal relationships, the MBTI serves as a key to unlocking more harmonious and productive interactions, enabling individuals to build stronger, more authentic connections that are rooted in a profound comprehension of human diversity and complexity.

The research study operates under the premise that the mentorship program's existing manual pairing process, carefully executed with expertise and in-depth knowledge, yields the most beneficial mentor-mentee matches. The program's leadership harnesses their profound insights into the company's personnel, as well as an understanding of the individual's professional and personal characteristics, to establish these pairings.

Given this context, the manual mentor-mentee pairings established by the program's leadership are considered the gold standard for optimal match quality, serving as a yardstick for evaluating the machine learning algorithm's recommendations. The algorithm's performance will be gauged by the degree of congruence between its automated pairings and the selections made by the company's leadership. This comparison aims to ascertain the machine learning model's effectiveness in emulating or potentially improving upon the results delivered by the human-led process.

The evaluation of the algorithm will involve several critical metrics. Accuracy is a primary measure, reflecting the frequency with which the algorithm's pairings coincide with the manual choices. A high accuracy suggests that the machine learning model is adept at recognizing pairings that correlate with the expertise-based human judgment.

Precision is another vital metric, indicating the fraction of algorithmic pairings that are accurate according to the benchmark. Precision is particularly indicative of the model's capability to identify the most fitting mentors for each mentee accurately.

Recall, also known as sensitivity, measures the algorithm's capacity to identify all actual positive pairings correctly. For this study, it would pertain to the model's ability to detect all potential high-quality matches from the available mentor pool for every mentee.

Additional evaluation parameters may include the diversity of the recommended matches, the satisfaction levels of participants with the algorithmic pairings, and the impact of these pairings on the participants' professional growth within the mentorship program.

The overarching objective of this research is to examine if a machine learning algorithm can enhance the pairing process while maintaining match quality. By comparing the model's predictions to expert-driven manual pairings, the study not only seeks to validate the machine learning model but also to explore the potential for augmenting the mentorship program with AI-driven tools. Should the algorithm demonstrate effectiveness, it could offer a scalable, efficient, and potentially more impartial approach to creating mentor-mentee connections, thereby improving the mentorship experience for all participants.

3.2 Features and Supervised Learning Method

A Convolutional Neural Network (CNN) represents a class of deep learning algorithms that are primarily employed for tasks such as image classification and recognition within the domain of machine learning. The design of CNNs is influenced by the hierarchical structure found in the human visual cortex and is widely applied to a range of computer vision challenges (Wang et al., 2023). CNNs utilize a series of layers composed of interconnected processing nodes that apply convolutional operations to input data to distill pertinent features. These extracted features are then processed through pooling layers, which serve to reduce the data's dimensionality while retaining essential information.

Collaborative filtering, meanwhile, is a well-established recommendation algorithm designed to forecast user preferences for various items by harnessing the likes and dislikes of similar users (Zaremarjal & Yiltas-Keplan, 2021). The integration of CNNs in collaborative filtering endeavors has recently shown considerable potential. Titles and headings within the text should refrain from using abbreviations unless their use is absolutely necessary.

Convolutional Neural Network (CNN) algorithms have garnered significant traction in Natural Language Processing (NLP), acclaimed for their proficiency in detecting local patterns and structural hierarchies in text-based data (Wang et al., 2019). These algorithms implement a series of filters to perform convolutions on textual input, extracting salient features across different layers. CNNs' distinctive capacity to dissect and interpret the contextual and semantic nuances of text empowers them to deeply analyze and understand the data.

Content-based filtering is a technique in machine learning that recommends items by scrutinizing their content features. It involves examining the content of the items and juxtaposing it with a user's expressed preferences to generate tailored recommendations (Othman et al., 2023).

Within the sphere of mentorship, content-based filtering can be leveraged to align mentors with mentees by assessing mutual interests and predilections. This method scrutinizes the characteristics of mentors and mentees, including facets like industry tenure, distinct skills, academic achievements, and vocational objectives, to facilitate personalized pairing suggestions. Utilizing machine learning algorithms, the framework can suggest mentors to mentees sharing analogous content attributes, thereby optimizing the potential for a fruitful mentorship bond. This strategy allows mentees to discover mentors with the requisite expertise and knowledge, whilst mentors can engage with mentees within their realms of specialization. In essence, content-based filtering equips both mentors and mentees with the opportunity to forge impactful and constructive mentorship engagements.

3.3 Proposed Hybrid Approach

The Convolutional Neural Network (CNN)-based recommendation system designed for this research adopts a pioneering hybrid approach, integrating the strengths of both collaborative and content-based filtering to refine the mentor-mentee matching process. This dual strategy leverages the synergistic advantages of the two filtering methods to deliver a sophisticated and discerning pairing mechanism.

Within the model, collaborative filtering is realized through the CNN's analysis of user interactions and relationships in the mentorship program. By reviewing historical mentor-mentee pairings, feedback, and the outcomes of these relationships, the CNN uncovers latent patterns and preferences that may not be overtly observable. This capability enables the CNN to forecast potential successful partnerships by drawing on the learned patterns and results within the dataset. The CNN's proficiency in identifying intricate patterns positions it as an apt choice for capturing the nuances of user interactions.

Conversely, content-based filtering concentrates on the specific attributes and traits of the mentors and mentees. These include the mentors' professional expertise, the mentees' career ambitions, and the personality types of both, as determined by assessments like the MBTI. This facet of the recommendation system provides personalized suggestions that resonate with each mentee's distinctive needs and inclinations, ensuring the recommended mentors have the appropriate expertise to foster the mentees' professional development and that their personalities are likely to synergize well.

The research by Vedaswi et al. (2023) highlights the efficacy of combining these approaches. By employing CNNs to discern complex relationships within collaborative filtering and merging it with the detailed attribute analysis of content-based filtering, a comprehensive recommendation system is developed.

The expected outcome of this integrated approach is a recommendation system that is both adaptive and individualized. It is adaptive because it can learn from the dynamic patterns of user interactions in the mentorship program, and individualized because it can cater to each participant's particular traits. By harmonizing the broad insights from collaborative filtering with the specific insights from content-based filtering, the CNN-based recommendation system is designed to offer contextually apt and nuanced mentor-mentee pairings. This, in turn, aims to improve the effectiveness and satisfaction of the mentorship experience for all participants involved.

Incorporating similarity computation methods is essential in enhancing recommendation systems' ability to accurately identify compatible entities, such as mentor-mentee pairs. Cosine similarity and Jaccard similarity are two prevalent techniques for quantifying the resemblance between entities based on their attributes or preferences.

Cosine Similarity: This method assesses the cosine of the angle between two attribute vectors within a multi-dimensional space. The cosine similarity score ranges from -1 to 1, with a value close to 1 indicating a high degree of similarity and a value near 0 or negative suggesting low similarity. In mentor-mentee matching, vectors may represent individuals' skills, experiences, or characteristics.

Jaccard Similarity: Jaccard similarity calculates the ratio of the intersection over the union of two attribute sets. This method is particularly useful when attributes can be represented as sets, such as the presence or absence of certain skills or personality traits. The Jaccard index ranges from 0 to 1, where 0 denotes no shared attributes and 1 signifies identical attribute sets.

The research by Mulyana et al. (2023) underscores the significance of similarity computation in improving recommendation system accuracy. By comprehensively assessing the relevance and commonality of attributes between users, recommendation systems can offer more precise and relevant matches.

For the current research, both cosine similarity and Jaccard similarity will be compared to determine which is more effective for mentor-mentee pairings. This comparative analysis will apply both methods to the same dataset and evaluate the match quality they yield.

The performance of each similarity computation technique will be evaluated against various metrics, including:

- Accuracy: The frequency with which the system's recommendations coincide with actual successful pairings.
- Precision: The fraction of the system's recommendations that are deemed suitable.
- Recall: The system's ability to capture all potential successful matches within the dataset.
- Robustness: The system's performance when dealing with incomplete or noisy data.

This comparative study will offer insights into the strengths and weaknesses of each similarity computation method relative to the mentorship program's specific data characteristics. If the program's data are binary (presence or absence of skills/traits), Jaccard similarity may be more appropriate. Alternatively, if the data involve gradations or levels of proficiency, cosine similarity could be more fitting.

The findings from this analysis will not only identify the more suitable algorithm for this dataset but also contribute to a broader understanding of optimizing similarity computation methods for varying types of recommendation systems. By matching the similarity computation method to the dataset's nature and the mentorship program's requirements, the research aims to further the development of sophisticated, precise, and personalized recommendation systems.

Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The TF part measures the raw frequency of a term in a document, while the IDF part scales down the importance of terms that appear frequently across documents, emphasizing unique terms. TF-IDF is beneficial for text analysis and information retrieval by allowing for the comparison of documents and identification of those with particular relevance based on the terms they contain.

In the context of mentor-mentee pairing, where the focus is on matching based on attributes like skills, objectives, and personality types, TF-IDF may not be the most applicable measure. This is due to the requirement of comparing profiles based on attributes rather than textual frequency. Thus, similarity measures such as cosine and Jaccard similarities are considered more fitting for this scenario.

Cosine similarity is particularly useful when mentor and mentee attributes can be represented as vectors in a multi-dimensional attribute space. Cosine similarity excels at assessing the orientation of these vectors, which corresponds to the pattern of attributes shared between profiles, as opposed to the magnitude of the vectors. This makes it an excellent choice for evaluating the similarity between the skill sets and experiences of mentors and mentees, as noted by Setiawan & Adnyana (2023).

On the other hand, Jaccard similarity is effective when dealing with binary or set-based data, such as whether a skill or trait is present or absent. It evaluates matches based on the proportion of shared attributes to the total number of unique attributes across both sets, making it suitable for scenarios where the presence of shared characteristics is the primary concern.

Overall, while TF-IDF is a powerful tool for text analysis, its application to the mentor-mentee pairing problem is limited. Instead, cosine and Jaccard similarities offer more relevant measures for this specific type of data, matching the attributes of individuals to form effective mentor-mentee pairs.

Jaccard similarity, grounded in set theory, is adept at applications where the order of elements is non-essential, and the primary concern is the shared content between sets. Panja & James (2020) highlight its effectiveness in tasks such as document deduplication, where the identification of duplicates relies on shared content rather than the sequence of that content. Recommendation systems also benefit from Jaccard similarity when identifying items with shared features, like movies sharing similar genres or products with common attributes. Furthermore, in clustering processes, Jaccard similarity is valuable for grouping items based on common characteristics, emphasizing the presence or absence of attributes over their order or frequency.

For the present study, the application of cosine and Jaccard similarities is critical for refining the recommendation system, especially for mentor-mentee matching. These methods are selected due to their suitability for the dataset's nature, contrasting with TF-IDF's focus on term significance within text documents.

Cosine similarity will be used to compare the orientation of attribute vectors in a multidimensional space, enabling the analysis of profile alignment between mentors and mentees. Jaccard similarity will be applied to feature sets, assessing the extent of shared attributes without the influence of attribute frequency or sequence.

The research will undertake a meticulous application of both similarity measures, followed by an extensive evaluation to ascertain their effectiveness in the mentorship program's recommendation system. This evaluation will clarify the impact of each method on match suggestion quality, guiding the creation of a more accurate and context-sensitive pairing mechanism. The strategic use of cosine and Jaccard similarities is anticipated to produce a recommendation system that is precise and attuned to the specific context, thereby improving the mentorship experience for participants in the IT company's program.

The study will employ cosine and Jaccard similarity techniques to create a specialized hybrid model designed to improve the mentor-mentee matching process, drawing from the rich research presented in the literature review. These techniques are selected for their ability to capture the complexities of the mentorship relationships within the dataset, focusing on the diverse attributes and preferences that characterize these relationships.

The hybrid model will synergistically combine cosine and Jaccard similarity:

- Cosine similarity will be utilized to analyze the quantitative aspects of mentors' and mentees' profiles, such as skills and expertise, by comparing the orientation of vectors in a multidimensional space.

- Jaccard similarity will be employed to examine qualitative attributes like interests or personality types, where the binary nature of the attributes (present or absent) is more significant than their frequency.

The algorithm's development is anchored in the Python programming language due to Python's comprehensive libraries and frameworks, which are conducive to machine learning and data analysis. Python's readability and flexibility are advantageous for constructing intricate algorithms and supporting collaborative efforts within the research community.

Jupyter Notebook is chosen as the interactive platform for the algorithm's development and evaluation, offering a conducive environment for code writing, execution, and modular organization. Jupyter Notebook's ability to blend live code with visualizations and narrative text makes it an invaluable tool for data exploration, code sharing, and result presentation. It allows the research team to conduct iterative improvements, visualize data, and meticulously document the research journey, enhancing reproducibility and transparency.

The documentation compiled in Jupyter Notebook will encompass the code itself and an elaborate exposition of the methods and insights acquired throughout the model's evolution.

This approach, employing Python and Jupyter Notebook to build the hybrid recommendation model, reflects an adherence to accuracy, clarity, and innovative research methodologies. The resultant algorithm will demonstrate the efficacy of merging established similarity measures with advanced programming resources to address the intricate task of refining mentorship pairings within a professional environment.

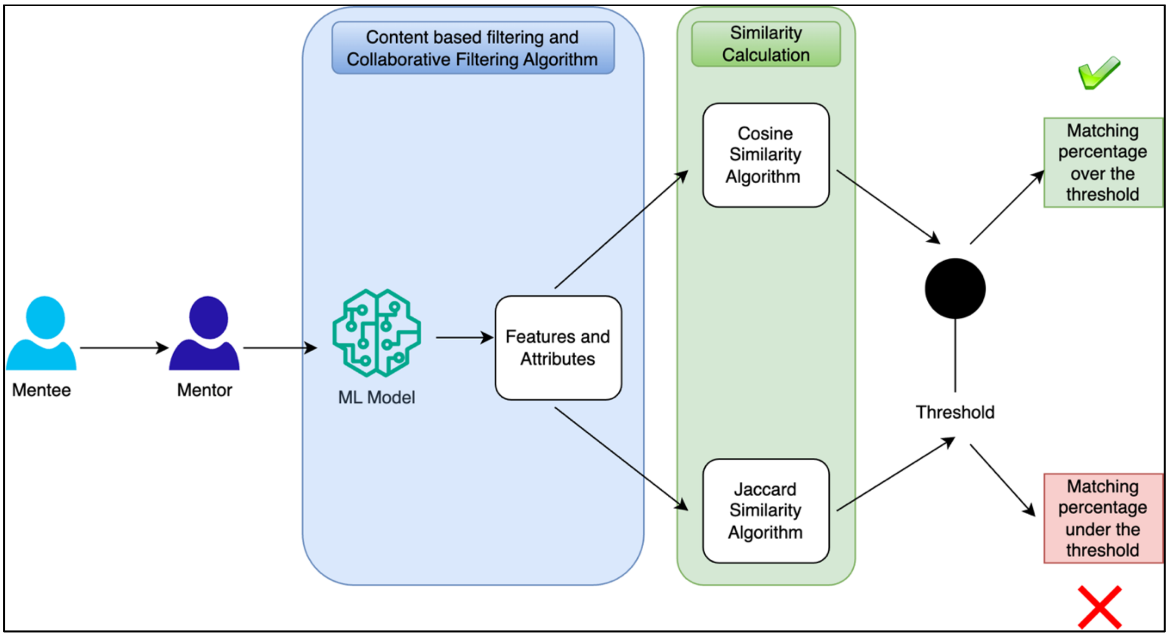

Content based filtering and Collaborative Filtering Algorithm
Similarity Calculation
Mentee
Mentor
ML Model
Features and Attributes
Cosine Similarity Algorithm
Jaccard Similarity Algorithm
Threshold
Matching percentage over the threshold
Matching percentage under the threshold

Figure 2 - Proposed Hybrid Approach

Chapter 4

Design and Development

4.1 Initial Data Format

The dataset provided by the corporation is fundamental for the creation and refinement of the machine learning recommendation system. The use of Comma-Separated Values (CSV) files is ideal for storing the tabular data required for this project due to its wide compatibility with various data analysis tools, including Python, which is instrumental for the research at hand.

Each row in the CSV files corresponds to a distinct entry, with critical data points that contribute to a detailed profile for each mentee. These data points encompass anonymized identifiers and a wealth of attributes, such as skill sets, career goals, and MBTI personality types, which are paramount for the recommendation process.

Similar data is included for mentors, detailing their competencies, job titles, and personality types, enabling the recommendation system to consider both skill alignment and interpersonal compatibility between mentors and mentees.

During data preprocessing, entries without a personality type were removed to ensure the recommendation system could evaluate the mentor-mentee dynamic comprehensively. Personality type information is vital for assessing the interpersonal aspect of the pairing process.

However, it is important to note that the dataset is relatively small, as it originates from a small business in New Zealand with only 22 mentor-mentee matches. This limitation necessitates the use of data augmentation techniques to enhance the training dataset. Data augmentation is a standard machine learning practice that increases the dataset's size and diversity, thereby mitigating overfitting and improving the model's ability to generalize to new data. Techniques include resampling, synthesizing new samples, or introducing variations to existing data.

The goal of data augmentation in this research is twofold: to prevent overfitting and to bolster the model's robustness by introducing a broader array of training examples. This approach is designed to ensure that the machine learning algorithm can make accurate and reliable recommendations despite the initial small dataset size.

By applying these data augmentation strategies, the research team intends to develop a fortified dataset that underpins the creation of a machine learning algorithm geared towards effective mentor-mentee matching. The enhanced dataset will facilitate the algorithm's analysis, leading to a recommendation system capable of navigating the intricacies of mentor-mentee pairings in a professional environment.

4.2 Calculate Jaccard and Cosine Similarity

The preprocessing phase is indeed essential for preparing the raw dataset for machine learning analysis. This phase involves a series of steps to clean, organize, and transform the data into a structured, machine-readable format. The transformation of the dataset into a DataFrame using Python's pandas library is particularly advantageous for this process due to pandas' powerful data manipulation and analysis capabilities.

After preprocessing and structuring the data into DataFrame format, similarity measures such as Jaccard similarity and cosine similarity are computed for each potential mentor-mentee pair. These measures quantify the similarity between mentor and mentee profiles based on the provided attributes.

Initially, the focus is on aligning skills and career goals. Cosine similarity is particularly effective here, as it measures the orientation of skill vectors in the vector space, identifying how closely the skills of mentors and mentees are aligned. Jaccard similarity can be used to evaluate the overlap in career goals, as it measures the similarity between sets, which is suitable if the career goals are categorized as such.

The second iteration incorporates personality types into the computation to assess compatibility on a deeper level. Since personality types impact communication, learning styles, and interpersonal relationships, their inclusion provides a more well-rounded analysis of potential mentor-mentee matches. Depending on how personality types are represented in the dataset, they may be encoded as categorical variables or vectors. Jaccard similarity might be used if the personality types are represented as binary attributes, indicating the presence or absence of certain traits. Cosine similarity could be helpful to discern patterns in personality dimensions and how they relate between mentors and mentees.

These iterative similarity computations, which increase in complexity by considering a broader range of attributes, demonstrate the study's dedication to creating a nuanced recommendation system. This approach ensures that the system considers all critical aspects of the mentorship relationship, enhancing the ability to suggest optimal pairings. Through this careful and progressive refinement, the research team aims to ensure the recommendation system is comprehensive and effective in facilitating successful mentor-mentee matches.

For the computation of Jaccard similarity, the intersection and union of sets were calculated using the formula:

$$Jaccard\ Similarity = \frac{Intersection}{Union}$$

The methodical calculation of Jaccard similarity scores for each potential mentor-mentee pair involves comparing their respective attribute sets, such as skills, expertise, career goals, and personality traits. The Jaccard index is computed by dividing the number of shared attributes (the intersection) by the number of all distinct attributes found in either set (the union), yielding a score between 0 (no similarity) and 1 (identical sets).

These individual Jaccard scores are then compiled into an overall compatibility score, reflecting the comprehensive match between mentors and mentees. This composite score is achieved by assigning weights to each attribute category, based on their relative importance as determined by the mentorship program's objectives. This weighted approach aligns with the findings of Panja & James (2020), who advocate for the differential weighting of attributes to more accurately reflect the factors that contribute to a successful mentor-mentee relationship.

For instance, should the mentorship program emphasize aligning mentors' and mentees' skills, the Jaccard score for this attribute category would carry more weight in the aggregated score. If personality traits are considered less crucial but still pertinent, their associated score would have a lesser weight. This weighting process ensures that the overall similarity score mirrors the program's emphasis on particular attributes.

The weighted aggregated scores are then used to rank potential mentor-mentee pairs, with higher scores indicating a greater degree of compatibility. This ranking is a crucial feature of the recommendation system, as it facilitates the identification of the most suitable matches based on the alignment of mentor attributes with mentee needs and preferences.

Incorporating weighted Jaccard similarity scores allows the recommendation system to create pairings that are not just based on commonalities but are also attuned to the mentorship program's strategic objectives. By thoughtfully adjusting the weight of each attribute, the system offers refined, data-informed match suggestions that are likely to result in productive and successful mentor-mentee relationships. This tailored matching process exemplifies the power of leveraging calculated metrics to foster well-aligned mentor-mentee pairings that align with the overarching goals of the mentorship initiative.

In the development of the mentor-mentee recommendation system, cosine similarity is used to quantify the textual compatibility between potential pairs. This process involves two iterations as outlined by Setiawan & Adnyana (2023). Initially, the focus is on the skills and career goals of mentors and mentees, which are converted into numerical vectors through TF-IDF, a technique that reflects the importance of words based on their frequency in a document relative to their frequency across a collection of documents.

Once the TF-IDF vectors are established, the cosine similarity between each mentor and mentee's skills and goals vectors is calculated. Cosine similarity measures the cosine of the angle between two vectors in a multidimensional space, providing a metric that ranges from -1 to 1, with values closer to 1 indicating a higher degree of similarity.

In the second iteration, personality types are integrated into the TF-IDF vectors, allowing the recommendation system to consider both professional alignment and the interpersonal dynamic between mentors and mentees. By accounting for personality types, the system aims to match individuals more holistically, increasing the chances of a successful mentorship relationship.

The mean cosine similarity for each potential pair is then determined from the individual cosine similarity scores, and this average measure reflects the overall compatibility across the various attributes included in the model. These mean cosine similarity scores are then added to the DataFrame as new columns, serving as labels that the machine learning algorithm will use for prediction and learning. The similarity scores inform the model about the relationships within the data that are indicative of successful pairings.

The final step involves evaluating the machine learning model's accuracy in predicting these similarity score labels. This assessment helps discern which similarity measure, be it Jaccard or cosine similarity, is more effective for the dataset. The model’s performance is scrutinized using precision, recall, and the F1-score, providing an in-depth understanding of its predictive strength.

The use of cosine similarity scores, alongside Jaccard similarity scores, presents a comprehensive approach to training the machine learning model. By analyzing various metrics, the research team can refine the recommendation system to ensure it is well-suited to identifying mentor-mentee pairs that have the highest potential for a successful, mutually beneficial mentorship experience.

4.3 Pre-processing the data

The tokenization of textual data is indeed an essential step in preparing the data for analysis by the machine learning algorithm. This process involves breaking down text into smaller units, or tokens, which represent words or phrases, making it easier for the algorithm to process and analyze the text.

For the mentorship program recommendation system, tokenization will be applied to the textual data fields within the DataFrame, such as the descriptions of skills, career goals, and personality types. The process might include normalization steps like converting text to lowercase, removing punctuation, and excluding common stopwords that typically don't carry significant meaning.

After tokenization, the resulting arrays of tokens for each entry in the DataFrame will likely vary in length. Given that machine learning models require inputs of consistent shape and size, the token arrays must be made uniform through padding. Padding involves filling the shorter arrays with a placeholder value (zeroes or a specific token) until they match the length of the longest array in the dataset. This uniformity is critical for batch processing in many machine learning algorithms, including neural networks.

Once the token arrays are padded to a consistent length, they become the input features for the machine learning model. This standardization is a key step in preparing the data, ensuring that the model can effectively learn from the patterns within the data without being affected by the variability in the original text lengths.

Tokenization and padding transform the unstructured text data into a format that is ready for machine learning analysis. This process is a crucial part of the data preparation pipeline, aiding in the development of an accurate and effective recommendation system. With the tokenized and padded DataFrame, the machine learning model is now set to be trained on the dataset, with the goal of accurately predicting successful mentor-mentee matches based on the similarity scores and other relevant features.

The division of the dataset into training and testing subsets is indeed a crucial step in machine learning model development. The 80/20 split you mentioned is a common approach that allocates the majority of the data (80%) for training the model, with the remaining 20% used to test and evaluate its performance.

During training, the machine learning model is exposed to the training data, which now includes the tokenized and padded features along with the similarity scores. The model learns by optimizing its internal parameters to reduce the discrepancy between its predictions and the actual outcomes. This optimization often involves minimizing a loss function—a measure of prediction error—using techniques like gradient descent.

Performance metrics are monitored throughout the training to gauge the model's accuracy. These may include precision, recall, F1-score, or mean squared error, depending on the specific task and model being used.

The testing phase is essential for validating the model's generalizability. The testing subset, which the model has not encountered during training, is used to assess how well the model can apply what it has learned to new, unseen data. The model's performance on the test set provides a more realistic indication of how it will perform in practical applications.

A representative test set ensures that the evaluation is meaningful and that the performance metrics reflect the model's likely behavior when deployed. A model that performs well on the test set is considered to have good generalization capabilities, a key quality for a recommendation system used in real mentor-mentee matching scenarios.

Splitting the data also aids in preventing overfitting, a common issue where a model learns the training data too well, including its noise and anomalies, to the detriment of its performance on new data. By holding out a portion of the data for testing, we can ensure that the model's predictions are truly based on the underlying patterns that are predictive of successful pairings, rather than memorizing the specific details of the training set.

The 80/20 dataset split strikes a balance that allows for comprehensive training of the model while maintaining a sufficient amount of data to evaluate its performance accurately. The insights gained from the test results play a vital role in fine-tuning the model before it is deployed to match mentors and mentees in the actual mentorship program.

4.4 Design the Machine Learning Model

Upon finalizing the data preparation, the construction of the CNN model for the machine learning algorithm commenced, incorporating the collaborative filtering approach. This model architecture was specifically crafted to handle the complex nature of mentor-mentee matching, leveraging the diverse input layers that include the skills, goals, and personality types of both parties. These input features were meticulously organized to allow for the content filtering method to analyze and match the profiles effectively.

The model's design utilized the previously calculated Jaccard and Cosine Similarity scores as the labels, guiding the CNN in its learning process. By treating these similarity measures as the target outcomes for each potential mentor-mentee pair, the model could focus on predicting the level of match based on the alignment of their respective attributes.

The training of the CNN model involved adjusting its parameters to optimize the prediction of these labels, striving to mirror the success of past mentor-mentee pairings. Through iterative training with the input data, the model 'learned' the intricate patterns that contribute to successful matches, with the end goal of applying this knowledge to suggest compatible pairings in a real-world environment. Once validated and tested, the CNN model stood ready to be an integral part of the recommendation system, harnessing the power of machine learning to enhance the mentorship program.

During the construction of the CNN model for the recommendation system, a sequential approach to model development was adopted, with successive iterations adding complexity to the input data.

In the initial iteration of the model, a single input layer was designed to accept and process the skills and goals of the mentors and mentees. This layer served as the foundation for the model, enabling it to learn from the most immediately quantifiable and objective attributes of the profiles. The skills and goals, likely represented as vectorized features, were the sole focus at this stage, providing a straightforward basis for the model to begin understanding the relationships between the different profiles.

Following the evaluation of the initial iteration, a subsequent iteration introduced an additional input layer to the CNN model. This new layer was specifically designed to account for the personality types of the mentors and mentees. By adding this layer, the model was expanded to take into consideration the more subjective and nuanced aspects of the profiles that could influence the compatibility of the mentor-mentee pairs. The inclusion of personality types aimed to create a richer, more rounded profile for each individual, allowing the model to explore deeper connections beyond professional skills and goals.

For both iterations, the calculated Jaccard and Cosine Similarity scores were incorporated as labels. This dual-label approach was critical for several reasons. Firstly, it allowed the model to learn to predict both types of similarity simultaneously, providing a more comprehensive measure of compatibility. Secondly, by including both labels, the research team could directly compare the model's performance with respect to each similarity measure. This comparison would be instrumental in determining which similarity measure—or combination of measures—most effectively captured the essence of a successful mentorship match within the context of the dataset.

The inclusion of both similarity labels also permitted a nuanced evaluation of the model's behavior. By analyzing the model's predictions against the Jaccard and Cosine Similarity scores, insights could be gained into how different features influenced the compatibility scores and, consequently, the pairing recommendations.

These iterative developments and the addition of multiple input layers and similarity labels were key to refining the CNN model's predictive capabilities. The approach ensured that the model was not only trained on a broad spectrum of relevant features but also that its effectiveness could be critically assessed, paving the way for an optimized and reliable recommendation system.

An epoch in machine learning is indeed one full cycle through the training dataset, where the model's parameters are adjusted to minimize the loss function. The loss function measures the discrepancy between the model's predictions and the actual target values, and minimizing this loss is key to enhancing the model's accuracy.

For the CNN model developed for the recommendation system, setting the number of training epochs to 15 is a strategic decision that takes into account the dataset's size and complexity, as well as the potential for overfitting. Overfitting is a concern when a model learns not just the underlying patterns but also the noise and outliers within the training data, which can impair its ability to generalize to unseen data. A limit of 15 epochs aims to enable adequate learning while minimizing the risk of the model becoming too tailored to the training set.

The batch size, which determines the number of training examples used before the model updates its parameters, is another important hyperparameter in the training process. A batch size of 32 represents a middle ground that allows for frequent updates, facilitating swift learning and better generalization due to the noise introduced by the smaller batch. This noise can act as a form of regularization, preventing the model from fitting the training data too closely.

By setting the number of epochs to 15 and the batch size to 32, the training process is optimized to balance learning efficiency with the quality of the model's generalization. These choices are designed to yield a CNN model that is well-equipped to provide accurate, reliable recommendations for the mentorship program, avoiding overfitting while ensuring that the model is sufficiently trained on the patterns present in the data.

4.5 Validating the Machine Learning Model

Once the model was created and trained, it underwent validation by incorporating the test data, which had been initially split into an 80/20 ratio. To enhance the robustness of the validation, the dataset was randomly divided during each iteration rather than employing a consistent separation method. This approach was adopted to bolster the validity of the model's performance metrics. Subsequently, the model made predictions on both labels. A comparative analysis was then conducted between the predicted values and the actual test values, thereby allowing the accuracy of each model to be determined.

The study by Schäfer et al., (2016) sheds light on an important aspect, namely that digitally matching mentors to mentees, as opposed to traditional matching methods, did not lead to improvements in mentor-mentee pairings. However, it did enhance the matching process by automating the traditional approach. Similar to the approach adopted in this paper, the traditional or manual matching process employed in this study was assumed to yield the most optimal outcomes. Consequently, the model results were evaluated against the original matching pairs. The high-level design of this study is illustrated in Figure 3.

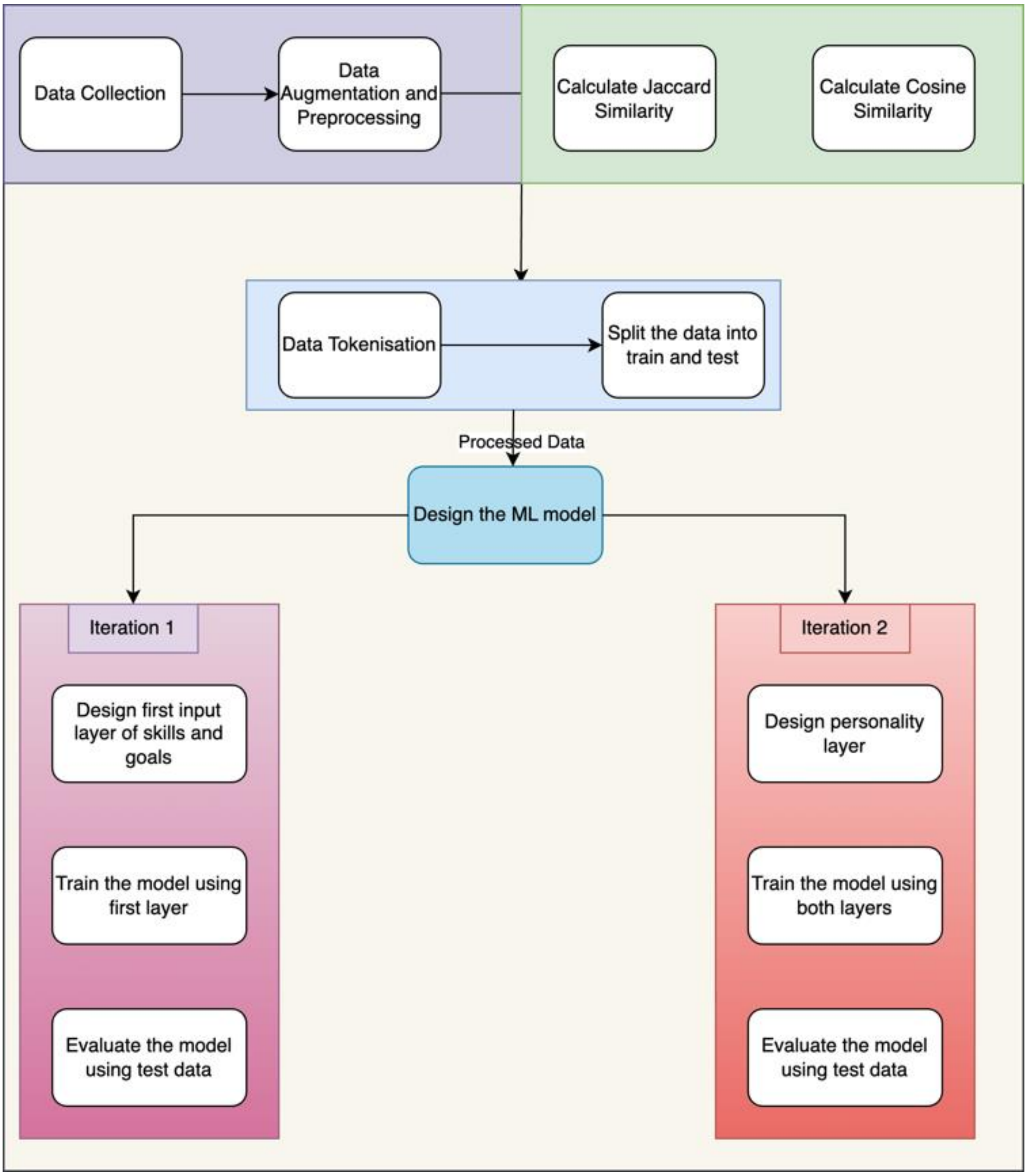

Data Collection
Data Augmentation and Preprocessing
Calculate Jaccard Similarity
Calculate Cosine Similarity
Data Tokenisation
Split the data into train and test
Processed Data
Design the ML model
Iteration 1
Design first input layer of skills and goals
Train the model using first layer
Evaluate the model using test data
Iteration 2
Design personality layer
Train the model using both layers
Evaluate the model using test data

Figure 3 - Methodology

Chapter 5

Results

In this study, a comprehensive dataset was constructed, documenting a myriad of critical attributes of individuals participating in a mentorship scheme. This compendium detailed the competencies, career aspirations, professional designations, and psychological archetypes of mentors and mentees. It is crucial to acknowledge that the submission of psychological archetype data by the participants was not mandatory within the framework of the mentorship scheme. Consequently, this led to the exclusion of certain dyads from the dataset due to incomplete information, thereby refining the scope of the investigation to 22 complete mentor-mentee dyads with all requisite data fields.

To counteract the constraints posed by the limited dataset size and to validate the sufficiency of the extant data for deploying machine learning techniques, data augmentation methods were utilized. These methods aim to synthetically enhance the dataset by creating additional, credible data points aligned with the original dataset, thereby amplifying its scope and heterogeneity.

Data augmentation is especially pivotal in machine learning contexts, where an expanded dataset can notably bolster the predictive efficacy and precision of the models in use. Through the deployment of advanced techniques to fabricate new mentor-mentee profiles, the research team was able to ensure that the dataset was ample enough to facilitate the development of potent machine learning models. These models are instrumental in discerning trends and forecasting results within the mentorship scheme with heightened accuracy.

The augmentation procedure entailed meticulous scrutiny of the extant data to generate supplementary, synthetic data points that accurately reflect the attributes of actual participants. This was accomplished by employing strategies such as resampling, perturbation, and the application of generative models, all aimed at preserving the authenticity and distribution of the original dataset while enlarging its magnitude.

The dataset, thus augmented, provided a more robust platform for in-depth analysis, granting a more intricate comprehension of the elements that underpin successful mentorship dyads. Consequently, the research team was empowered to extract more profound insights and propositions for the structuring and execution of prospective mentorship schemes, with the objective of nurturing more efficacious and gratifying mentor-mentee connections.

5.1 Iteration 1 – Hybrid Model without Personality Attribute

In the initial stage of the investigation, the research collective embarked on establishing a methodology to quantify the congruence between mentors and mentees concerning their respective competencies and professional ambitions. This endeavor leveraged two distinct similarity indices: Jaccard similarity and cosine similarity. These indices facilitate the quantitative assessment of the extent to which the aptitudes and aspirations of mentees correspond with the mentors' expertise.

To optimize the machine learning model for the study, the input stratum was augmented to incorporate the skills and objectives as features. The similarity values, derived from juxtaposing the mentors' skills with the mentees' skills and objectives, were then adopted as markers to orient the machine learning algorithm in its predictive efforts. This structured approach endowed the model with the capacity to discern patterns in efficacious mentorship pairings, predicated on the compatibility of the mentee's ambitions and abilities with the mentor's proficiency.

Within the exposition of the study, Figure 4 delineates the preliminary precision of the training model after the integration of these similarity indices. The figure likely elucidates the adeptness of the model in forecasting successful pairings by correlating the projected similarity scores with the actual, historically verified successful mentor-mentee matches. The initial precision provides a metric for gauging the model's effectiveness, premised on the designated features and markers.

The incorporation of these similarity computations in the inaugural iteration is pivotal as it lays down the fundamental parameters for the model's learning trajectory. By scrutinizing the initial precision as depicted in Figure 4, the researchers are equipped to pinpoint areas where the model may necessitate further refinement or supplementary data to enhance its prognostic accuracy. This cyclical process of model training and evaluation is instrumental in cultivating an efficacious mechanism to aid in the establishment of productive mentor-mentee affiliations.

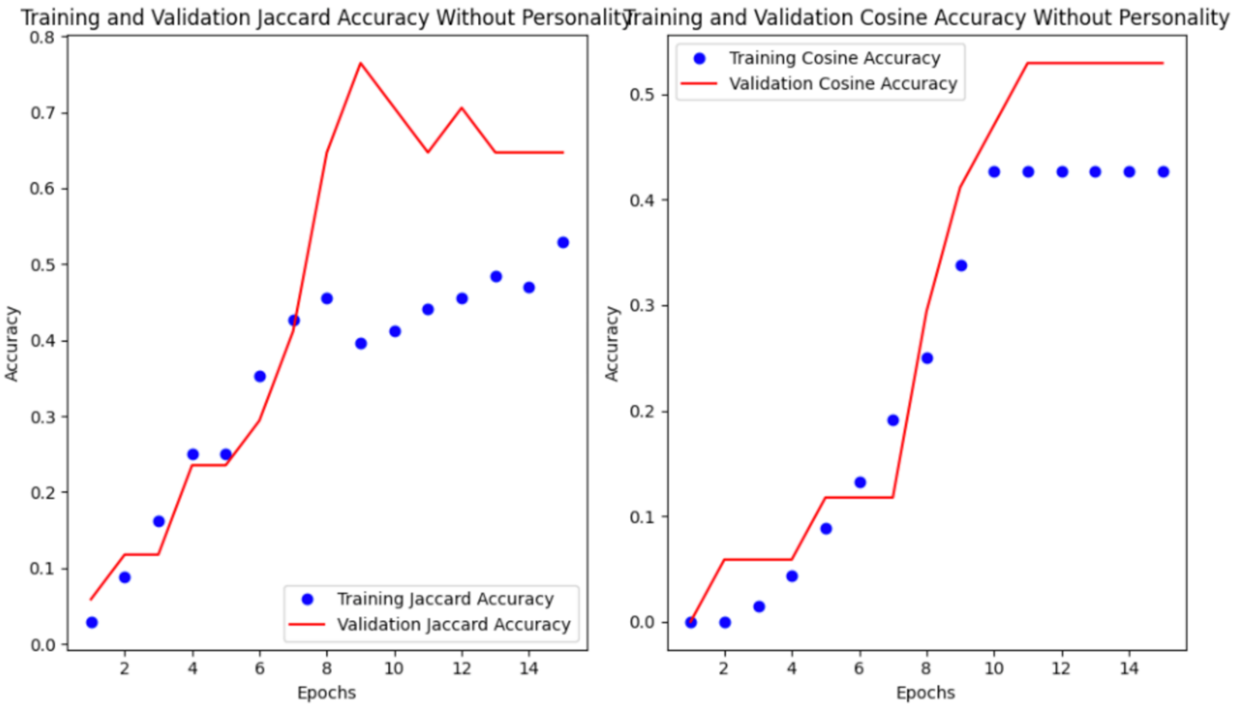


Figure 4 - Accuracy graph of Jaccard and Cosine Similarity Model without Personality

In Figure 4, which serves as a visual representation of the machine learning model's performance, two lines are plotted to represent different aspects of the model's accuracy. The blue line depicts the model's accuracy in relation to the training data, reflecting the proportion of mentor-mentee pairings that were correctly classified as successful or unsuccessful by the model during its training phase. This line is indicative of the model's capacity to discern and assimilate patterns and associations present in the training dataset, where the dataset comprises the skills and goals of mentees in alignment with the skills of mentors.

The model's accuracy on the training data is a pivotal indicator of its competence in interpreting and operationalizing the correlations learned from the provided input. A high training accuracy implies that the model has proficiently captured the intrinsic structure of the dataset and can consistently predict outcomes based on the features it has been exposed to. Nevertheless, vigilance against overfitting is warranted, as it represents a condition in which the model is excessively specialized to the training data, ultimately compromising its ability to generalize.

Conversely, the red line delineates the model's accuracy on a validation dataset, which comprises data points that were withheld during the training process and thus acts as a surrogate for novel, unseen data. The red line offers a gauge of the model's capability to extrapolate its learned knowledge to disparate data, which is reflective of its generalization aptitude. The validation accuracy is a critical barometer for forecasting the model's prospective real-world performance, where it will encounter data beyond what was explicitly presented during training.

A model that achieves high accuracy on both training and validation datasets is regarded as possessing robust generalization qualities. Such a model can accurately categorize new instances that were not incorporated into its training data. In contrast, a model exhibiting high training accuracy but diminished validation accuracy may be indicative of overfitting, signifying that the model may be memorizing the training dataset rather than truly comprehending the underlying patterns.

The interplay between the blue and red lines in Figure 4 is integral to evaluating the machine learning model's overall effectiveness. Ideally, both lines should closely align, demonstrating the model's consistent performance across training and validation datasets. The convergence of these lines at a high accuracy level would imply that the model is not only assimilating effectively but is also applying that knowledge to accurately predict outcomes in diverse contexts, thereby serving as a valuable instrument for identifying successful mentor-mentee pairings.

The divergence between the red line, representing the accuracy on the validation dataset, and the blue line, signifying the accuracy on the training dataset, is a salient sign of overfitting within the machine learning model's performance. Overfitting arises when a model internalizes the training data to such an extent that it includes the incidental noise and specificities, thereby undermining its capacity to generalize to novel data. The model becomes overly attuned to the training dataset, and its predictive effectiveness wanes when applied to previously unencountered data.

As the model undergoes additional epochs, which are iterative cycles over the training data that allow for the refinement of its weights and optimization of performance, the discrepancy between the red and blue lines is often observed to widen. This increasing gap, alongside the scatter of blue dots that might signify separate instances of accuracy measurements throughout the training, indicates that the model is progressively aligning itself more closely with the nuances of the training data instead of absorbing the more generalizable patterns pertinent to the validation dataset.

Furthermore, when evaluating the performance of two distinct models—one leveraging Jaccard similarity and the other employing cosine similarity—the empirical evidence points to a marginally enhanced validation performance by the model predicated on Jaccard similarity metrics. This observation suggests that Jaccard similarity may be more adept at elucidating the key attributes that lead to effective mentor-mentee pairings within this specific dataset and the domain of the problem. The more favorable validation performance of the Jaccard-based model intimates that it may be delivering a more pertinent measure of similarity between the characteristics of mentors and mentees, thus yielding improved generalization for predicting new pairings.

The comparative efficacy of the Jaccard model relative to the cosine model is significant for discerning which similarity metric is more suitable for the task. Each metric provides a distinct lens through which to view the data, with Jaccard focusing on the intersection relative to the union of sets, and cosine assessing the orientation between vectors in multidimensional space. The superior validation performance of one metric over the other can guide future decisions on model architecture and feature engineering.

To rectify overfitting, it is essential to employ strategies that foster a model's generalizability despite high training accuracy. Methods such as cross-validation, regularization, and pruning are instrumental in mitigating overfitting. By refining the model to diminish overfitting, researchers can bolster the reliability of the model's forecasts and affirm its utility as an instrument for facilitating successful mentor-mentee pairings in practical settings.

During the assessment phase of the investigation, the Jaccard Similarity model and the Cosine Similarity model underwent comprehensive testing to evaluate their predictive capabilities. This phase entailed utilizing each model to prognosticate outcomes on a test dataset that had not been introduced during the training stage. The test labels constituted the benchmark, or 'ground truth,' which the models' predictions were measured against to ascertain accuracy.

The Jaccard Similarity model demonstrated a mean accuracy of 75%, indicating that it accurately predicted successful mentor-mentee pairings with this frequency on average. In contrast, the Cosine Similarity model exhibited a marginally lower mean accuracy, recorded at 73%. This difference implies that the Jaccard Similarity model has a superior ability to predict test labels accurately, suggesting a more effective congruence of its similarity measurements with the determinants of successful mentor-mentee matches within the confines of the dataset used.

An insightful finding from the evaluation was the Jaccard Similarity model's relatively minor deviation in accuracy between its training and prediction phases. This characteristic is particularly noteworthy as it suggests the model's performance remains relatively consistent when transitioning from known (training) data to novel (testing) data. Such consistency is a hallmark of a model with reduced susceptibility to overfitting and enhanced capacity for generalization across diverse datasets.

The Jaccard Similarity model's more impressive performance in the first iteration of the research, marked by its higher average accuracy and sustained predictive quality, highlights its potential appropriateness for the mentorship program's dataset. It implies that the Jaccard Similarity metric's approach, which evaluates the shared attributes between mentee goals and skills in comparison to mentor skills, is particularly conducive for this specific application.

These insights are invaluable for informing the future development cycles of the predictive model. They provide evidence that the Jaccard Similarity model might serve as a more dependable predictor of successful mentorship pairings and that its continued refinement could yield even more precise and steadfast forecasts. Moreover, the performance metrics gleaned from this preliminary evaluation establish a baseline for measuring enhancements and modifications in future iterations of the model.

5.2 Iteration 2 – Hybrid Model with Personality Attribute

In the subsequent iteration of the investigation, the research cadre aimed to augment the predictive efficacy of the Convolutional Neural Network (CNN) model through the incorporation of an added dimension of data—namely, the personality types of mentors and mentees. This element was integrated into the CNN's input layer to provide a more encompassing perspective of the mentor-mentee dyad, factoring in not only their competencies and ambitions but also their psychological and behavioral predispositions.

The rationale for including personality types rests on the understanding that compatibility in mentorship transcends the mere intersection of professional skills and shared objectives. It also entails the interpersonal dynamics, communicative propensities, and rapport-building capabilities, all of which are potentially influenced by individual personality traits. The hypothesis was that with personality types considered, the model would discern underlying patterns and connections that might remain obscured when exclusively focusing on professional attributes.

Additionally, the personality data was integrated into both Jaccard Similarity and Cosine Similarity calculations, facilitating a more layered comparison of mentors and mentees. This approach expanded the evaluation to include not only technical and aspirational congruence but also character compatibility. The Jaccard Similarity continued to measure the shared traits relative to the collective set of traits, while the Cosine Similarity determined the orientation of the personality trait vectors in a multidimensional context.

The study's Figure 5 portrays the accuracy graph of the refined CNN model, illustrating the influence of personality type inclusion on the accuracy of mentor-mentee pairings using the two similarity metrics. The graph is likely to reveal distinct lines or bar charts representing the model's performance employing Jaccard Similarity and Cosine Similarity, both with and without the introduction of personality types.

The anticipation surrounding the integration of personality type data into the model's input layer and similarity assessments is an improvement in accuracy, predicated on the model's enhanced sensitivity to the multifaceted nature of human interactions within mentorship. An elevation in accuracy depicted in the graph would imply that personality types are indeed a substantial variable in forecasting the success of mentorship pairings.

The model's performance, when enhanced by these supplementary factors, could unveil significant revelations regarding the role of personality congruence in mentorship. For example, a marked increase in model accuracy following the introduction of personality types would emphasize the criticality of accounting for individual differences when shaping effective mentor-mentee connections. Such outcomes would not only affirm the benefit of incorporating a broader spectrum of data into the model but could also influence the structuring of mentorship programs, highlighting the significance of personality matching as well as the alignment of skills and objectives.

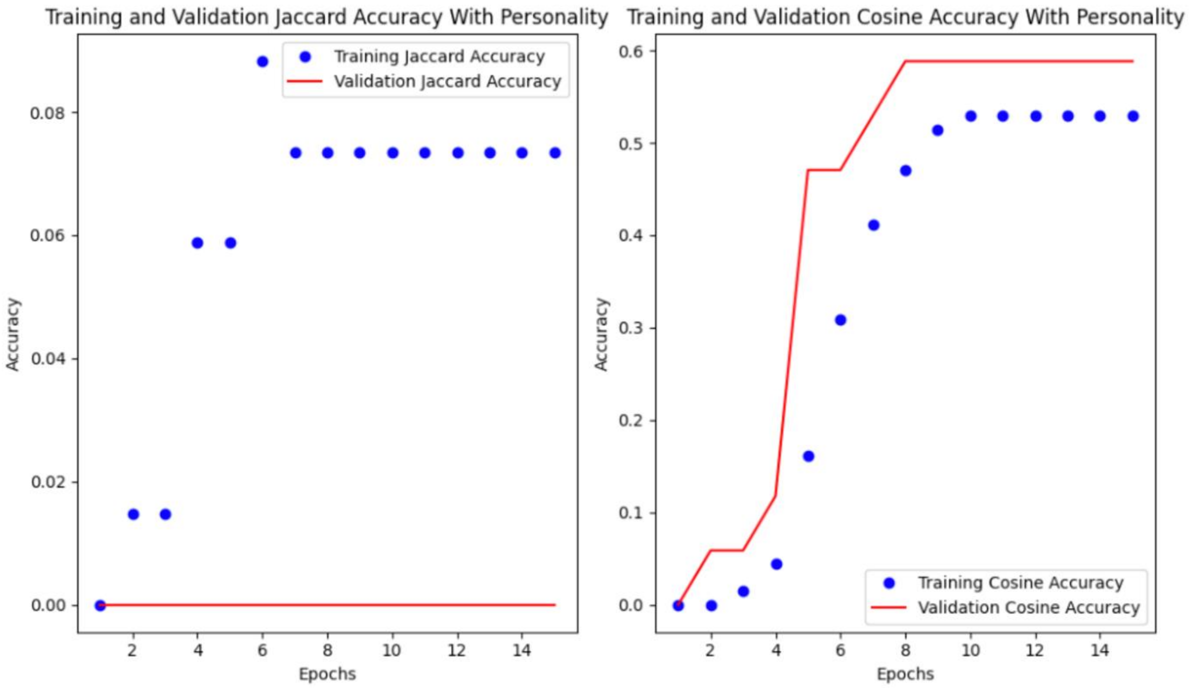


Figure 5 - Accuracy graph of Jaccard and Cosine Similarity Model with Personality

As the study progressed into its second iteration, the incorporation of personality dimensions into the similarity metrics led to an unexpected shift in the model's performance. Contrary to initial predictions, the inclusion of personality traits within the Jaccard similarity calculations resulted in a decline in effectiveness, producing less favorable outcomes compared to those derived without the personality data. This could suggest complexities in effectively integrating personality information into a set-based similarity measure like Jaccard or that personality traits might not be as indicative of mentorship success as initially thought.

Conversely, the Cosine similarity model, which now considered personality traits in addition to skills and goals, exhibited improved performance relative to its first iteration. This enhancement indicates that the underlying vector space model of the Cosine similarity metric may be better suited to encapsulate and leverage the subtleties of personality data. Given that Cosine similarity evaluates the cosine of the angle between two vectors, it is plausible that it more aptly accommodates the multi-dimensional aspects of personality data, thus providing a more nuanced gauge of mentor-mentee compatibility.

The distinctions in performance are likely illustrated in Figure 5, which would show the accuracy graph of the updated models. Although the gap between the validation accuracy (red line) and the training accuracy (blue dots) persisted, it was less extensive than in the first iteration, hinting at reduced overfitting and better generalization to the validation dataset.

Furthermore, the more consistent correlation between the red line and the blue dots suggests that the model's training accuracy is becoming a more dependable indicator of its performance on unseen data. If the validation accuracy closely mirrors the training accuracy, it implies that the model has internalized patterns that are generalizable and hold true across diverse data subsets.

In essence, the second iteration's investigation into the role of personality traits in mentor-mentee similarity assessments yielded mixed results. The Jaccard similarity model did not benefit from the inclusion of personality data, but the Cosine similarity model's predictive accuracy was positively influenced. The diminished discrepancy between training and validation accuracy, along with the validation accuracy's congruence with the training trend, indicates progress towards a model that more accurately and reliably assesses mentor-mentee compatibility. These outcomes may prompt additional research into the most effective methods of integrating personality data into models for mentorship matching and could inform the enhancement of mentorship programs.

During the validation stage of the second iteration, the predictive performance of the enhanced mentor-mentee matching models was scrutinized using a set of test labels. This involved applying the Jaccard Similarity model and the Cosine Similarity model to forecast outcomes on a test dataset not previously encountered during the training or validation phases. These test labels act as a metric for assessing the models' genuine predictive power in scenarios analogous to real-world applications, where the model interacts with new data.

The predictive results disclosed a stark contrast in the performance of the two models. The Jaccard Similarity model yielded a mean accuracy of 66%, a decrease from the 75% accuracy rate achieved in the first iteration. This downturn in performance suggests that the integration of personality data within the Jaccard similarity calculations may not have been as beneficial as expected, potentially due to the Jaccard Similarity metric's unsuitability for complex personality traits in conjunction with skills and goals, or an indication that personality compatibility's impact on mentorship success is not as linear as assumed by the model.

On the flip side, the Cosine Similarity model's mean accuracy improved significantly to 81%, up from 73% in the first iteration. This uptick in performance implies that the Cosine Similarity model was more adept at using the additional personality data to predict successful pairings. The enhancement of the Cosine Similarity model's accuracy suggests that its underlying vector space model of similarity is better equipped to handle the added dimensionality of personality attributes. It also points to the potential significance of personality data in forming a more rounded understanding of successful mentor-mentee pairings.

The varying performance between the two iterations, especially the advancement in the Cosine Similarity model, underscores the importance of the chosen similarity calculation method when new data dimensions are incorporated. The Cosine Similarity metric, which evaluates the angle between vectors in a multidimensional space, may be more versatile in adjusting to the inclusion of diverse attributes that surpass conventional skills and goals. This could mean that the Cosine Similarity measure is more robust in integrating complex data, which is especially relevant in the intricate area of human relationships and mentorship matching.

These outcomes provide critical insights into the selection of suitable similarity metrics for models that match mentorship pairs. They highlight the necessity of choosing a metric that is not only congruent with the data types being compared but is also capable of embracing additional complexity. The improved accuracy of the Cosine Similarity model, when enriched with personality data, points towards a promising pathway for future research iterations and the potential development of more sophisticated models that can accurately determine mentor-mentee compatibility across an expanded range of attributes.

Chapter 6

Discussion

The research aims to optimize the process of creating optimal mentor-mentee pairings by evaluating a broad spectrum of factors encompassing skills, career objectives, and personality attributes. To this end, the study introduces a cutting-edge hybrid machine learning model that synergizes the methodologies of collaborative filtering and content-based filtering. Collaborative filtering is a commonly implemented technique in recommendation systems that forecasts user preferences by aggregating choices from numerous users, while content-based filtering generates recommendations through a comparison between the content of items and a user profile.

The assessment of skills and professional goals is of paramount importance, as it empowers mentors to customize their mentorship to address the distinct developmental requisites and ambitions of mentees. Concurrently, the comprehension of personality traits is crucial in the pairing process, influencing the rapport between individuals and the efficiency of their communication. The amalgamation of these components is essential for nurturing a mentorship dynamic that is both reinforcing and conducive to achievement.

In preparation for the dataset analysis, the research employed sophisticated Natural Language Processing (NLP) methods for preprocessing. These NLP techniques facilitate the distillation and refinement of pertinent information from textual data, recasting it into a form amenable to machine learning algorithms. This preprocessing phase is critical to guarantee that the data is devoid of impurities, uniform, and organized in a manner that is favorable to analysis.

A portion of the preprocessed data was then utilized to compute Jaccard Similarity and Cosine Similarity scores, which quantitatively gauge the congruence between the skills and objectives of mentees with the competencies of mentors, as well as the harmony of their personality profiles. These similarity scores were subsequently adopted as labels in the hybrid machine learning model, directing its learning trajectory in predicting successful pairings.

The research executed two separate iterations of the model to elucidate the effect of personality types on the pairing mechanism. While the initial iteration omitted personality type data, the second iteration included it, enabling a comparative analysis to pinpoint the effect of incorporating personality traits on the model's efficacy.

Findings from both iterations elucidated that the integration of personality type information markedly bolstered the model's predictive capabilities. The hybrid model utilizing Cosine Similarity as the label, especially with the addition of personality types, was found to exhibit higher accuracy. In contrast, the Jaccard Similarity-based model did not experience the same benefit from the personality data inclusion and, in fact, saw a diminished performance.

Ultimately, the study culminates in an enhanced hybrid machine learning model which, by integrating personality type data and leveraging Cosine Similarity as the label, attains an impressive mean accuracy of 81%. This elevated accuracy rate validates the model's effectiveness and emphasizes the significance of an all-encompassing approach to pairing mentors with mentees—one that considers not only professional skills and goals but also the influential aspect of personality traits.

The Jaccard similarity metric, which is predicated on the intersection over the union of two sets, exhibits a heightened sensitivity to the size of those sets. As noted by Vorreuther & Warin (2021), minor alterations in the composition of either set can precipitate considerable variations in the resultant similarity score. This characteristic can be particularly challenging when comparing sets of divergent sizes, where the inclusion or exclusion of a few elements might disproportionately sway the outcome. Within the realm of mentorship pairing, where sets might symbolize skills or personality traits, this implies that trivial differences in the attributes of mentors or mentees could engender marked discrepancies in perceived compatibility, potentially skewing the results.

Conversely, cosine similarity assesses the cosine of the angle between two vectors in a multidimensional space and is thus unaffected by the magnitude of the vectors. This attribute renders it capable of effectively managing vectors of unequal lengths and establishes it as a robust option for comparing attributes that can be quantified numerically. The distinction between Jaccard and cosine similarity became particularly salient when personality data was integrated into the similarity assessments. Given that personality trait data can be inherently dimensional and possibly continuous, vector representations are well-suited, which elucidates the enhanced performance observed with cosine similarity.

Moreover, cosine similarity is especially apt for data that can be articulated as vectors, such as text data transformed into TF-IDF (Term Frequency-Inverse Document Frequency) representations. TF-IDF is a statistical measure that determines the significance of a word to a document in a corpus, accounting for the word's document frequency and its inverse corpus frequency. Such data is intrinsically multivariate and often real-valued, aligning well with the strengths of cosine similarity.

The research by Wangwiwattana & Tongvivat (2022) corroborates the observations of this study. Their examination confirms that cosine similarity is proficient at handling multivariate and real-valued data, positing it as a fitting metric for contexts where such attributes are common. In the context of mentor-mentee matching, this pertains to datasets where skills and personality traits are recorded in a manner amenable to vectorization, enabling refined comparisons.

The technical conclusions drawn from this study are buttressed by the collective insights from existing scholarly works, which consistently advocate for the efficacy of cosine similarity in the analysis of vectorizable data types where the magnitude of vectors does not constitute a chief concern. This concurrence with prior research fortifies the choice to employ cosine similarity as the metric of choice within the hybrid machine learning model developed in this research. The congruence with earlier findings intimates that cosine similarity is poised to continue delivering superior outcomes in analogous applications, notably those involving intricate, multidimensional datasets such as those utilized for evaluating mentor-mentee compatibility.

The research by Örtenblad et al. (2017) posits a skepticism around the connection between personality types and their effect on student learning experiences, focusing narrowly on the direct relationship between these two factors. This approach precludes the intricate interplay of additional elements such as course content, pedagogical approaches, and the level of student engagement, which are all pivotal in shaping the educational environment. The limitation of this scope might obscure the potential nuances and synergies that exist within the broader educational context.

In contrast, the current study expands upon this premise by incorporating a more holistic methodology. This approach not only acknowledges the importance of personality types but also integrates them with other influential factors, such as the skills and ambitions of both mentors and mentees. It is predicated on the understanding that mentorship, much like education, is an inherently complex interaction influenced by a multitude of factors that extend beyond the bounds of personality traits alone.

Through the inclusion of multiple dimensions of data, this study aims to develop a more sophisticated and accurate model for predicting successful mentor-mentee pairings. The empirical evidence from the data analysis substantiates this multi-faceted approach. The integration of personality types as a standalone variable yielded a model with an accuracy of 73%. However, the accuracy significantly escalated to 81% when the model considered the interplay of personality types with skills and goals.

This notable increase in model accuracy reinforces the concept that personality should not be examined in isolation when evaluating potential mentor-mentee relationships. Instead, the convergence of professional competencies and objectives with personal traits offers a more robust framework for understanding and predicting the dynamics of successful mentorship.

The findings of this study challenge the conclusions drawn by Örtenblad et al., suggesting that the impact of personality types on learning outcomes may indeed be significant when analyzed in conjunction with other relevant factors. This comprehensive approach aligns with the multifaceted reality of mentorship interactions, where success is often predicated on a harmonious blend of professional alignment and interpersonal compatibility.

Consequently, this research contributes to the discourse by highlighting the need for inclusivity of diverse data points in predictive modeling within educational and professional development contexts. By transcending a unidimensional perspective, it paves the way for more nuanced analyses and interventions that can enhance the efficacy of mentorship programs and, by extension, the learning and development experiences of individuals.

Understanding personality traits is essential for personal and professional growth, as it allows individuals to introspect on their intrinsic predispositions that shape their information processing, decision-making, and interactions. Two prominent personality dimensions often considered are Intuition and Sensing, which form a cognitive style continuum that influences how people perceive and comprehend their environments.

Intuitively inclined individuals tend to trust their instincts, look past immediate data to future possibilities, abstract ideas, and the overarching view. They are naturally innovative and creative, favoring the exploration of new concepts. Intuitive types are imaginative and have a visionary mindset, skilled at discerning patterns and often flourish in environments that value brainstorming and the development of novel solutions or strategies.

Conversely, individuals with a Sensing preference are anchored in reality's concrete and tangible elements. They prioritize empirical data and observable facts, concentrating on the here-and-now and what is accessible through their senses. Sensing types are detail-focused, pragmatic, and systematic, usually excelling in tasks requiring acute attention to detail or hands-on experience.

The significance of recognizing Intuition and Sensing personality dimensions was underscored in Hemdan et al.'s (2022) study, which explored the influence of personality on creativity. By identifying students' personality traits, educators can customize their teaching to accommodate diverse learning styles. The study likely discovered that environments aligned with students' personality traits could bolster their creative capacities, enhancing their overall educational experience.

Similarly, Brownfield (1993) investigated the Sensing/Intuition scale's importance, which evaluates individuals' preferred information absorption mode. This scale is pivotal as it influences how people learn, assimilate information, and utilize their knowledge. By discerning an individual's propensity towards Sensing or Intuition, one can predict their favored learning methods—whether they lean towards concrete, experiential learning or abstract, theoretical thinking.

In this study's context, factoring in personality traits into the mentor-mentee matching model is particularly crucial. By accounting for mentors' and mentees' Intuition and Sensing preferences, the model can more accurately forecast a successful partnership's likelihood. Aligning a mentee with a mentor who shares or complements their information processing and decision-making style can result in a more efficacious and rewarding mentorship. It allows mentors to offer advice that harmonizes with the mentee's cognitive preferences, fostering a relationship conducive to growth and productivity.

This intricate comprehension of personality traits and their practical ramifications emphasizes the importance of integrating such traits into educational and mentorship programs. Utilizing personality psychology insights, these programs can be more precisely tailored to meet the distinct needs of their participants, leading to improved personal and professional development outcomes.

In mentorship relationships, the congruence of personality traits between mentors and mentees, particularly how they perceive and process information, can significantly influence the effectiveness of their interactions. When individuals have similar personality traits in areas such as information perception and processing, they often share compatible communication styles and learning and problem-solving approaches. This synergy can cultivate a mentorship dynamic that is both productive and rewarding for both parties.

For example, mentors and mentees with a pronounced intuitive tendency may naturally connect over their joint propensity for abstract thought and conceptualizing. These pairs might thrive in dialogues about innovation, future possibilities, and theoretical constructs, sharing an enthusiasm for creativity and a vision that looks beyond immediate, tangible data. Pairs like these may excel in environments that promote exploration, experimentation, and the development of new ways to tackle challenges.

Conversely, mentor-mentee pairs with a strong sensing preference may benefit most from a mentorship where both participants favor a practical, detail-focused approach. Such pairs might appreciate mentorship activities that involve tangible data, direct observation, and concrete facts. They may prefer pragmatic, hands-on problem-solving and activities that involve engaging directly with their environment. This mutual emphasis on the present and the practical application of knowledge can form a solid, pragmatic bond between mentor and mentee.

Intuitive-Intuitive or Sensing-Sensing pairings can be beneficial because they enable an effortless exchange of ideas and strategies that both individuals inherently understand and value. This reduces the likelihood of miscommunication and misaligned expectations, as both mentor and mentee share a similar cognitive approach. Such cognitive congruity can lead to a more harmonized learning experience, where the mentor can intuitively respond to the mentee's needs and preferences with guidance that profoundly aligns with the mentee's natural thought processes.

However, it's also crucial to recognize that while matching personality traits can simplify some mentorship aspects, diversity in personality types can introduce complementary strengths to a pairing. An Intuitive mentor, for instance, might encourage a Sensing mentee to consider broader implications of their work, whereas a Sensing mentor could help an Intuitive mentee maintain focus on concrete results. The success lies in appreciating and harnessing these differences beneficially.

Ultimately, the goal of considering personality traits in mentorship pairings is not to find identical personalities but to forge connections that enrich the experience and growth of both participants. Whether through shared traits or through the balance of differing qualities, the objective is to foster a relationship conducive to the mentee's development and the mentor's fulfillment in their guiding role. By acknowledging the intricate ways individuals perceive and absorb information, mentorship programs can be crafted to optimize the potential of each pairing.

The preliminary examination of mentor-mentee matching data has uncovered a distinct pattern: a considerable proportion of the pairings, around 60%, featured intuitive mentees matched with intuitive mentors. This pattern is reflected in the pairing of sensing individuals, indicating a pronounced preference for pairing individuals with similar methods of information processing.

This predilection for homogeneous pairings aligns with insights from the literature review, underscoring the significance of matching teaching and learning styles in educational and developmental contexts. The literature posits that when mentors and mentees have congruent learning and problem-solving approaches, knowledge and skill transfer can be more effective, potentially leading to a more harmonious and fruitful relationship.

The observed correlation between matching personality traits and mentor-mentee compatibility suggests an innate tendency for individuals to connect with those who have a similar information processing style. For example, intuitive mentors and mentees, sharing a common ground in abstract thinking and conceptual dialogue, are likely to find it easier to exchange ideas, establish objectives, and appreciate each other's viewpoints. In the same vein, sensing pairs, with a mutual emphasis on practicality and detail, may find knowledge exchange more direct and grounded in tangible experiences and facts.

Incorporating this personality dimension into the machine learning model capitalizes on these observed tendencies to refine the model's predictive capacity for successful mentor-mentee pairings. By including personality traits, the model goes beyond matching based on skills and career aspirations, also considering the participants' engagement and interpretation of their environment. This enriched data layer deepens the model's understanding of mentor-mentee relationship dynamics, facilitating more sophisticated predictions.

The outcome of this data integration has been a notable enhancement in the model's predictive accuracy. The improved model performance validates the literature's focus on the congruity of teaching and learning styles. By integrating the personality dimension, the model more accurately captures the intricate array of variables that bolster successful mentorships. The data supports the concept that effective mentorship hinges on not just the content of what is taught but also the manner in which it is communicated and perceived.

These insights have profound implications for structuring mentorship programs. They suggest that a data-informed approach, acknowledging the complex nature of human interactions, can lead to more efficacious mentor-mentee pairings and, consequently, more successful mentorship endeavors. By recognizing the importance of aligning mentors and mentees based on personality traits, organizations can cultivate deeper and more impactful learning experiences.

The challenge of limited data is indeed prevalent in machine learning, particularly when dealing with the intricate and multifaceted nature of human relationships, such as mentor-mentee dynamics. Developing a model with strong predictive capabilities and generalizability requires a dataset that captures the breadth and depth of the interactions and patterns inherent in these relationships.

In the context of mentor-mentee matching, a larger number of pairings would enrich the dataset, enabling the model to discern a broader range of patterns and subtleties within the pairing process. The complexity of matching mentors with mentees goes beyond mere skill alignment; it involves a careful examination of CVs, skill sets, and other pertinent criteria to determine compatibility, including personality fit and learning styles. These elements are crucial for fostering a conducive atmosphere for professional growth and development.

The pairings crafted by the organization's leadership, which are considered to be optimal, act as a benchmark for assessing the machine learning model's performance. If the model's predictions are consistent with the leadership's selections, it can be deemed effective. However, a limited dataset can hinder the model's learning, potentially causing overfitting to the few examples it has been exposed to or failing to generalize to new, unseen pairings.

Expanding the dataset would grant the model access to a more diverse array of mentor and mentee profiles, encompassing a wider variety of skills, goals, and personality types. With more data, the model can learn more robust patterns, enhancing its adaptability to different pairing scenarios. Additionally, a larger dataset would facilitate more rigorous validation methods, like cross-validation, reinforcing the model's generalizability and dependability.

In conclusion, augmenting the dataset is essential for improving the machine learning model used for mentor-mentee matching. With a more comprehensive dataset, the model can more accurately replicate the nuanced human-driven pairing process, offer more reliable predictions, and become an invaluable asset for organizations seeking to create impactful professional development experiences.

Implementing a systematic feedback mechanism is a strategic approach to enriching the understanding of mentor-mentee relationship dynamics within an organization. By administering a survey to both mentors and mentees after a set period of participation in the mentoring program, the enterprise can collect valuable qualitative data on the participants' satisfaction with their pairings and the perceived effectiveness of the relationship.

The survey should be thoughtfully constructed to draw out rich, detailed feedback covering the mentorship experience's multifaceted nature. Questions might explore the alignment of communication styles, the relevance and helpfulness of the guidance provided, progress toward goals, and the personal and professional development of the mentee. Additionally, it should inquire about the quality of the rapport established, any challenges faced, and suggestions for program improvement.

The qualitative data amassed from these surveys would provide the enterprise with in-depth insights. By correlating the subjective experiences shared by mentors and mentees with the predictions of the machine learning model, patterns and mismatches can be identified. Should the model's predicted successful pairings repeatedly result in less than satisfactory survey feedback, it would highlight areas where the model might be refined.

On the other hand, a consistent correlation between the model's predictions and high satisfaction levels reported in the surveys would affirm the model's effectiveness. Nonetheless, a thorough analysis of the feedback could reveal additional subtleties and elements that the model currently does not consider. These findings could then be used to enhance the model's training data, thereby improving its predictive accuracy.

Incorporating feedback from surveys into the machine learning algorithm allows the enterprise to adopt a continuous improvement approach, where the model is iteratively refined. This not only enhances the model's precision but also ensures that it stays attuned to the participants' changing needs and expectations in the mentorship program.

Beyond optimizing the model, the feedback surveys also offer critical insights to the leadership team overseeing mentor-mentee pairings. Understanding the subjective experiences of participants can help identify successful strategies and common stumbling blocks, guiding the organization to make more informed decisions and adjustments to the mentoring program's design and execution.

Integrating a feedback survey into the mentor-mentee matching process yields twofold advantages: it acts as a tool for both validating and enhancing the machine learning model and provides the organization with actionable data to elevate the efficacy and satisfaction of its mentoring program.

The proposition of creating a recommendation system grounded on the predictive model holds significant potential to elevate the mentorship program. This system would act as an intermediary that streamlines the matching process, capitalizing on the model's capacity to sift through intricate data and pinpoint appropriate mentor-mentee pairs.

For this recommendation system to function effectively, mentees would be required to submit comprehensive personal information, encompassing professional expertise, career objectives, and personality nuances. A standardized inventory or questionnaire might be employed to ascertain personality traits in a uniform and measurable way.

Once the requisite information is acquired, the model would scrutinize the data to discern feasible matches. It would generate a hierarchy of mentors sorted by compatibility scores, based on the model's calculated output percentages. The top-ranking mentors, such as the top three, would be those whose compatibility percentages align most closely with the mentee's profile and the learned patterns from the model.

The deployment of such a recommendation system would not only hasten the pairing procedure but also involve mentees more directly by offering them a selection of potential mentors, thereby fostering a sense of agency in their mentor selection. Moreover, it would alleviate the organizational burden of manually pairing mentors and mentees.

To continuously refine the system and preserve the model's accuracy and pertinence, the recommendation system could be integrated into the mentorship program's regular operations. Periodic participant surveys could be conducted to collect feedback on the effectiveness of the pairings, focusing on relationship quality, progress toward goals, the mentorship's perceived value, and areas for program enhancement.

This feedback would act as ongoing validation for the model. Positive survey results would suggest that the model is effectively identifying suitable matches that meet participants' expectations and contribute to their professional growth.

Should the feedback reveal persistent issues or dissatisfaction, it would highlight aspects of the model that may need improvement. These findings could be channeled back into the system in a feedback loop, enabling model adjustments and improvements based on actual outcomes and user experiences.

By adopting such a system, the enterprise would not only validate the model's effectiveness through real-world application but also cultivate a culture of continuous enhancement. The recommendation system, driven by the predictive model and enriched through participant feedback, would progressively evolve, ultimately becoming a more accurate and invaluable asset for fostering successful mentor-mentee pairings within the mentorship program.

6.1 Broader Implications and Applications

Human Resources and Recruitment: The hybrid machine learning model for mentor-mentee pairing can significantly enhance human resources and recruitment processes. In employee onboarding, similar algorithms can match new hires with suitable mentors or buddies, ensuring a smoother integration into the company culture and accelerating their productivity. Furthermore, this model can be instrumental in forming project teams by aligning complementary skills, career aspirations, and personality compatibility. This approach not only fosters more cohesive and effective teams but also improves job satisfaction and employee retention.

Educational Settings: In educational environments, the algorithm can revolutionize student-teacher and peer tutoring pairings. By matching students with advisors or mentors who align with their academic goals, learning styles, and personality traits, the model can optimize educational guidance and support. Additionally, peer-to-peer tutoring can be enhanced by pairing students needing assistance in certain subjects with those who excel in them and have compatible personalities. This personalized approach can improve academic performance and foster a supportive learning community.

Healthcare: The healthcare sector can also benefit from this advanced pairing model. In personalized care or mental health support, the algorithm can match patients with caregivers or therapists who possess the relevant expertise and a compatible approach to care, thereby enhancing patient outcomes. Additionally, forming support groups based on similar experiences and personality compatibility can significantly improve the effectiveness of group therapy or support networks, providing patients with a more empathetic and understanding environment for their recovery and growth.

Chapter 7

Conclusion

The research project focuses on crafting a state-of-the-art hybrid machine learning model that combines the strengths of collaborative filtering and content-based filtering approaches. Collaborative filtering predicts preferences based on user similarities, while content-based filtering suggests items by comparing item features with a user's past preferences.

The goal of the project is to discern optimal mentor-mentee matches by meticulously analyzing the skills, career goals, and personality traits of participants in the mentorship program. The model is designed to foster mentorship pairings that are congruent professionally and compatible on a personal and interpersonal level, enhancing the likelihood of a successful and enriching mentorship experience.

Skills and goals analysis is pivotal, providing a snapshot of where mentors and mentees are in their professional trajectories. This analysis allows mentors to tailor their guidance to the mentees' developmental needs, ensuring that their expertise is directly beneficial to the mentees' professional growth. This skill-goal alignment is essential to establish a focused and effective learning atmosphere within the mentorship program.

Incorporating personality traits into the model introduces an additional layer to the matching process by considering the mentees' learning preferences, which can vary widely based on their personality profiles. A mentee who flourishes with hands-on learning might be best paired with a mentor who can provide such experiences. Conversely, mentors' teaching styles should also be considered, as they can significantly impact the mentorship's effectiveness.

Evaluating personality traits also helps measure the potential for communicative harmony between mentors and mentees. Communication is fundamental to mentorship success, and the capacity to relate and understand each other personally can lead to more profound interactions and a solid mentorship bond. Factoring in personality compatibility allows the model to predict pairings that are more likely to engage in meaningful dialogue, understand each other, and build a strong mentorship foundation.

In conclusion, the hybrid machine learning model developed through this research serves as a multifaceted instrument to optimize mentor-mentee pairings, carefully weighing skills, goals, and personality traits. By adopting an all-encompassing approach that recognizes the importance of both professional and personal compatibility, the model aims to significantly improve the mentorship process and outcomes for participants.

The implementation of the collaborative filtering component of the hybrid machine learning model utilized a Convolutional Neural Network (CNN) architecture, which is commonly recognized for its prowess in image processing. However, CNNs have also demonstrated their capability in recommendation systems by identifying patterns and features within complex datasets. In this model, the CNN was trained using a dataset provided by an IT firm based in New Zealand, consisting of detailed records of past mentor-mentee pairings and the outcomes of these relationships.

Complementing the collaborative filtering, content-based filtering was incorporated through two distinct input layers in the model. The first input layer focused on aligning the skills and goals of mentors and mentees, examining professional capabilities, experiences, and career aspirations to predict successful mentorship matches.

The second input layer concentrated on the compatibility of personality types between mentors and mentees. This was achieved by utilizing the Myers-Briggs Type Indicator (MBTI) personality assessment, a prevalent psychological tool that classifies individuals into sixteen personality types across four dichotomies: Introversion/Extraversion, Sensing/Intuition, Thinking/Feeling, and Judging/Perceiving. The inclusion of MBTI results allowed the model to pair mentors and mentees based on compatibility in their work styles, communication methods, and interpersonal interactions.

Despite the limitations presented by the dataset's size, the model achieved an accuracy rate of 78%. This is significant, as the scope and diversity of data are critical factors in a model's ability to learn and make generalizations. A limited dataset might not represent the full range of mentor-mentee relationship possibilities, potentially constraining the model's predictive capabilities. Nevertheless, the respectable accuracy rate achieved indicates that the model has effectively learned key patterns and features from the given dataset.

The efficacy of this hybrid machine learning model is rooted in its comprehensive approach. By merging collaborative and content-based filtering, the model leverages historical pairing patterns while matching individual profile attributes. Consequently, it has proven to be a valuable tool for evaluating and predicting mentor-mentee pairings, with the prospect of further enhancements as additional data is gathered. This research underscores the potential of hybrid models in personalized recommendation systems, particularly in the realm of professional mentorship programs.

The academic exploration of mentorship is extensive, detailing the critical components that drive the success of mentor-mentee relationships. Scholarly research underscores the profound benefits of well-matched mentorship, such as enriched learning experiences, accelerated professional growth, and heightened job satisfaction. Several studies within this rich body of literature recommend methods to enhance mentorship programs, often pointing to the importance of compatibility in fostering more effective mentor-mentee dynamics.

Simultaneously, the machine learning domain has seen substantial advancements, particularly in recommendation systems capable of predicting user preferences in various areas, including media, e-commerce, and social networking. Within mentorship, machine learning research has shown that algorithms can streamline the pairing process by identifying potential matches through historical data and specific matching criteria.

This study introduces an innovative approach to the mentor-mentee matching dilemma, deviating from conventional methods that typically involve mentees selecting mentors based on subjective research or personal inclinations. It presents a machine learning algorithm designed to systematically assess pairings by evaluating a comprehensive set of factors, such as professional skills, career goals, and personality types, offering a more intricate matching mechanism than those traditionally explored or employed by existing tools.

The integration of both collaborative and content-based filtering within a hybrid machine learning model marks a novel application of AI in mentorship pairing processes. This model extends beyond the straightforward matching of similarities by scrutinizing the distinct characteristics and demands of mentors and mentees. It assesses the compatibility of skills and objectives to ensure mentors can effectively guide mentees toward their professional aspirations. In parallel, it considers personality traits, which can impact communication and collaboration, to forecast the potential for a productive and harmonious partnership.

This study's innovation lies in its algorithmic, data-centric pairing approach, which aims to supplement or advance traditional, human-led methods of forming mentor-mentee relationships. This method opens avenues for the creation of sophisticated search tools or recommendation systems explicitly designed for mentorship programs. Such systems could offer mentees algorithmically curated mentor suggestions, thereby simplifying the pairing process and potentially elevating the quality and success rate of mentorship pairings.

Ultimately, this study enhances the mentorship literature by merging insights from academic research on mentorship with the cutting-edge developments in machine learning. It introduces a progressive methodology that harnesses AI's capabilities to improve the mentorship process, bridging theoretical knowledge with pragmatic application to forge optimal mentor-mentee pairings.

REFERENCES

Ahmed, T., Johnson, J., Latif, Z., Kennedy, N., Javier, D., Stinson, K., & Vishwanatha, J. K. (2021). MyNRMN: A national mentoring and networking platform to enhance connectivity and diversity in the biomedical sciences. *Biomedical Education*, 3, 497-509. https://doi.org/10.1096/fba.2020-00102

Avecilla, P. A. S., Capina, X. E. R., & Javier, A. Y. (2023). Teacher's Teaching Style as Perceived by Students and its Influence on Students' Level of Self-Regulation and Motivation in Learning Psychology. *Technium Social Sciences Journal*, 43, 213-240.

Barrett, J. L., Mazerolle, S. M., & Nottingham, S. L. (2017). Attributes of effective mentoring relationships for novice faculty members: perspectives of mentors and mentees. *Athletic Training Education Journal*, 12(2), 152-162. https://doi.org/10.4085/1202152

Bielczyk, N., Veldsman, M., Ando, A., Caldinelli, C., Makary, M. M., Nikolaidis, A., Scelsi, M. A., Stefan, M., & Badhwar, A. (2018). Establishing online mentorship for early career researchers: Lessons from the Organization for Human Brain Mapping International Mentoring Programme. *European Journal of Neuroscience*, 49(1069-1076). https://doi.org/10.1111/ejn.14320

Bjursell, C., & Sädbom, R. F. (2018). Mentoriship programs in the manufacting industry. *European Journal of Training and Development,* 42, 455-469. https://doi.org/10.1108/EJTD-05-2018-0044

Deng, C., Gulseren, D. B., & Turner, N. (2022). How to match mentors and proteges for successful mentorship programs: a review of the evidence and recommendations for practitioners. *Leadership & Organization Development Journal*, 43(3), 386-403. https://doi.org/10.1108/LODJ-01-2021-0032

Haas, C., Hall, M., & Vlasnik, S. L. (2018). Finding optimal mentor-mentee matches: A case study in applied two-sided matching. *Heliyon*, 4. https://doi.org/10.1016/j.heliyon.2018. e00634

Haas, C., & Hall, M. (2019). Two-Sided Matching for mentor-mentee allocations—Algorithms and manipulation strategies. *PLoS ONE*, 14(3). https://doi.org/10.1371/journal.pone.0213323

Hagler, M. A., & Rhodes, J. E. (2018). The Long-Term Impact of Natural Mentoring Relationships: A Counterfactual Analysis. Am J Community Psychol, 62(172-188). https://doi.org/10.1002/ajcp.12265

Hee, J., Toh, Y. L., Yap, H. W., Toh, Y. P., Kanesvaran, R., Mason, S., & Krishna, L. K. R. (2020). The Development and Design of a Framework to Match Mentees and Mentors Through a Systematic Review and Thematic Analysis of Mentoring Programs Between 2000 and 2015. Mentoring & Tutoring: Partnership in Learning, 28(3), 340-364. https://doi.org/10.1080/13611267.2020.1778836

Heppe, E. C. M., Kupersmidt, J. B., Kef, S., & Schuengel, C. (2018). Does having a similar disability matter for match outcomes?: A randomized study of matching mentors and mentees by visual impairment. *Journal of Community Psychology*, 47, 210-226. https://doi.org/10.1002/jcop.22116

Hoenen, S., & Kolympiris, C. (2019). The Value of Insiders as Mentors: Evidence from the Effects of NSF Rotators on Early-Career Scientists. *The Review of Economics and Statistics*, 102(5), 852-866. https://doi.org/10.1162/rest_a_00859

Kim, M. (2021). Intensive Learning Experience - Development of STEM Mentorship Program for High School Gifted Students. Gifted Child Today, 44(4), 228-235. https://doi.org/10.1177/10762175211030522

Li, H., Tan, E. L. Y., Wong, M. L., & Ong, M. M. A. (2022). Tackling study-work chasm: Perceptions of the role of mentorship in the healthcare workplace. *Medical and health professions education*, 7(3), 10-22. https://doi.org/10.29060/TAPS.2022-7-3/OA2539

Matching mentors with mentees: Practical, evidence-based recommendations. (2022). *Human Resource Management International Digest*, 30(7), 43-45. https://doi.org/10.1108/HRMID-08-2022-0231

Myers, I. B., & Myers, P. B. (2010). Gifts differing understanding personality type. *Nicholas Brealey Publishing*.

Othman, Z., Samah, K. A. F. A., Zain, N. H. M., & Zulkifli, A. F. (2023). Optimizing Sports Centre Recommendation System in Malaysia Through Content-Based Filtering Technique and Web Application. 14th Control and System Graduate Research Colloquium. https://doi.org/10.1109/ICSGRC57744.2023.10215432

Panja, S., & James, A. P. (2020). Belief Index for Fake COVID19 Text Detection. *2020 IEEE Recent Advances in Intelligent Computational Systems (RAICS) Intelligent Computational Systems (RAICS)*, 63-67. https://doi.org/10.1109/RAICS51191.2020.9332508

Pinilla, S., Pander, T., Borch, P. v. d., Fischer, M. R., & Dimitriadis, K. (2015). 5 years of experience with a large-scale mentoring program for medical students. *GMS Zeitschrift für Medizinische Ausbildung*, 32(1).

Poulsen, K. M. (2013). Mentoring programmes: learning opportunities for mentees, for mentors, for organisations and for society. *Industrial and Commercial Training*, 45(5), 255-263. https://doi.org/10.1108/ICT-03-2013-0016

Rehman, R., Khan, F., Kayani, N., & Ali, T. S. (2022). Reflection of mentors and mentees at initiation of Faculty Mentorship Program at Aga Khan University: A persepective. Pak J Med Sci, 38(6), 1691-1695. https://doi.org/10.12669/pjms.38.6.5454

Schäfer, M., Pander, T., Pinilla, S., Fischer, M. R., Borch, P. v. d., & Dimitriadis, K. (2016). A prospective, randomised trial of different matching procedures for structured mentoring programmes in medical education. *Medical Teacher*, 38(9), 921-929. https://doi.org/10.3109/0142159X.2015.1132834

Schwartz, L. P., ́nard, J. F. L., & David, S. V. (2022). Impact of gender on the formation and outcome of formal mentoring relationships in the life sciences. *PLoS Biology*, 20(9). https://doi.org/10.1371/journal.pbio.3001771

Setiawan, G. H., & Adnyana, M. B. (2023). Improving Helpdesk Chatbot Performance with Term Frequency-Inverse Document Frequency (TF-IDF) and Cosine Similarity Models. *Journal of Applied Informatics and Computing*, *7*(2), 252-257. https://doi.org/10.30871/jaic.v7i2.6527

Taylor, C. J. (2020). Mentee and mentor perceptions of a mentoring court for high-risk probationers. *Probation Journal*, 67(3), 214-227. https://doi.org/10.1177/0264550520939174

Vance, E. A., LaLonde, D. E., & Zhang, L. (2017). The Big Tent for Statistics: Mentoring Required. The American Statistician, 71(1), 15-22. https://doi.org/10.1080/00031305.2016.1247016

Vance, E. A., Tanenbaum, E., Kaur, A., Otto, M. C., & Morris, R. (2017). An Eight-Step Guide to Creating and Sustaining a Mentoring Program. The American Statistician, 71(1), 23-29. https://doi.org/10.1080/00031305.2016.1251493

Vedaswi, K., Krishna, N. V., Poojitha, T. V., Lokesh, P., K, A., & Kuma, A. (2023). Movie Recommendation using Collaborative filtering and Content-based Filtering Approach. *International Conference on Inventive Computation Technologies*. https://doi.org/10.1109/ICICT57646.2023.10134213

Wang, S., Liao, X., & Ma, K. (2023). Analysis and Detection of Orange Images Based on Improved Faster R-CNN Algorithm and Feature Data Analysis. *International Conference of Electronic Communications, Internet of Things and Big Data*. https://doi.org/10.1109/ICEIB57887.2023.10170465

Wang, S., Wang, J., Han, Y., & Zhao, Q. (2019). A Semi-supervised Knowledge Assessment Paradigm Based on T-CNN Algorithm for the Industrial Massage System. *6th International Conference on Systems and Informatics*, 499-504

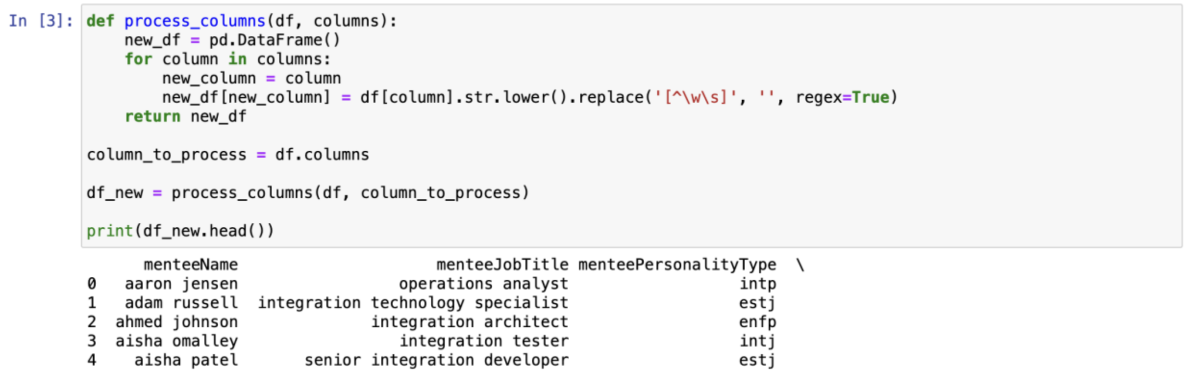
Wulf, K., Borges, N., Huggett, K., Blanco, M., Binkley, P., Moore-Clingenpeel, M., & Hurtubise, L. (2021). Personality Compatibility Within Faculty Mentoring Dyads and Perceived Mentoring Outcomes: Survey Results of Academic Medicine Institutions in the USA. Medical Science Educator, 31, 345-348. https://doi.org/10.1007/s40670-020-01191-w

Zaremarjal, A. Y., & Yiltas-Kaplan, D. (2021). Semantic Collaborative Filtering Recommender System using CNNs. *8th International Conference on Electrical and Electronics Engineering*. https://doi.org/10.1109/ICEEE52452.2021.9415931

Appendix

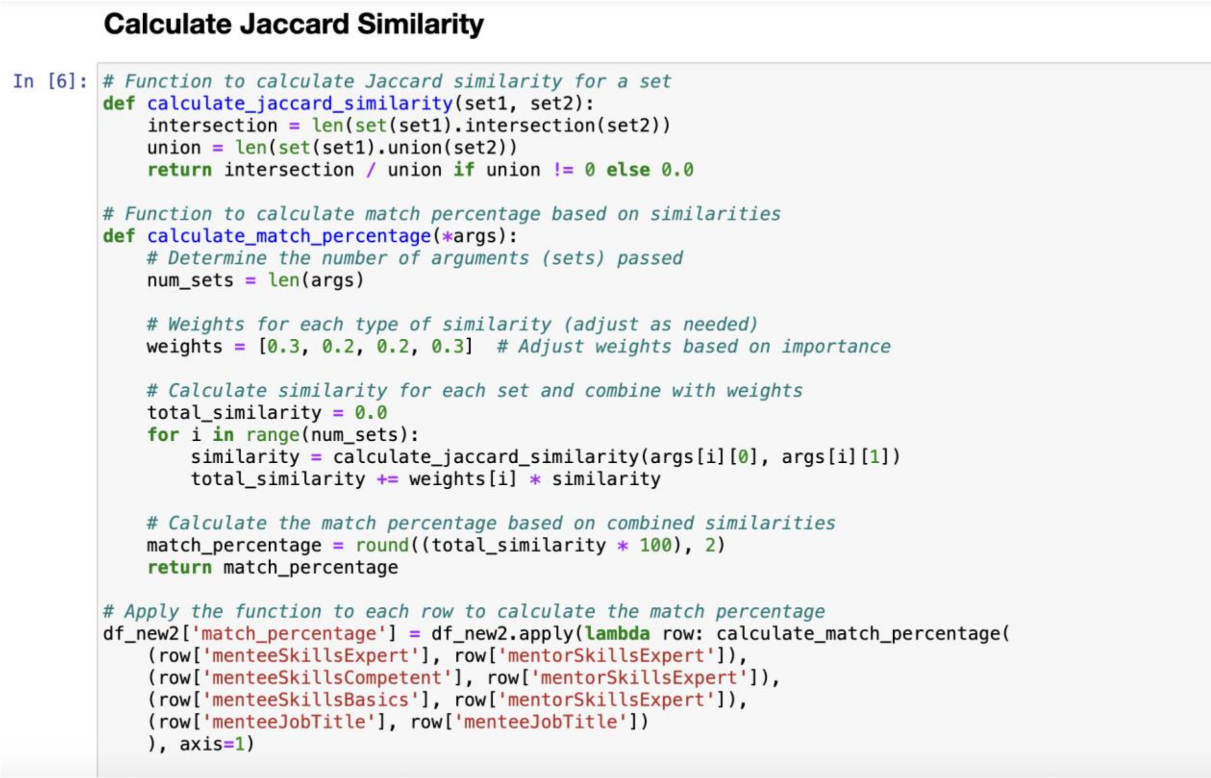
Original Code – Training the Model

Processing the Columns



```
In [3]: def process_columns(df, columns):
            new_df = pd.DataFrame()
            for column in columns:
                new_column = column
                new_df[new_column] = df[column].str.lower().replace('[^\w\s]', '', regex=True)
            return new_df

        column_to_process = df.columns

        df_new = process_columns(df, column_to_process)

        print(df_new.head())

             menteeName                      menteeJobTitle menteePersonalityType  \
        0  aaron jensen                  operations analyst                  intp
        1   adam russell  integration technology specialist                  estj
        2 ahmed johnson               integration architect                  enfp
        3 aisha omalley                  integration tester                  intj
        4    aisha patel       senior integration developer                  estj
```

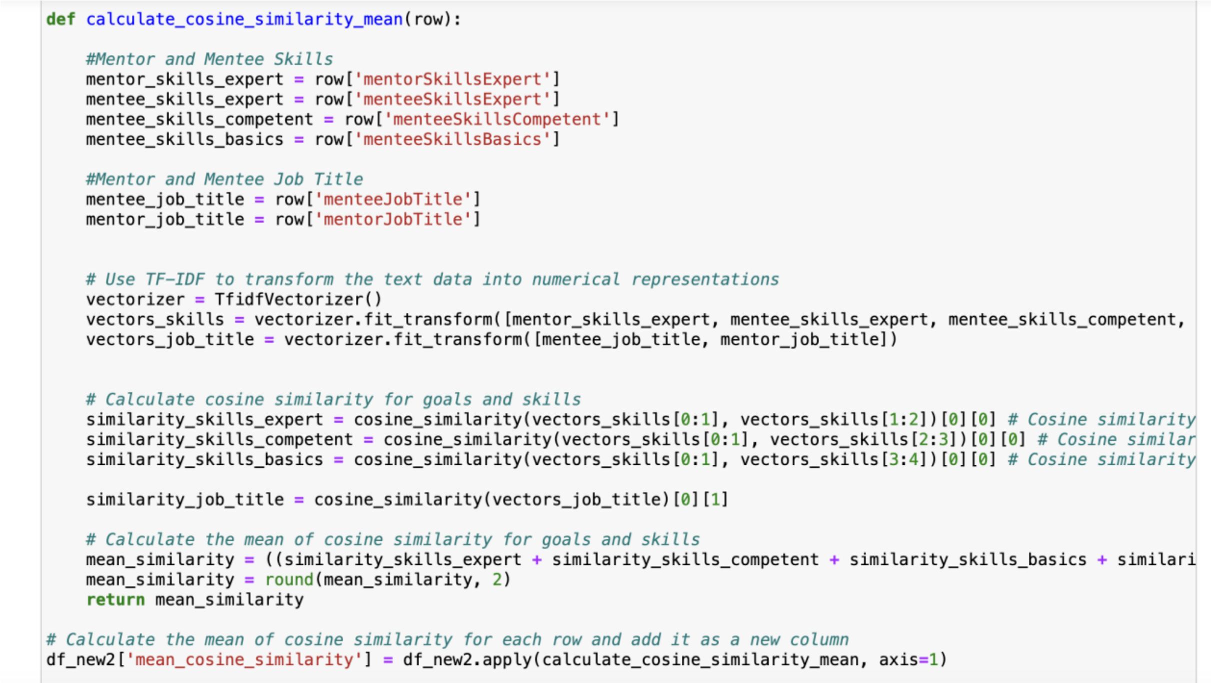
Calculate Jaccard Similarity



Calculate Jaccard Similarity

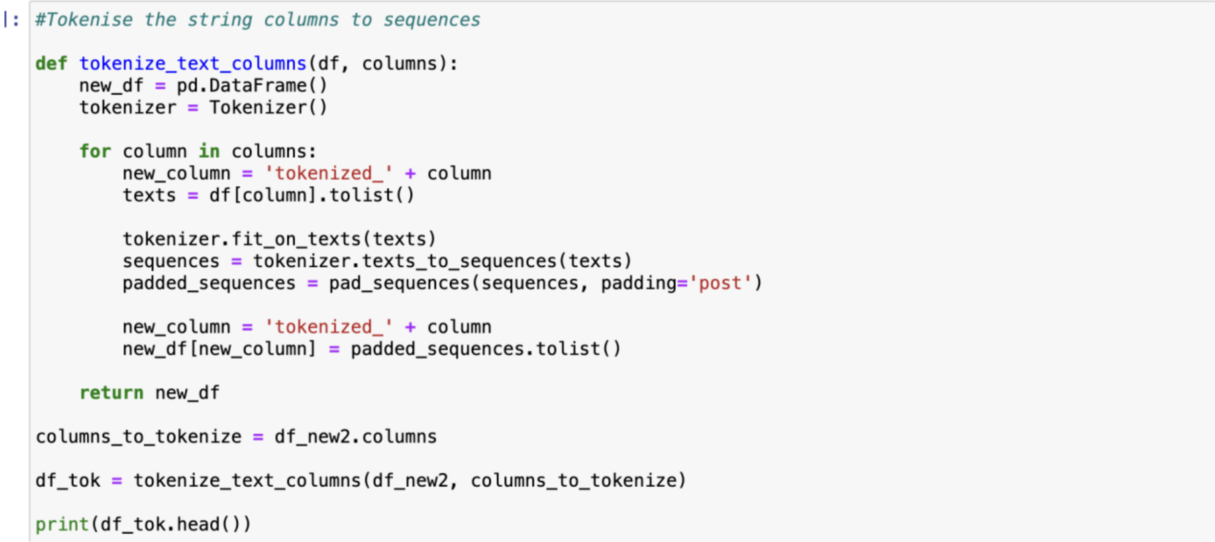
```
In [6]: # Function to calculate Jaccard similarity for a set
        def calculate_jaccard_similarity(set1, set2):
            intersection = len(set(set1).intersection(set2))
            union = len(set(set1).union(set2))
            return intersection / union if union != 0 else 0.0

        # Function to calculate match percentage based on similarities
        def calculate_match_percentage(*args):
            # Determine the number of arguments (sets) passed
            num_sets = len(args)

            # Weights for each type of similarity (adjust as needed)
            weights = [0.3, 0.2, 0.2, 0.3]  # Adjust weights based on importance

            # Calculate similarity for each set and combine with weights
            total_similarity = 0.0
            for i in range(num_sets):
                similarity = calculate_jaccard_similarity(args[i][0], args[i][1])
                total_similarity += weights[i] * similarity

            # Calculate the match percentage based on combined similarities
            match_percentage = round((total_similarity * 100), 2)
            return match_percentage

        # Apply the function to each row to calculate the match percentage
        df_new2['match_percentage'] = df_new2.apply(lambda row: calculate_match_percentage(
            (row['menteeSkillsExpert'], row['mentorSkillsExpert']),
            (row['menteeSkillsCompetent'], row['mentorSkillsExpert']),
            (row['menteeSkillsBasics'], row['mentorSkillsExpert']),
            (row['menteeJobTitle'], row['menteeJobTitle'])
            ), axis=1)
```

Calculate Cosine Similarity



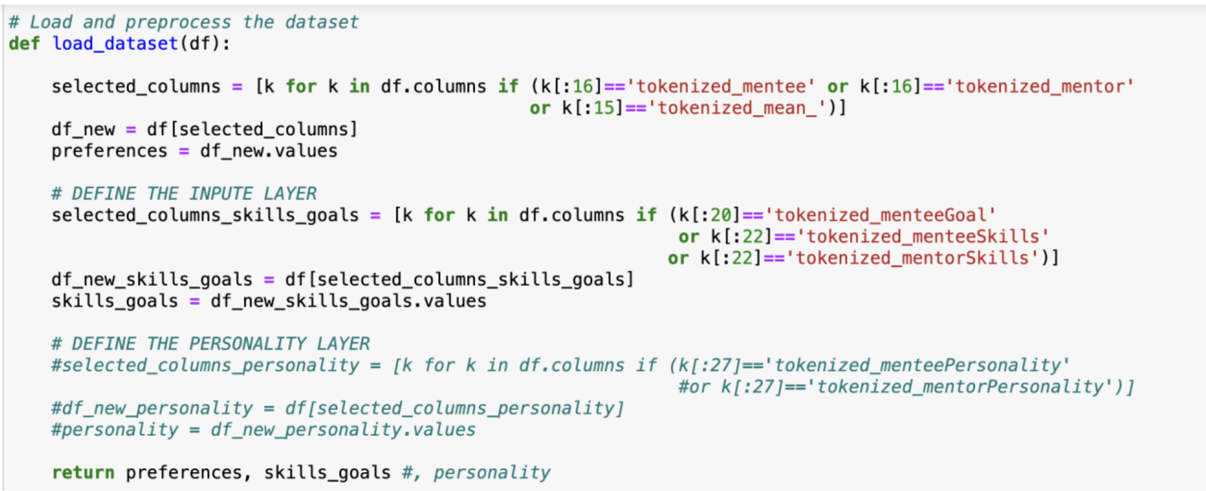
```
def calculate_cosine_similarity_mean(row):

    #Mentor and Mentee Skills
    mentor_skills_expert = row['mentorSkillsExpert']
    mentee_skills_expert = row['menteeSkillsExpert']
    mentee_skills_competent = row['menteeSkillsCompetent']
    mentee_skills_basics = row['menteeSkillsBasics']

    #Mentor and Mentee Job Title
    mentee_job_title = row['menteeJobTitle']
    mentor_job_title = row['mentorJobTitle']


    # Use TF-IDF to transform the text data into numerical representations
    vectorizer = TfidfVectorizer()
    vectors_skills = vectorizer.fit_transform([mentor_skills_expert, mentee_skills_expert, mentee_skills_competent,
    vectors_job_title = vectorizer.fit_transform([mentee_job_title, mentor_job_title])


    # Calculate cosine similarity for goals and skills
    similarity_skills_expert = cosine_similarity(vectors_skills[0:1], vectors_skills[1:2])[0][0] # Cosine similarity
    similarity_skills_competent = cosine_similarity(vectors_skills[0:1], vectors_skills[2:3])[0][0] # Cosine similar
    similarity_skills_basics = cosine_similarity(vectors_skills[0:1], vectors_skills[3:4])[0][0] # Cosine similarity

    similarity_job_title = cosine_similarity(vectors_job_title)[0][1]

    # Calculate the mean of cosine similarity for goals and skills
    mean_similarity = ((similarity_skills_expert + similarity_skills_competent + similarity_skills_basics + similari
    mean_similarity = round(mean_similarity, 2)
    return mean_similarity

# Calculate the mean of cosine similarity for each row and add it as a new column
df_new2['mean_cosine_similarity'] = df_new2.apply(calculate_cosine_similarity_mean, axis=1)
```

Tokenise the columns



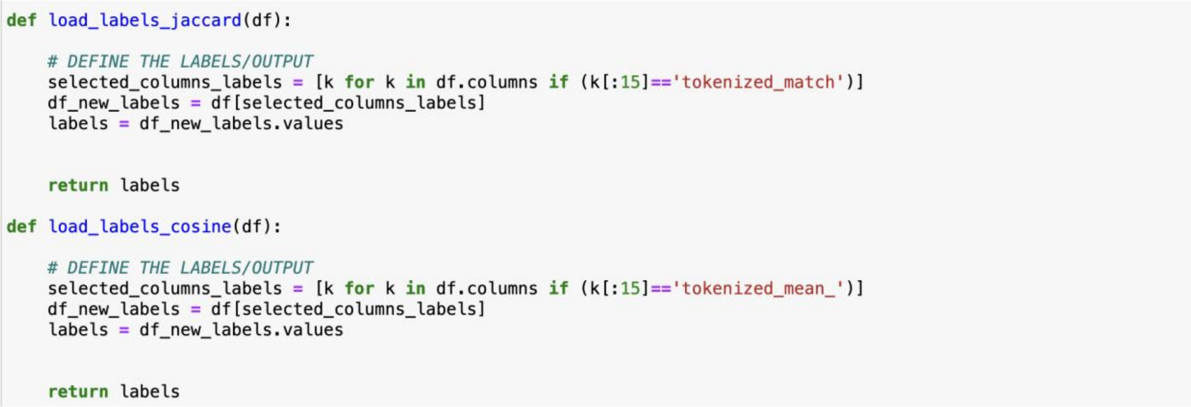
```
#Tokenise the string columns to sequences

def tokenize_text_columns(df, columns):
    new_df = pd.DataFrame()
    tokenizer = Tokenizer()

    for column in columns:
        new_column = 'tokenized_' + column
        texts = df[column].tolist()

        tokenizer.fit_on_texts(texts)
        sequences = tokenizer.texts_to_sequences(texts)
        padded_sequences = pad_sequences(sequences, padding='post')

        new_column = 'tokenized_' + column
        new_df[new_column] = padded_sequences.tolist()

    return new_df

columns_to_tokenize = df_new2.columns

df_tok = tokenize_text_columns(df_new2, columns_to_tokenize)

print(df_tok.head())
```

Load Dataset



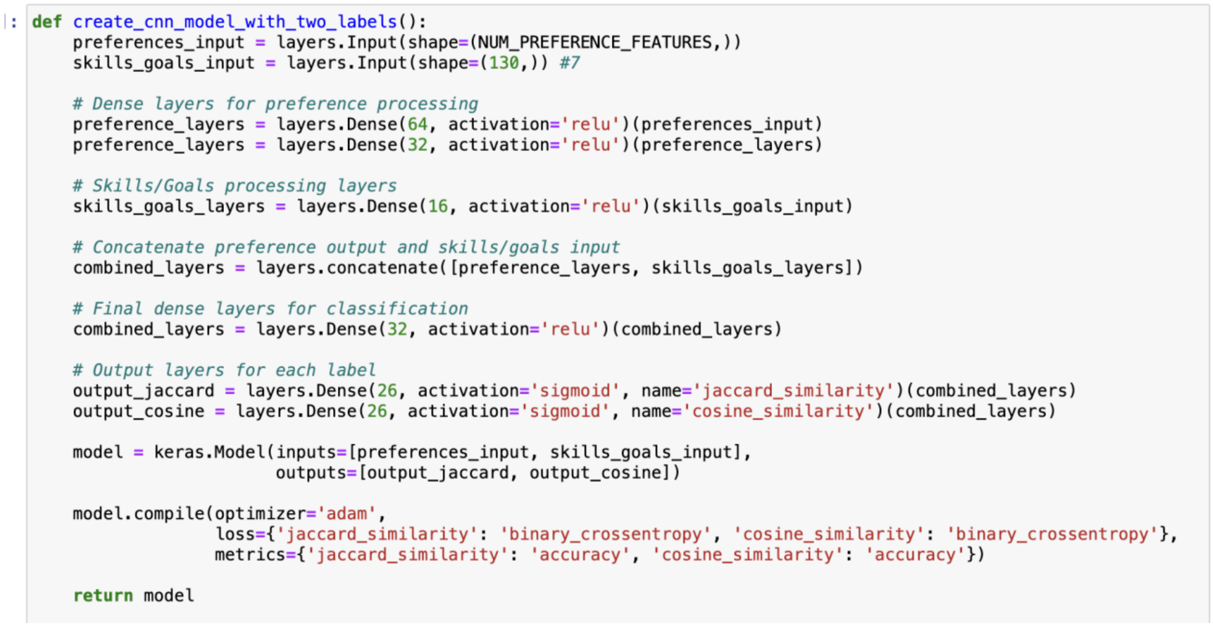
```
# Load and preprocess the dataset
def load_dataset(df):

    selected_columns = [k for k in df.columns if (k[:16]=='tokenized_mentee' or k[:16]=='tokenized_mentor'
                                                 or k[:15]=='tokenized_mean_')]
    df_new = df[selected_columns]
    preferences = df_new.values

    # DEFINE THE INPUTE LAYER
    selected_columns_skills_goals = [k for k in df.columns if (k[:20]=='tokenized_menteeGoal'
                                                               or k[:22]=='tokenized_menteeSkills'
                                                              or k[:22]=='tokenized_mentorSkills')]

    df_new_skills_goals = df[selected_columns_skills_goals]
    skills_goals = df_new_skills_goals.values

    # DEFINE THE PERSONALITY LAYER
    #selected_columns_personality = [k for k in df.columns if (k[:27]=='tokenized_menteePersonality'
                                                               #or k[:27]=='tokenized_mentorPersonality')]
    #df_new_personality = df[selected_columns_personality]
    #personality = df_new_personality.values

    return preferences, skills_goals #, personality
```

Load y1 and y2 outputs



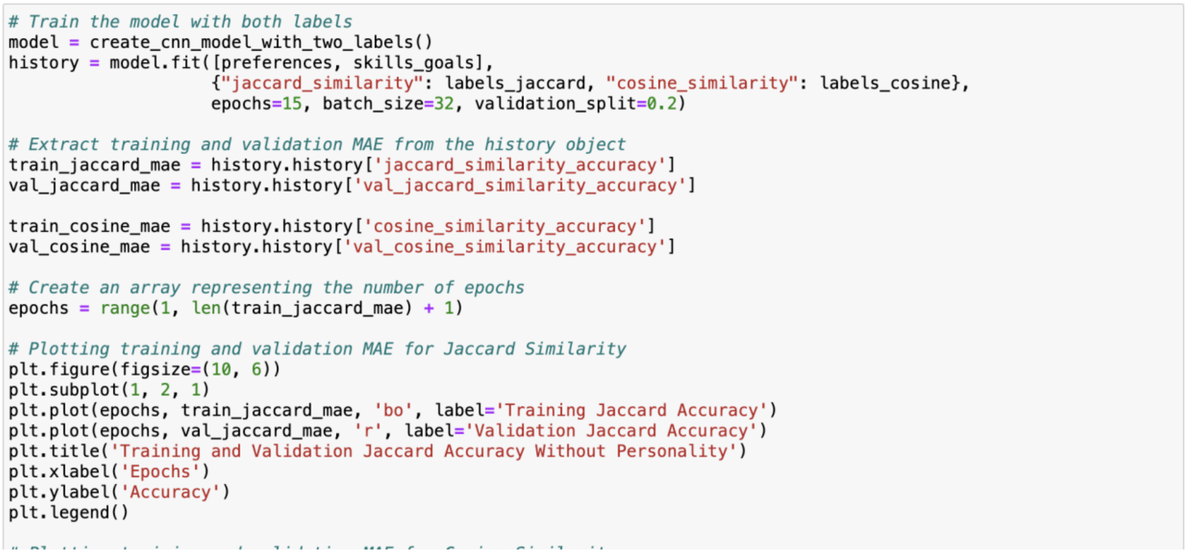
```
def load_labels_jaccard(df):

    # DEFINE THE LABELS/OUTPUT
    selected_columns_labels = [k for k in df.columns if (k[:15]=='tokenized_match')]
    df_new_labels = df[selected_columns_labels]
    labels = df_new_labels.values


    return labels

def load_labels_cosine(df):

    # DEFINE THE LABELS/OUTPUT
    selected_columns_labels = [k for k in df.columns if (k[:15]=='tokenized_mean_')]
    df_new_labels = df[selected_columns_labels]
    labels = df_new_labels.values


    return labels
```

CNN Model

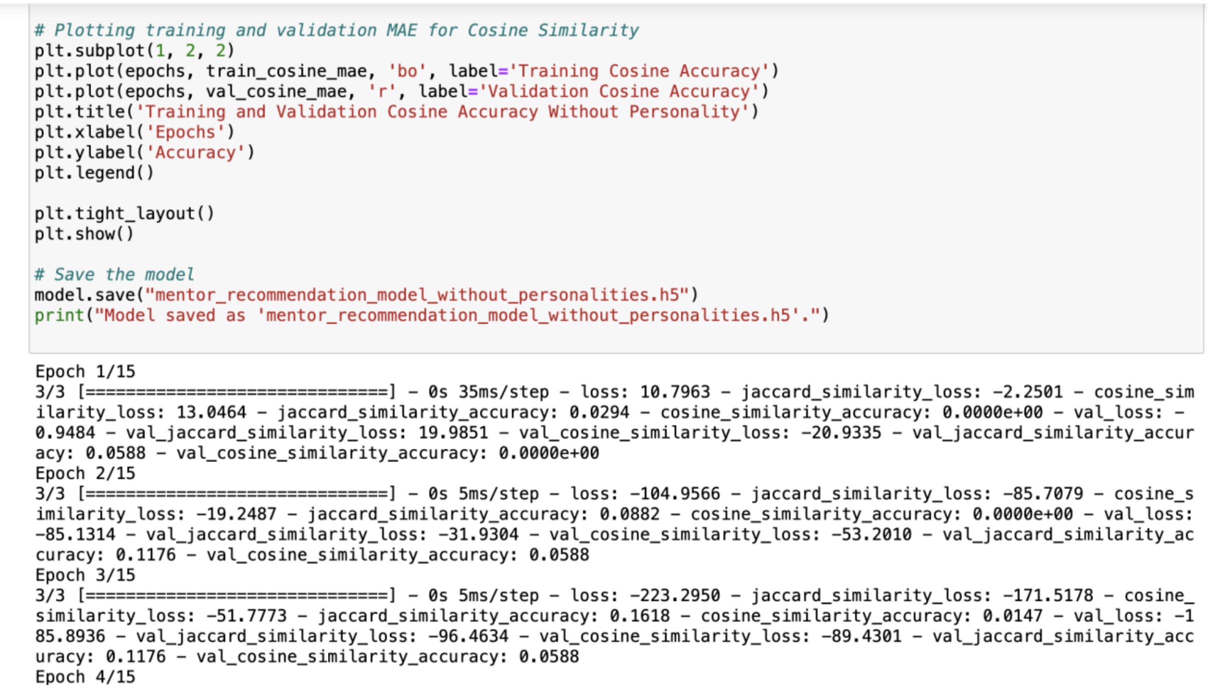


```python
def create_cnn_model_with_two_labels():
    preferences_input = layers.Input(shape=(NUM_PREFERENCE_FEATURES,))
    skills_goals_input = layers.Input(shape=(130,)) #7

    # Dense layers for preference processing
    preference_layers = layers.Dense(64, activation='relu')(preferences_input)
    preference_layers = layers.Dense(32, activation='relu')(preference_layers)

    # Skills/Goals processing layers
    skills_goals_layers = layers.Dense(16, activation='relu')(skills_goals_input)

    # Concatenate preference output and skills/goals input
    combined_layers = layers.concatenate([preference_layers, skills_goals_layers])

    # Final dense layers for classification
    combined_layers = layers.Dense(32, activation='relu')(combined_layers)

    # Output layers for each label
    output_jaccard = layers.Dense(26, activation='sigmoid', name='jaccard_similarity')(combined_layers)
    output_cosine = layers.Dense(26, activation='sigmoid', name='cosine_similarity')(combined_layers)

    model = keras.Model(inputs=[preferences_input, skills_goals_input],
                        outputs=[output_jaccard, output_cosine])

    model.compile(optimizer='adam',
                  loss={'jaccard_similarity': 'binary_crossentropy', 'cosine_similarity': 'binary_crossentropy'},
                  metrics={'jaccard_similarity': 'accuracy', 'cosine_similarity': 'accuracy'})

    return model
```

Running the model



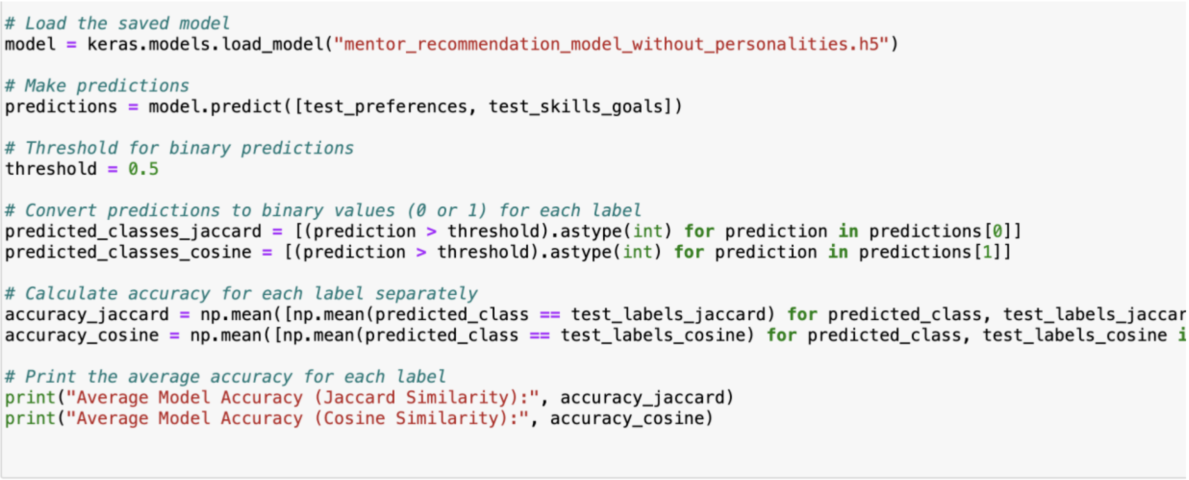
```python
# Train the model with both labels
model = create_cnn_model_with_two_labels()
history = model.fit([preferences, skills_goals],
                    {"jaccard_similarity": labels_jaccard, "cosine_similarity": labels_cosine},
                    epochs=15, batch_size=32, validation_split=0.2)

# Extract training and validation MAE from the history object
train_jaccard_mae = history.history['jaccard_similarity_accuracy']
val_jaccard_mae = history.history['val_jaccard_similarity_accuracy']

train_cosine_mae = history.history['cosine_similarity_accuracy']
val_cosine_mae = history.history['val_cosine_similarity_accuracy']

# Create an array representing the number of epochs
epochs = range(1, len(train_jaccard_mae) + 1)

# Plotting training and validation MAE for Jaccard Similarity
plt.figure(figsize=(10, 6))
plt.subplot(1, 2, 1)
plt.plot(epochs, train_jaccard_mae, 'bo', label='Training Jaccard Accuracy')
plt.plot(epochs, val_jaccard_mae, 'r', label='Validation Jaccard Accuracy')
plt.title('Training and Validation Jaccard Accuracy Without Personality')
plt.xlabel('Epochs')
plt.ylabel('Accuracy')
plt.legend()

# Plotting training and validation MAE for Cosine Similarity
```



```
# Plotting training and validation MAE for Cosine Similarity
plt.subplot(1, 2, 2)
plt.plot(epochs, train_cosine_mae, 'bo', label='Training Cosine Accuracy')
plt.plot(epochs, val_cosine_mae, 'r', label='Validation Cosine Accuracy')
plt.title('Training and Validation Cosine Accuracy Without Personality')
plt.xlabel('Epochs')
plt.ylabel('Accuracy')
plt.legend()

plt.tight_layout()
plt.show()

# Save the model
model.save("mentor_recommendation_model_without_personalities.h5")
print("Model saved as 'mentor_recommendation_model_without_personalities.h5'.")
```

```
Epoch 1/15
3/3 [==============================] - 0s 35ms/step - loss: 10.7963 - jaccard_similarity_loss: -2.2501 - cosine_sim
ilarity_loss: 13.0464 - jaccard_similarity_accuracy: 0.0294 - cosine_similarity_accuracy: 0.0000e+00 - val_loss: -
0.9484 - val_jaccard_similarity_loss: 19.9851 - val_cosine_similarity_loss: -20.9335 - val_jaccard_similarity_accur
acy: 0.0588 - val_cosine_similarity_accuracy: 0.0000e+00
Epoch 2/15
3/3 [==============================] - 0s 5ms/step - loss: -104.9566 - jaccard_similarity_loss: -85.7079 - cosine_s
imilarity_loss: -19.2487 - jaccard_similarity_accuracy: 0.0882 - cosine_similarity_accuracy: 0.0000e+00 - val_loss:
-85.1314 - val_jaccard_similarity_loss: -31.9304 - val_cosine_similarity_loss: -53.2010 - val_jaccard_similarity_ac
curacy: 0.1176 - val_cosine_similarity_accuracy: 0.0588
Epoch 3/15
3/3 [==============================] - 0s 5ms/step - loss: -223.2950 - jaccard_similarity_loss: -171.5178 - cosine_
similarity_loss: -51.7773 - jaccard_similarity_accuracy: 0.1618 - cosine_similarity_accuracy: 0.0147 - val_loss: -1
85.8936 - val_jaccard_similarity_loss: -96.4634 - val_cosine_similarity_loss: -89.4301 - val_jaccard_similarity_acc
uracy: 0.1176 - val_cosine_similarity_accuracy: 0.0588
Epoch 4/15
```

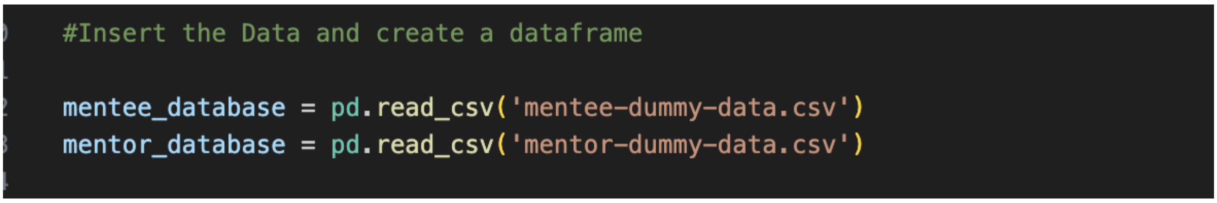
Test Predictions



```
# Load the saved model
model = keras.models.load_model("mentor_recommendation_model_without_personalities.h5")

# Make predictions
predictions = model.predict([test_preferences, test_skills_goals])

# Threshold for binary predictions
threshold = 0.5

# Convert predictions to binary values (0 or 1) for each label
predicted_classes_jaccard = [(prediction > threshold).astype(int) for prediction in predictions[0]]
predicted_classes_cosine = [(prediction > threshold).astype(int) for prediction in predictions[1]]

# Calculate accuracy for each label separately
accuracy_jaccard = np.mean([np.mean(predicted_class == test_labels_jaccard) for predicted_class, test_labels_jaccar
accuracy_cosine = np.mean([np.mean(predicted_class == test_labels_cosine) for predicted_class, test_labels_cosine i

# Print the average accuracy for each label
print("Average Model Accuracy (Jaccard Similarity):", accuracy_jaccard)
print("Average Model Accuracy (Cosine Similarity):", accuracy_cosine)
```

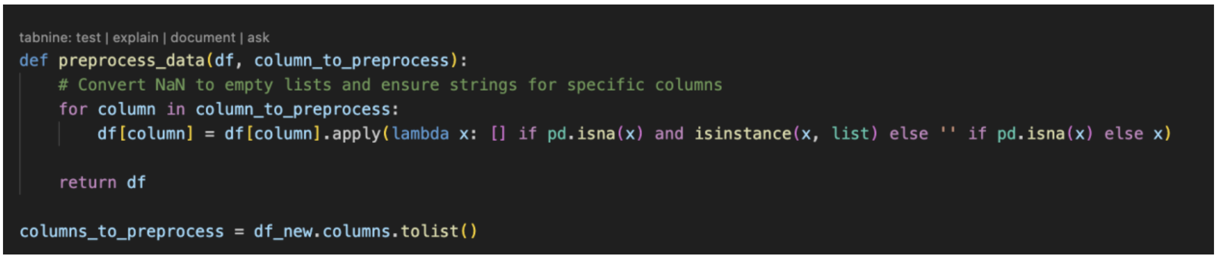
Implementing the Model with a dummy data

Using mentor and mentee data from the form



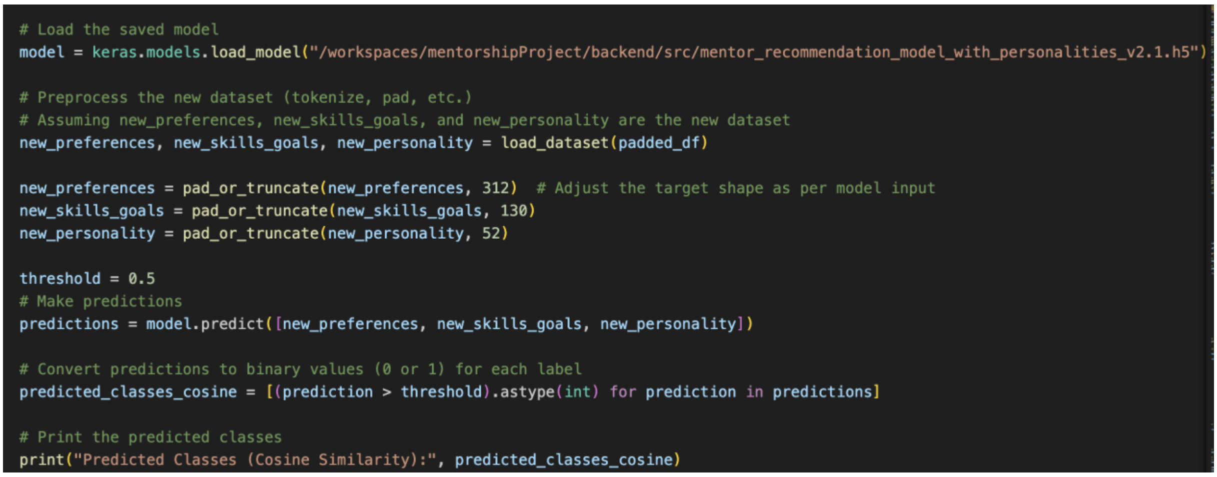
```
#Insert the Data and create a dataframe

mentee_database = pd.read_csv('mentee-dummy-data.csv')
mentor_database = pd.read_csv('mentor-dummy-data.csv')
```

Pre process the data

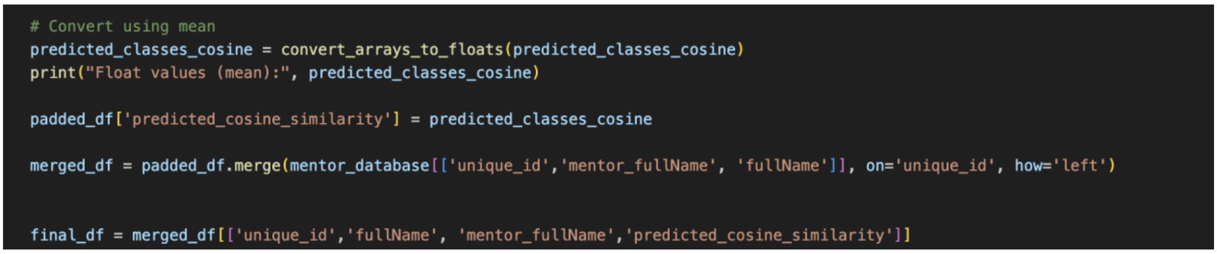


```
tabnine: test | explain | document | ask
def preprocess_data(df, column_to_preprocess):
    # Convert NaN to empty lists and ensure strings for specific columns
    for column in column_to_preprocess:
        df[column] = df[column].apply(lambda x: [] if pd.isna(x) and isinstance(x, list) else '' if pd.isna(x) else x)

    return df

columns_to_preprocess = df_new.columns.tolist()
```

Call the model

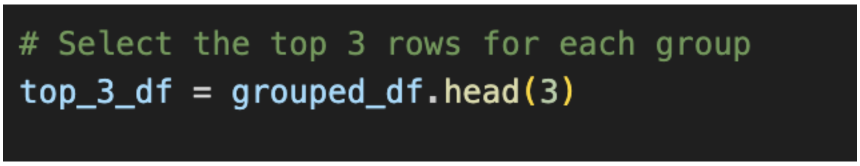


```
# Load the saved model
model = keras.models.load_model("/workspaces/mentorshipProject/backend/src/mentor_recommendation_model_with_personalities_v2.1.h5")

# Preprocess the new dataset (tokenize, pad, etc.)
# Assuming new_preferences, new_skills_goals, and new_personality are the new dataset
new_preferences, new_skills_goals, new_personality = load_dataset(padded_df)

new_preferences = pad_or_truncate(new_preferences, 312)  # Adjust the target shape as per model input
new_skills_goals = pad_or_truncate(new_skills_goals, 130)
new_personality = pad_or_truncate(new_personality, 52)

threshold = 0.5
# Make predictions
predictions = model.predict([new_preferences, new_skills_goals, new_personality])

# Convert predictions to binary values (0 or 1) for each label
predicted_classes_cosine = [(prediction > threshold).astype(int) for prediction in predictions]

# Print the predicted classes
print("Predicted Classes (Cosine Similarity):", predicted_classes_cosine)
```

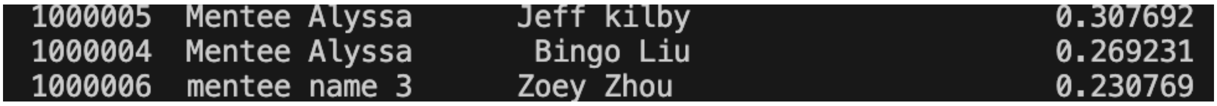


```
# Convert using mean
predicted_classes_cosine = convert_arrays_to_floats(predicted_classes_cosine)
print("Float values (mean):", predicted_classes_cosine)

padded_df['predicted_cosine_similarity'] = predicted_classes_cosine

merged_df = padded_df.merge(mentor_database[['unique_id','mentor_fullName', 'fullName']], on='unique_id', how='left')


final_df = merged_df[['unique_id','fullName', 'mentor_fullName','predicted_cosine_similarity']]
```

Top 3 Matched Mentors for Mentees

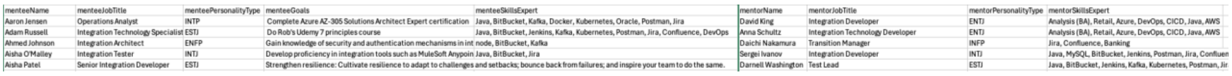


```
# Select the top 3 rows for each group
top_3_df = grouped_df.head(3)
```



```
1000005  Mentee Alyssa      Jeff kilby             0.307692
1000004  Mentee Alyssa       Bingo Liu             0.269231
1000006  mentee name 3      Zoey Zhou              0.230769
```

Screenshot of a Sample Data



menteeName	menteeJobTitle	menteePersonalityType	menteeGoals	menteeSkillsExpert	mentorName	mentorJobTitle	mentorPersonalityType	mentorSkillsExpert
Aaron Jensen	Operations Analyst	INTP	Complete Azure AZ-305 Solutions Architect Expert certification	Java, BitBucket, Kafka, Docker, Kubernetes, Oracle, Postman, Jira	David King	Integration Developer	ENTJ	Analysis (BA), Retail, Azure, DevOps, CICD, Java, AWS
Adam Russell	Integration Technology Specialist	ESTJ	Do Rob's Udemy 7 principles course	Java, BitBucket, Jenkins, Kafka, Kubernetes, Postman, Jira, Confluence, DevOps	Anna Schultz	Integration Technology Developer	ENTJ	Analysis (BA), Retail, Azure, DevOps, CICD, Java, AWS
Ahmed Johnson	Integration Architect	ENFP	Gain knowledge of security and authentication mechanisms in int	node, BitBucket, Kafka	Daichi Nakamura	Transition Manager	INFP	Jira, Confluence, Banking
Aisha O'Malley	Integration Tester	INTJ	Develop proficiency in integration tools such as MuleSoft Anypoin	Java, BitBucket, Jira	Sergei Ivanov	Integration Developer	INTJ	Java, MySQL, BitBucket, Jenkins, Postman, Jira, Confluen
Aisha Patel	Senior Integration Developer	ESTJ	Strengthen resilience: Cultivate resilience to adapt to challenges and setbacks; bounce back from failures; and inspire your team to do the same.		Darnell Washington	Test Lead	ESTJ	Java, BitBucket, Jenkins, Kafka, Kubernetes, Postman, Jir