FYP2-Report-01072026

by Anis Hazirah BINTI MOHAMAD SABRY

Submission date: 01-Jul-2024 09:19PM (UTC+0800)
Submission ID: 2411202268
File name: ID-2530-PIACC-ProblemSolving-Anis-1211300373_REPORT_V1_CHECK.pdf (2.78M)
Word count: 17435
Character count: 90820

Chapter 1 Introduction

Chapter 1.1 Overview

In this era of technological advancements, countries across the globe are in a rat race to be able to maximise the possible avenues of their country as efficiently as possible. For this to happen, mass amounts of data concerning the population is required to uncover and disseminate information The Organisation for Economic Growth and Development (OECD) took the initiative to make this a possibility with the implementation of a comparative survey of adults titled 'The Survey of Adult Skills'. With this they had developed the Programme for the International Assessment of Adult Competencies (PIAAC), an international level survey that divulges into the critical need about the distribution of knowledge, skills, and characteristics that have become vital key factors for an optimised participation in modern societies. The PIAAC data aims to provide policymakers a baseline of their populations' levels of knowledge, skills, and competencies as elements towards personal and societal success, creating a gauge of economic outcomes and recommendations towards enhancing the countries' human capital. They provide a deeper understanding of the processes behind skill gains, loss, and retainment, allowing a more dynamic knowledge base in regards to these issues.

Adults aged 16 to 65 were surveyed over 29 countries between the years 2011-2017. Out of these countries, 24 were able to report the results of their assessments. The countries involved were Australia, Austria, Canada, Cyprus, Czech Republic, Denmark, United Kingdom (England and Northern Ireland), Estonia, Finland, Flanders (Belgium), France, Germany, Ireland, Italy, Japan, South Korea, Netherlands, Norway, Poland, Russian Federation, Slovak Republic, Spain, Sweden, and the United States.

To comprehend PIAAC and be able to dissect it at its utmost peak, a background of the surveys' assessment must be considered. The survey collects a myriad of background information of the participants, considering their qualifications, work experiences, skill usage, and additional training whether it is formal or informal. The survey is designed as a computer-based assessment, having participants answer in

the comforts of their own home or via another location that has been agreed upon by the interviewer. No time limits were imposed during the assessment, but instead were taken into consideration and used as an additional variable to be considered.

PIAAC assesses three domains of cognitive skills- literacy, numeracy, and problem solving skills in technology-rich environments (PSTRE). Problem solving skills is "using digital technology, communication tools and networks to acquire and evaluate information, communicate with others and perform practical tasks."(OECD, 2019, p. 55) In the first wave of PIAAC however, it focuses on "the abilities to solve problems for personal, work and civic purposes by setting up appropriate goals and plans, and accessing and making use of information through computers and computer networks" (OECD, 2019, p. 55). PSTRE covers a domain of competence that intersects with computer literacy and the cognitive skills involved to solve problems associated with technology-rich environments. Having said that, the prime is not to assess the usage of information and communication technology (ICT) tools or applications on its own, but rather the level of comprehension required to process, evaluate, and analyse information with the usage of these peripherals. (Kirch, Thorn, Educational Testing Service (ETS), and OECD, 2019).

Economic growth within the Asian region proves that efficiency for work is a critical factor in keeping up with the rapid expansion. For instance, South Korea is attributed to its economic stability after a rough patch suffering poverty and socio-economic stability, historically marking it as an achievement of sustainable growth (Lee, Jeong, & Hong, 2018). In the current era, the sustainability of the growth can be viewed more prominently through younger generations as opposed to older generations in South Korea via their productivity levels (Lee, Kwak, & Song, 2022).

Several factors contribute to the stability of productivity. Workplace learning is one of the measurements in regards to raising productivity defined by learning by working. This subject matter, although important in terms of improving work efficiency, has not been heavily explored with PIAAC data (Olsen & Tikkanen, 2018). Importantly, there is an empirical difference that can be observed between the culture

of Asia and that of Western countries. This cultural distinction adds unique layers to the challenges faced in sustaining economic growth. Examining the cultural differences, it becomes evident that, since the publication of the PIAAC dataset, research focusing on Asian countries has not been proportionate to studies conducted in European or American contexts (Maehler, Konradt, Morozova, Dickson, Braun, & Jakowatz, 2023). This imbalance raises the need for a more comprehensive understanding of the region as a whole as opposed to observing a single unique country.

In recognising the scarcity of research done on Asian countries, particularly those that encompass the region as whole, this project aims to dive deeper towards understanding the sustainability of Asia's economy. Particularly in regards to the role of adaptive problem solving as a key to personal growth and success in the context of the Asian working environment. Through an analysis of the relationship between education levels, working environments, and problem-solving skills, the research aims to illuminate the key features influencing work efficiency in the region.

Chapter 1.2 Problem Statement

Work efficiency in Asia is a multifaceted issue that is obscured by layers of challenges, warranting thorough exploration and analysis. At its core, the cultural differences pose a fundamental question: Why is Asia consistently linked with a reputation for hardworking individuals? This inquiry goes beyond a mere observation; it represents a complex issue with profound implications for the composition and dynamics of the region's labour force.

The motivation behind embarking on this task to unravel and understand the intricacies inherent in Asian working culture stems from the necessity to comprehend the profound significance of work ethics in this region. By delving into the driving forces behind Asian work culture, we seek to uncover patterns and insights crucial for addressing challenges related to employment, skill utilisation, and the educational landscape of the workforce.

Chapter 1.3 Research Objectives

- Examining factors that influence working over the required working hours and under the required working hours.
- Conducting association rule analysis to identify factors influencing the mismatch between highest qualifications and employment qualifications.
- Investigating how individuals address the skill gap when they lack the necessary qualifications in their current profession

Chapter 1.4 Project Scope

The study focuses on analysing occupational mismatch, with the focus of problem solving skills, using the first cycle (2011 to 2017) of PIAAC dataset for Asian countries. Those countries are Japan, South Korea, Kazakhstan, and Singapore. This project intends to:

1. Examine the factors and relationships concerning work attributes. This involves researching variables that would influence working over or under the average working hours.
2. Identifying if a skill mismatch impacts working hours and comparing the information across different work sectors.
3. Classifying the worker mismatch by their education mismatch and working hours.

Chapter 1.5 Gantt Charts

The first gantt chart details the timeline of actions taken during the first phase of FYP. The first 5 weeks were the exploration of the problem statement. For the literature review, it took place from week 2 and ended in week 10, comprising of reviewing research done on international and Asia-centric papers. At the same time, a preliminary data exploration was conducted during week 6 to week 10. The actual project implementation began in week 7 with the cleaning and pre-processing of the data Applying machine learning models began in week 9. Finally, the process of writing the report began in week 12 until week 14.

FYP1 GANTT CHART	FYP ID:	2530
	FYP TITLE:	Skill Analytics for Personal Growth and Success
	STUDENT NAME:	ANIS HAZIRAH BINTI MOHAMAD SABRY
	STUDENT ID:	1211300373

No.	Task title	WEEK												
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Problem Statement														
2	Background Study/Literature Review														
2.1.	International papers														
2.2.	Asia-centric papers														
3	Preliminary Data Exploration														
4	Project Implementation														
4.1.	Cleaning Data														
4.2.	Pre-processing Data														
4.3.	Modeling the Data														
5	Report Writing														

Figure 1.1 FYP1 Gantt chart

For the second phase of FYP, the process of refining the dataset took place from week 1 to week 4, these include feature selection and conducting an exploratory data analysis. The second segment, data mining implementation, began at week 2 and ended on week 10, with reassessment of the framework, applying machine learning models, and applying association rule mining being the activities done in that span of time. The remaining weeks were left for assessing the model results and writing the report.

FYP2 GANTT CHART	FYP ID:	2530
	FYP TITLE:	Skill Analytics for Personal Growth and Success
	STUDENT NAME:	ANIS HAZIRAH BINTI MOHAMAD SABRY
	STUDENT ID:	1211300373

No.	Task title	WEEK													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Dataset Refinement														
1.1	Feature Selection														
1.2	Exploratory Data Analysis														
2	Data Mining Implementation														
2.1	Reassessment of Framework														
2.2	Applying Machine Learning Models														
2.3	Applying Association Rule Mining														
3	Report Writng														
3.1	Investigation and Comparision of Model Results														
3.2	Report Writing														

Figure 1.2 FYP2 Gantt chart

Chapter 1.5 Chapter Organisation

The initial chapter starts off with introducing the problem that will be tackled in this project. This includes an overview of the project in its entirety, the problem statement itself, the three research objectives of this project, and the gantt charts for both phases of FYP.

The following chapter, literature review, details on the background research found for this project. It comes with three subchapters; those are the literature review of papers concerning analysis of the PIAAC data as a whole, papers concerning analysis of PIAAC data for the Asia region only, and the analysis of the Asia PIAAC PSTRE data in the context of work.

The third, fourth, and fifth chapters explains the theoretical framework, that is the concepts that will be tackled in the projects, the research methodology which details the plan of action for this project, and how the project is implemented. Concepts that will be explained includes the PIAAC data usage itself, the concept of skill mismatch and working hours in Asia, the machine learning models that will be used and the association rule mining. The next two chapters details on the findings acquired and answers the concepts and problems in the prior chapters.

Chapter 2. Literature Review

Chapter 2.1 International Analysis of Problem Solving Skills in the Context of Work

An investigation is conducted to evaluate the methods and factors associated with the correlation between problem-solving skills and efficiency in a workplace setting. This section encompasses studies conducted on a global level or in regions outside of Asia.

A study by Liao, He, and Jiao (2019) showcases that education has a high correlation with a high level of problem solving. In regards to work variables, it was observed that individuals who were employed and received payment, had exposure to relevant experiences, regularly dealt with problem solving tasks, possessed advanced computer skills, worked in more specialised occupations, and earned a higher monthly income were more likely to achieve a high score in terms of problem solving.

With this in mind, it should be noted that using this as a means of measurement may not be equitable for individuals from diverse backgrounds. Immigrants, particularly in the United States, frequently face racial prejudice. Pivovarova and Powers (2022) found that Asian Americans and black workers often possess credentials that far exceed the minimum requirements for their respective occupational categories. This phenomenon is particularly prevalent among first-generation immigrants in comparison to third-generation immigrants.

Similarly, Vocational Education and Training (VET) workers face distinct and ever-changing challenges, which differ from those encountered by individuals pursuing general education courses. These workers are prone to having lower problem-solving skill scores and often require additional support compared to others in order to enhance their professional competencies. The reason for this is the discrepancy between VET courses and general education courses (Hämäläinen, De Wever, Nissinen, & Cincinnato, 2017).

On a similar note, Olsen and Tikkanen (2018) conducted a literature review to evaluate the influence and significance of workplace learning (WPL) and the extent to which the PIAAC data enables a more comprehensive analysis of this subject. WPL is an essential component of lifelong learning (LLL) since the advancement of new technologies necessitates a greater level of expertise in applying skills to adapt to them. Out of the 60 articles found, only 7 of them had relevance towards WPL. These topics range from VET, opportunities to informal learning, and participation in learning at work. The authors note the scarcity of research based on WPL and advocate for its untapped potential. According to them, the PIAAC data has the potential to offer a more thorough examination of WPL, including interesting factors like learner profiles and behaviours, to gain a deep understanding of how an individual implements WPL.

Further complexities arise once the occupational sector is considered. Many educators frequently demonstrate a skill mismatch, which is characterised by a discrepancy between the demands of their current job and their highest level of education. The misalignment leads to a perception of educators having reduced levels of professionalisation (Grotlüschen, Stammer, and Sork, 2020).

Another perspective to consider is investigation of the PSTRE via process data. He, Shi, and Tighe (2023) uses the process data of particpants from the United States who partook in the second cluster of the PIAAC PSTRE. A multiclass hierarchal classification was conducted with the models Random Forest Classifier and Support Vector Machine (SVM). Their study concludes that hierarchical classification models moderately outperform flat classification models in predicting proficiency levels. With the emergence of Artificial Intelligence (AI) tools being widespread, the authors heavily emphasise the use of process data as a validity check to see if the response was entirely human.

Table 2.1 Tabulated Analysis of PIAAC as a Whole

Year	Title	Author(s)	Region	Scope	Method
2023	Predicting Problem-Solving	Qiwei He, Qingzhou Shi,	United States	Predicting proficiency levels of	Using process data

	Proficiency with Multiclass Hierarchical Classification Using Process Data: A Machine Learning Approach	Elizabeth L. Tighe		PSTRE	Random Forest and Support Vector Machine (SVM)
2022	Do Immigrants Experience Education-Job Mismatch? New Evidence from the U.S. PIAAC	Margarita Pivovarova, Jeanne M. Powers	United States	Employed Immigrant Workers	Multinomial Logistic Regressions
2020	People Who Teach Regularly: What Do We Know From PIAAC about Their Professionalization?	Anke Grotlüschen, Christopher Stammer, Thomas J Sork	Japan, Republic of Korea, Singapore, Denmark, Finland, Germany, Czech Republic, Slovenia, Ireland, United Kingdom, Italy, Spain, Turkey, and Canada	Educators in the mentioned countries	IDB Analyzer SPSS programmes
2019	Mapping Background Variables With Sequential Patterns in Problem-	Dandan Liao, Qiwei He, Hong Jiao	United States	Employed Workers	Regression Analysis, Chi-square Focused primarily on U02

	Solving Environments: An Investigation of United States Adults' Employment Status in PIAAC				'Meeting Room Assignment'
2018	The Developing Field of Workplace Learning and the Contribution of PIAAC	Dorothy Sutherland Olsen, Tarja Tikkanen	All	Workplace Learning	Literature review
2017	Understanding Adults' Strong Problem-Solving Skills Based on PIAAC	Raija Hämäläinen, Bram De Wever, Kari Nissinen, Sebastiano Cincinnato	Europe	VET workers	Binary Logistic Regression

Chapter 2.2 Analysis of PIAAC Data for Asian Regions

It can be argued that the PIAAC data is not a proper representation of the human capital found in Asian regions. Komatsu and Rappleye (2019) voice out the lack of consideration towards the cultural differences in OECD'S reports. OECD does not take into the account the historical attributes that would result in the economic prosperity that blooms from the years of detrimental conditions these countries had the unfortunate luck of experiencing. They commented that policy recommendations are simplistic in its views and too ideological, as opposed to being rooted in reality.

The authors also observed that education is not the primary determinant of economic growth, but rather a contributing factor. They propose that policymakers

should not prioritise education as the primary basis for development. Instead, they advocate for a more moderate approach in devising solutions to achieve their goals.

Table 1.2 Tabulated Analysis of PIAAC in Asian Regions

Year	Region	Title	Author(s)	Scope	Method
2019	Asia, with focus in South Korea and Singapore	Refuting the OECD-World Bank Development Narrative: Was East Asia's 'Economic Miracle' Primarily Driven by Education Quality and Cognitive Skills?	Hikaru Komatsu, Jeremy Rappleye	Economic Growth	Pearson Correlation Coefficient (r)

Chapter 2.3 Analysis of Problem Solving Skills Asia in the Context of Work

To fully understand the roles at play to output the best possible productivity levels in a work environment, Lee, Kwak, et al. (2022) builds upon previous research of examining age effects on productivity, the effect of ICT skills on productivity, followed by assessing how job training participation improves productivity. They find that as workers get older, their productive levels and skills may only be maintained with adequate participation in job training.

Lim, Ryu, and Jin (2020) further explore this subject by categorising older workers based on typology. The authors examine the characteristics of older workers by analysing their engagement in learning activities, which encompass their readiness to learn, their involvement in informal learning inside the workplace, and their participation in informal learning opportunities. Their research uncovers three distinct categories: The Dormant Workers with Low Skills, The Educated Workers in the Public Sector, and The Educated Workers in Flexible Working Conditions. They emphasise the need for organisations to embrace a non-traditional approach towards ageing workers, allowing flexibility in order for them to fully satisfy their needs and

job objectives. Consequently, the support given to these workers aids in maintaining a competent level of problem solving skill.

In their study, Yoon, Hur, and Kim (2020) reported consistent findings regarding varying levels of problem-solving competencies across different age cohorts. Typically, older workers tend to exhibit a decrease in their problem solving abilities, indicating a strong correlation between age and experience. Another argument for this reasoning is that adolescents are exposed to a significantly greater number of problem-solving opportunities in their educational experiences. Furthermore, the rapid introduction of new technologies poses challenges for older generations who struggle to comprehend them as swiftly as their younger counterparts. The authors demonstrated that engaging in interactions with colleagues within the workplace is able to counteract this effect.

In addition, Lee, Han, and Son (2019) propose another way to address the degradation of problem solving competencies in older generations. They vouch for conduction of vocational training as a supplement in preparing workers with adequate skills. This fosters the workers' understanding of the latest technological advancements, ensuring they stay informed and enhancing their career opportunities by incorporating life-long learning.

Additionally, a study by Jyung, Lee, Park, Cho, and Choi (2020) claims skill usage is the most effective way to avoid deterioration of skill competencies. They demonstrate that the utilisation of skills has a beneficial effect in both professional and domestic settings. Moreover, their research showcases how problem solving skills are often associated with basic background information, cultural capital defined by the number of books kept at home, and skill usage whether it is in a personal setting or a work setting.

In looking at the issue of skills and job mismatches in the labour market within the context of an Asian region. Umurzakova (2021) reveals that the influence of education on work satisfaction is relatively smaller compared to the significance of workplace skills. On the other hand, the combination of these two factors leads to

incomes that are lower. In the case of over qualified workers however, there is a reduction in productivity the longer they are employed, most likely due to the lack of opportunities available to enhance their skillset. The results of this suggest that those that are under qualified may eventually progress into the modal average of their occupation overtime. It is important to note however that these factors vary depending on the sector of occupation.

Lee and Wei (2017) examine the likelihood of being employed in Japan and South Korea. The authors analysed individuals aged 25 to 55, irrespective of their employment situation. Their research demonstrates that an increase in years of formal education and informal learning has a substantial effect on the likelihood of employment for people in both nations, however the influence is more pronounced in Japan compared to South Korea. Nonetheless, they regarded a comprehensive focus on enhancing skill development as a more advantageous investment for workers, highlighting that their education and training systems have room for improvement to be able to keep up with the demands of their current and future needs for skill usage.

Table 2.2 Tabulated Analysis of PIAAC in Asian Regions Concerning Work

Year	Region	Title	Author(s)	Scope	Method
2022	South Korea	Can Older Workers Stay Productive? The Role of ICT Skills and Training	Jong-Wha Lee, Do Won Kwak, Eunbi Song	Workers	Logistic Regression
2021	Kazakhstan	Effects of Skills Mismatches on Job Satisfaction in Kazakhstan: Evidence from PIAAC Data	Tolganay Umurzakova	Employed Workers	Probit Model

2020	South Korea	A Latent Class Analysis of Older Workers' Skill Proficiency and Skill Utilization in South Korea	Doo Hun Lim, Hyunok Ryu, Bora Jin	Older Workers	Latent Class Analysis (LCA) Mplus version 7.4
2020	South Korea	An Analysis of the Factors on the Problem-Solving Competencies of Engineering Employees in Korea	Jiyoung Yoon, Eun Jung Hur, Minsun Kim	Engineering Employees	Random Effect Multi-level Model (MLM) Stata MP 15.0
2020	Japan and South Korea	Factors Affecting Employees' Problem-Solving Skills in Technology-Rich Environments in Japan and Korea	Chyul-Young Jyung, Yoowoo Lee, Sunyoung Park, Eunhye Cho, Romi Choi	Workers	Item Response Theory, IDB Analyser, SPSS 23.0
2019	South Korea	The Effects and Challenges of Vocational Training in Korea	Jong-Wha Lee, Jong-Suk Han, Eunbi Son	VET Workers	Regression model
2017	Japan and South Korea	Returns to Education and Skills in the Labor Market: Evidence	Jong-Wha Lee, Dainn Wie	Respondents over 25 years old with at least college level education	Decomposition Analysis

		from Japan and Korea			

Chapter 3 Theoretical Framework

Chapter 3.1 The PIAAC Data

The PIAAC data collection comprises three assessment cycles. As of writing, the PIAAC has completed the first cycle of data collection which took place between 2011 to 2017. The second cycle is currently underway and is expected to take place between 2024 to 2029. The survey evaluates participants aged 16 to 65 in various countries through either traditional pen and paper methods or digitally using a computer. (OECD, 2019)

In total, 37 countries took part in the first cycle. Australia, Austria, Flanders in Belgium, Canada, Czechia, Denmark, Estonia, Finland, France, Germany, Ireland, Italy, Japan, Korea, Netherlands, Norway, Poland, Russian Federation, Slovak Republic, Spain, Sweden, United Kingdom, that is England and Northern Ireland, and the United States participated in the first round between August 2011 to March 2012. In the second round, Chile, Greece, Indonesia, Israel, Lithuania, New Zealand, Singapore, Slovenia, and Turkey participated between April 2014 to March 2015. The last round of the cycle took place between July to December of 2017 and was participated by Ecuador, Hungary, Kazakhstan, Mexico, Peru, and the United States.

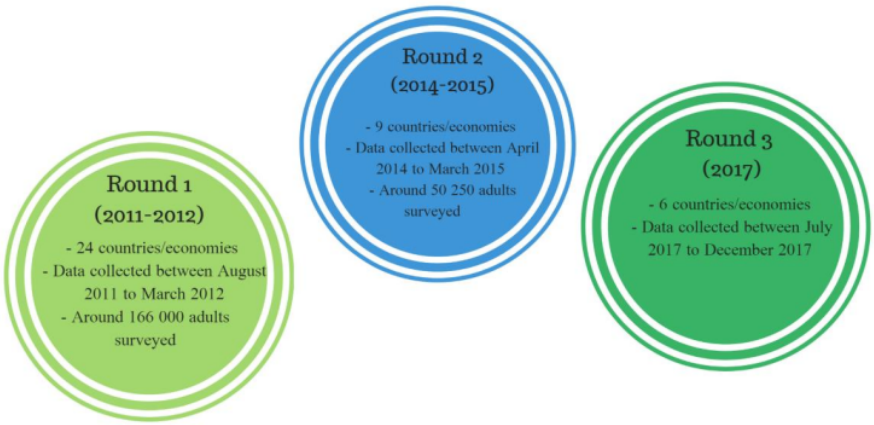


Figure 3.1 from (OECD, n.d.)

In this research, the countries that were analysed are the four Asian countries that are available in the dataset. The countries mentioned are South Korea, Singapore,

Japan, and Kazakhstan. Although Indonesia is an Asian country, their method of data collection was only via the use of pen and paper due to a majority of the population not being familiar with the use of a computer and thus, did not partake in the PSTRE assessment. Despite Turkey and Israel being geographically located within West Asia, they are considered Middle East. Alongside that, the current ongoing political imbalance between those countries may affect the results of this study and thus, are not included.

The combined number of participants from these countries amounted to 23,463. It should be noted that the Singapore dataset does not have an adequate amount of data for the ages of its participants. Hence, the variable of age will not be taken into account in this study.

Chapter 3.1.1 Questionnaire

The PIAAC data is an extensive repository of information, offering a diverse range of inquiries to enhance comprehension of an individual. The questionnaire is divided into ten distinct sections labelled with the letters A to J in alphabetical order. The various categories encompass demographic data, including age, gender, education, training, current status, and work history. Additionally, it includes information about participants' current and previous jobs, the skills utilised in their work, the application of three cognitive skills in both work and home settings, questions concerning themselves such as learning strategies and social trust, and background details such as parental education status.

The features utilised for this study were obtained from Hämäläinen, et al. (2017), where the authors categorised the available features into four distinct subgroups. Those subgroups are demographic characteristics, work and education, work skill use and learning, and everyday life skill use and learning. These subgroups are further refined to only include features that take into account problem solving or ICT questionnaires. This aids in generating a more complete picture of the participants, adding on more possibilities to comprehend and analyse skill sets within the context of other factors that could possibly influence it.

The demographic characteristics subgroup contains features that paint the background of a participant. Features from this subgroup include their age, the country of which the participant took the assessment in and whether or not they were born in that country, the languages spoken at home, and the highest level of education for the participants' parents or guardians.

Table 3.1 Table of demographic characteristic variables

Demographic Characteristics	
Features	**Description**
AGE_R	Age of the participant
CNTRYID	Country of the participant taking the assessment
GENDER_R	Gender of the participant
I_Q08	Self-reported health state of the participant
J_Q04a	Whether or not the participant was born in the country taking the assessment
J_Q06b	The highest education level of participants' mother/female guardian
J_Q07b	The highest education level of participants' father/male guardian
LNG_HOME	Language spoken at home
NATIVELANG	Whether the assessment language is the same as the participants' native language
NATIVESPEAKER	Whether or not the participant is a native speaker of the country's language

In the work and education subgroup, we consider the variables that indicate the highest level of education the participant has, along with their area of study. We also consider if the participants' formal education is job related and whether or not it is useful for their job. Besides that, we dive into the participants' work features which includes questions such as their employment status, the type of work they have, and hours of work per week. The full list of features from this subgroup is in table 3.2.

Table 3.2 Table of work and education variables

Work and Education	
Feature	**Definition**
B_Q01a	The highest level of qualification possessed
B_Q01a3	The highest level of foreign qualification possessed
B_Q01b	The area of study for the participants' highest qualification
B_Q05c	Whether or not the participants' formal education is job related
B_Q10c	Whether or not the participants' formal education is useful for their job
C_D06	The type of job the participant current has, whether it is a paid job or a family business
C_Q07	Subjective status of the participants' work situation
D_Q03	Participants' occupations' economic sector
D_Q04	Whether or not the participant is employed or self-employed
D_Q07a	Whether or not employees work for the participant
D_Q07b	Number of employees working for the participant
D_Q08a	Whether or not the participant manages other employees
D_Q08b	Number of employees the participant is managing
D_Q10	Current hours at work per week
D_Q12a	The participants' current work education requirement
D_Q14	Self reported job satisfaction
ISCO1C	Classification of participants' job based on 1-digit level from ISCO 2008
ISCO2C	Classification of participants' job based on 2-digit level from ISCO 2008
ISCOSKIL4	Classification of participants' job based on 4 skill based job categories
NFEHRS	Number of hours participating in informal education

VET	Whether or not the participants' highest level of education is vocationally oriented

For the work skill use and learning subgroup, we consider the general skill use at work, the ICT skill use at work, the index of ICT usage at work, and the informal learning hours for job-related reasons. The 'F' category in the assessment pertains to questions about general skill use at work. The 'G' category contains questions of skill usage at work. The questions within this category are split between all three of the cognitive skill domains. For this research, only questions that are relevant towards ICT will be taken into account. Other variables include 'ICTWORK' that denotes the index of ICT usage at work, and 'NFEHRSJR' that denotes the hours of informal learning for job-related reasons.

Table 3.3 Table of work and skill use variables

Work Skill Use and Learning	
Feature	**Description**
F_Q01b	Frequency of the participant cooperating with co-workers
F_Q02a	Frequency of the participant sharing work-related info
F_Q02b	Frequency of the participant teaching people
F_Q02c	Frequency of the participant conducting presentations
F_Q02d	Frequency of the participant participating in sales
F_Q02e	Frequency of the participant advising people
F_Q03a	Frequency of the participant planning their own activities
F_Q03b	Frequency of the participant planning activities
F_Q03c	Frequency of the participant organising their own time
F_Q04a	Frequency of the participant influencing other people
F_Q04b	Frequency of the participant negotiating with people
F_Q05a	Frequency of the participant solving simple problems

F_Q05b	Frequency of the participant solving complex problems
F_Q06b	Frequency of the participant working physically for long hours
F_Q06c	Frequency of the participant using hands or fingers
F_Q07a	Whether or not the participant feels challenged at work
F_Q07b	Whether or not the participant feels that more training is required
G_Q04	Whether or not the participant has an experience with computers during work
G_Q05a	Frequency of the participant using the internet for emails
G_Q05c	Frequency of the participant using the internet to search for work related information
G_Q05d	Frequency of the participant conducting transactions via the internet
G_Q05e	Frequency of the participant producing spreadsheets via the use of a computer
G_Q05f	Frequency of the participant using Word on the computer
G_Q05g	Frequency of the participant using programming languages
G_Q05h	Frequency of the participant having real-time discussions via the use of a computer
G_Q06	Self reported level of computer use
G_Q07	Whether or not the participant has the skills for using a computer
G_Q08	Whether or not a lack of computer skills affects the participants' career
ICTWORK	ICT skill use at work index
NFEHRSJR	Number of hours participating in informal education for job-related reasons

In the everyday skill use and learning subgroup, we consider the skill usage of ICT in everyday life which can be found in the 'H' category. For learning strategies, the 'I' category of the questionnaire pertains to questions regarding the participants' own selves. Here, we use only those that ask about the participants' learning strategies. Besides that, the ICT usage index at home which is under the 'ICTHOME' feature will

be used as well as 'NFEHRSNJR' which denotes the informal learning hours outside of job reasonings.

Table 3.4 Table of everyday skill use and learning variables

Everyday Skill Use and Learning	
Feature	**Description**
H_Q04a	Whether or not the participant has ever used a computer
H_Q04b	Whether or not the participant has experience with using a computer in their everyday life
H_Q05a	Frequency of the participant using the internet for emails
H_Q05c	Frequency of the participant using the internet to better understand various issues
H_Q05d	Frequency of the participant conducting transactions via the internet
H_Q05e	Frequency of the participant producing spreadsheets via the use of a computer
H_Q05f	Frequency of the participant using Word on the computer
H_Q05g	Frequency of the participant using programming languages
H_Q05h	Frequency of the participant having real-time discussions via the use of a computer
I_Q04b	The extent of which the participant relates new ideas into real life
I_Q04d	The extent of which the participant likes learning new things
I_Q04h	The extent of which the participant attributes something new
I_Q04j	The extent of which the participant gets to the bottom of difficult things
I_Q04l	The extent of which the participant configures how different ideas fit together
I_Q04m	The extent of which the participant looks for additional info as a method of learning
ICTHOME	ICT skill use at home index

NFEHRSNJR	Number of hours participating in informal education for non-job-related reasons

In order to maintain the dataset's relevance to our research, we have excluded participants who were unemployed from the dataset. In addition, the PIAAC data is presented in an encoded format. Therefore, the codebook was utilised to decode each value, ensuring that the data is easily readable and maximally effective for this research.

Chapter 3.1.2 Skill Domains

The OECD (2019) asserts that the PIAAC assessment examines three cognitive skills domains: literacy, numeracy, and PSTRE. Literacy is a cognitive ability that involves understanding and interpreting written information in various forms and styles, and being able to respond to it appropriately. Numeracy is a skill that involves not just the ability to solve mathematical problems, but to contend with an array of possible representations of numerical issues, not limited to merely arithmetic knowledge and computation.

As for PSTRE, this domain primarily revolves around a specific class of problems dealing with ICT tools. Often considered as 'computer literacy', this skill encompasses the capacity of an adults' use with these tools and applications, as well as the ability to maintain and handle technology-rich environments. Having said that, the main objective of analysing this skill is to assess the capacity of accessibility, process, evaluation and analysis capabilities of adults within the realm of ICT. The cognitive dimensions used to assess PSTRE include goal setting, monitoring progress, planning, accessing and evaluating information, and selecting, organising, and transforming information. Environments such as web, spreadsheet, and email environments were used to influence the difficulty in performing each of these tasks.

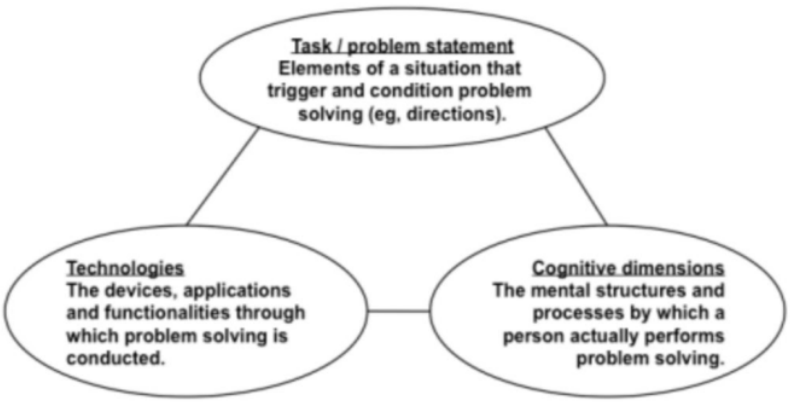


Figure 3.2 PSTRE from OECD (2019)

In the case of literacy and numeracy, all participants successfully completed both assessments. For PSTRE however, participants who do not have access to or have no experience with using a computer were unable to participate in this assessment.

Table 3.5 Definition of each PIAAC skill domain

Skill	Definition
Literacy	Understanding, evaluating, using and engaging with written texts to participate in society, to achieve one's goals, and to develop one's knowledge and potential
Numeracy	The ability to access, use, interpret and communicate mathematical information and ideas, in order to engage in and manage the mathematical demands of a range of situations in adult life
Problem Solving in Technology Rich Environment (PSTRE)	The ability to solve problems for personal, work and civic purposes by setting up appropriate goals and plans, and accessing and making use of information through computers and computer networks

Hämäläinen, et al. (2017) asserts that PSTRE can be divided into four proficiency levels that determine the knowledge and skills needed to accomplish tasks at each level. The levels are derived from the problem solving scale score using plausible values, indicated as PVPSL1 to PVPSL10. The lowest level includes scores from 241 to 290 points, the second lowest includes scores from 291 to 340, and the

highest level includes a scoring that is from 341 and above. Each of these are labelled as 'weak performers', 'moderate performers', and 'strong performers' respectively. Those that obtain a scoring below the first level are considered 'at risk'.

In order to optimise the dataset's performance to answer the problems in this research, the dataset is refined to include only those that have their problem-solving skills variable filled out. This is because no imputation will be involved to keep the data as accurate as possible. Those who do not have problem solving skill scores are participants who do not have access to computers or do not have any experience with modern technology and thus are not relevant in our research. Along with these features, the tenth problem solving variable, PVPSL10, will be used. Hämäläinen, et al. (2017) states that this variable is sufficient to represent all 10 problem solving variables. An additional variable that dictates the performance level of the participant based on their PVPSL10 score will also be included.

Chapter 3.2 Work Environment of Asian Countries

Chapter 3.2.1 Skill Mismatch

Skill mismatch refers to the type of imbalance between the skills possessed by a worker and the skills required to perform the demands of their current job. The evaluation method encompasses a wide range of discrepancies, encompassing both qualitative and quantitative aspects, as well as formal and informal education. Umurzakova (2021) stresses how this imbalance forms challenges at different stages of a worker's career. Consequently, the labour force is affected by a decrease in the number of workers who are no longer able to meet the requirements of their jobs.

The PIAAC data uses the International Standard Classification of Education (ISCED) 1997 to label each of the participants' education levels. ISCED level ranges from 1 to 6, with subcategories of A, B, or C. As each country has its own unique education system, each of them has been converted into its appropriate ISCED value to ensure uniformity within the PIAAC dataset. (OECD, 2019)

Table 3.6 PIAAC mapping of ISCED

ISCED level	Japan	Kazakhstan	South Korea	Singapore
School starting age	6	5	6	6
No formal qualification or below ISCED 1	No formal education, Dropped out of elementary school	No formal qualification	No formal education, Below elementary	No formal education, Lower primary
ISCED 1	Elementary school Special education school	Primary education	Elementary education	Primary education
ISCED 2	Lower secondary school Lower division of secondary education school, Lower secondary department of special education school	Basic general education	Middle school education	Lower secondary education
ISCED 3C shorter than 2 years	Short term course of upper secondary school/upper division secondary education school special education school			
ISCED 3C 2 years or more 5	Specialised course in upper secondary school/upper		High school vocational education	

	division secondary education school special education college			
ISCED 3A-B	General/Integrated upper secondary school/secondary education school/special education school/, College of technology (1st-3rd year)	Secondary general education	High school college preparation education	
ISCED 3 (without distinction A-B-C, 2 years and above)	Passed upper secondary school equivalency examination	Professional and technical education after grade 9		Upper secondary
ISCED 4C				
ISCED 4A-B		Post-secondary education after grade 11 (mass technical and service professions)		Post-secondary non-tertiary vocational education
				Post secondary non-tertiary academic education
ISCED 4 (without distinction A-B-C)	Advanced course of upper secondary school/upper division of secondary education, Short term			

	course of junior college/university			
ISCED 5B	Regular course of junior college/college of technology, Advanced course of junior college/college of technology, Specialised training college	Post-secondary education after grade 11 (industries associated with high technology and professional activities)	Master's degree for specialised/vocational graduate school	Diploma
			2-3 year college	
			4 year college of education for Bachelor's degree	
ISCED 5A, bachelor degree	Undergraduate programs, Advanced course of university	Higher education, Bachelor/Specialist degree	4 year university for Bachelor's degree	Bachelor's degree
ISCED 5A, master degree	Master's programs, Lower division doctoral programs, Professional degrees programs of university/law school	Higher education, Master/Candidate of Science degree	Master's degree at general universities	Master's degree
	Completed all work for doctoral program except doctoral thesis			

ISCED 6	Doctoral programs	Post-university education, PhD/Doctor of Science	Doctoral degree	Phd/Doctorate
	Specialised training, Miscellaneous school			

The OECD (2019) also provides a table that entails which of these particular ISCED levels are vocational. For Japan, the ISCED levels that are associated with vocational schooling are ISCED 3C shorter than 2 years, ISCED 3C 2 years or more, and ISCED 4 (without distinction A-B-C). In the case of Kazakhstan, those levels are ISCED 3 (without distinction A-B-C, for more than 2 years), ISCED 4A-B, and ISCED 5B. In Korea, ISCED 3C for 2 years or more is the only level that is considered vocational. Lastly, Singapore denotes levels only ISCED 4A-B as vocational. The reason for this distinction is to consider whether those that obtained a vocational qualification have different requirements compared to those that do not.

In order to conduct our research, we will categorise the skill mismatch based on the definition used by Pivovarova et al. (2022). An overmatch is defined as possessing credentials that exceed the average expected credentials in their occupational sector, whereas an undermatch is defined as possessing credentials lower than the expected average of their occupational sector. Individuals whose educational qualifications align with the job requirements will be regarded as 'equal. The study will take the individuals' highest level of education and the qualifications necessary for their job to define them. Those with a foreign level of education will be converted to the relevant ISCED levels.

Chapter 3.2.2 Working Hours Across Asian Countries

This research will consider the relevant legislation on working hours in each country to accurately determine the appropriate working hours. Normal working hours refer to the standard hours of work for an employee, excluding any instances where

the employee has agreed to work additional hours with their employer. The maximum possible working hours will be considered as working hours that include those factors.

According to the Labor Standards Act (1947), article 32 in chapter 4 (Working Hours, Breaks, Days Off, and Annual Paid Leave) states that an employee's maximum working hours must not exceed 40 hours per week. However, these working hours may extend up to 42 hours for businesses that are smaller in size.

According to Article 50 in Chapter 4 (Work Hours and Recess) of South Korean labour laws, the weekly maximum limit for work hours is set at 40 hours. The Labor Standards Act (2012) also stipulates that the time spent under the guidance or oversight of supervisors is also classified as working hours. However, according to article 52 and article 53, the working hours can be increased to a maximum of 48 hours and 52 hours, respectively, if mutual consent is present between the employer and the employee. It should be noted that the aforementioned articles only apply to workers above the age of 18. (Labor Standards Act, 2007)

As per Article 77 of Chapter 7 (Working Hours) of the Labour Code of the Republic of Kazakhstan (2007), the weekly working hours should not exceed 40 hours. Article 89 states that a maximum of two hours of additional work beyond regular working hours is allowed, as long as the total amount of overtime work does not exceed 12 hours per month or 120 hours per year. Each of these working hours may vary in accordance with the employee's respective occupation and any contractual agreement that has been made between them and their employer. Article 78 notes that there is a reduction in working hours for those under 18 but does not specify the maximum hours for people of this age group.

In Singapore, article 38 in part 4 (Rest Day, Hours of Work and Other Condition of Service) of the Employment Act (1968) states employees can work at most 48 hours per week, if they do not exceed 88 hours within 2 weeks continuously. This case is only applicable when there is mutual consent between the employer and

the employee. Under normal circumstances, employees are required to work a maximum of 44 hours per week.

For this research, the range that will be used as the average working hours will differentiate between countries. In Japan, the average range falls between 40 and 42. In the case of Singapore, the average range will be between 44 and 48. Both South Korea and Kazakhstan have the same average range of 40 to 52. Individuals who work less hours per week than the national average will be categorised as 'below average', while those who work more hours per week than the national average would be categorised as 'above average'.

Table 3.7 Tabulated normal and maximum possible working hours per week

Country	Japan	South Korea	Kazakhstan	Singapore
Normal working hours per week	40	40	40	44
Maximum possible hours per week	42	52	52	48

Chapter 3.3 Machine Learning Models

Three machine learning models will be used and finetuned for the purpose of this project. Those models are random forest classifier, decision tree classifier, and SVM. Each of these models will be finetuned to achieve the best possible outcome for this project.

The random forest classifier is an algorithm that combines multiple decision trees into a singular ensemble. It grows each decision tree by feeding it random sampling subsets. Any data not included is used for validation. As each tree grows, the nodes splits into a subset of predictor variables and stops when a predefined number of leaf nodes or impurity threshold has been achieved. The output is determined by aggregating the results from each individual tree, thus improving the accuracy and reducing the chances of overfitting compared to using a singular decision tree classifier. (He, et al. 2023)

The next model is a decision tree. A decision tree is as the name implies, a tree of which consists of branches, nodes, and leaves that aid in both regression and classification tasks. It is the tree that makes up the ensemble model, random forest. Similar to the random forest model, it has three criterions for splitting. Namely the gini impurity which measures the likelihood of incorrect classification of a randomly chosen element from the data, entropy which measures the amount of uncertainty of nodes, and log loss or logarithmic loss which measures the uncertainty of predictions. While both entropy and log loss measure similar things, entropy guides the decision tree for more purer subsets, while the log loss penalises incorrect predictions for a better overall probability prediction. (Datamapu, 2023)

The last machine learning model is an SVM. In a SVM, kernels are considered optimal boundaries that classifies the dataset in different regions. A kernel's function is to transform the inputted data into its' required form. Some of these kernels include linear kernels such as the linear kernel and nonlinear kernels such as the polynomial kernel, Gaussian radial basis function (RBF) kernel, and the sigmoid kernel. (He, et al. 2023)

Chapter 3.4 Association Rule Analysis

Colloquially called 'market basket analysis', association rule mining will be implemented in this research to uncover factors of work mismatch between a workers' highest qualification and work qualification. Association rule mining functions by sorting item sets within a dataset and unveiling associations between each distinct element. Key metrics of support, confidence, and lift are used to define and identify these associations.

The associations produced are referred to as 'rules', which consist of itemsets representing both the antecedent (events that occur) and the consequent (outcome). A support is defined as the frequency of an itemset occurring, while confidence is defined as the probability of a consequent having existing antecedents. Finally, there is lift. Mathematically, it is defined as the likelihood ratio of having an additional element

given the presence of another element that may be associated with it. Lifts can introduce the strength of an association between two elements. It refines the association rules even further by preventing misinterpretation of rules with a high confidence metric but low association. Lifts with a value greater than 1 indicate a positive association, while lesser than 1 indicates a negative association. (Garg, 2019)

$$Lift(\{X\} \rightarrow \{Y\}) = \frac{(Transactions\ containing\ both\ X\ and\ Y)/(Transactions\ containing\ X)}{Fraction\ of\ transactions\ containing\ Y}$$

Figure 3.3 Formula for lift by Garg (2019)

The association rule mining is applied on to the dataset to find common factors that are associated with a mismatch in educational qualification and work requirements, alongside with identifying how under qualified workers can mitigate and retain their current employment.

Chapter 4 Research Methodology

Deciphering the dataset and finding the patterns to solve the research objectives involves understanding the codebook provided by the OECD. In the codebook, there are columns that denote the sequence of the item in the dataset, the name of the item, the label which dictates the assessment question, the type of data (be it an integer or a string), the level which can be nominal, ratio, scale, or ordinal, the missing scheme indicator, and a link to a separate spreadsheet that entails all the true meanings behind each encoded answer.

Chapter 4.1 Preliminary Exploratory Analysis of the PIAAC Data

A preliminary exploration of the dataset was conducted to have an initial look at the data that will be further analysed. The histogram displays that for the country of Kazakhstan, the count of age peaks at around 30, while the age in Japan peaks near the 40th mark. For Korea however, several peaks in ages are found in the dataset, those peaks can be found near 20th, 30th, and the highest being by the 40th mark in age.

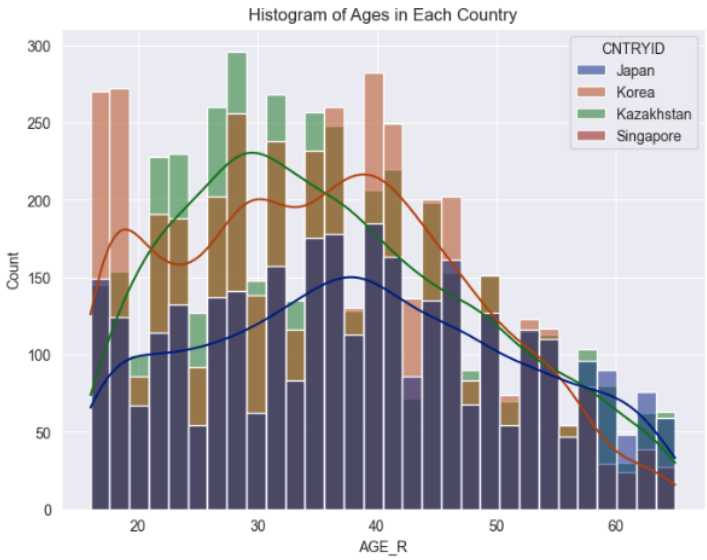


Figure 4.1 Histogram of ages in each country

Seeing as how the age varies across the dataset, a closer inspection of it against the hours of informal learning showcases how most of the participants are between 0 to 500 hours of informal learning regardless of age. This case is most true for participants from Kazakhstan. For the case of Korea however, the informal learning hours vary regardless of age, with some having almost 2000, while others are spread across 0 to 2000 informal learning hours. In Japan, the informal learning hours are not as high as other countries, with a smaller amount being above 500 hours when compared to Korea and Kazakhstan.

Figure 4.2 Pairplot of age and informal learning

A countplot of the participants' subjective status of their current work or work history shows that a majority of them are full time workers, with 1943 from Japan, 2489 from South Korea, 2607 from Kazakhstan, and 2612 from Singapore. Meanwhile those who are part-time workers have a count of 490, 596, 560, and 204 from Japan,

South Korea, Kazakhstan, and Singapore respectively. The dataset also features students, having 389 from Japan, 747 from South Korea, 394 from Kazakhstan, and 606 from Singapore. Only 3 participants in total did not state their current work status. Other variables include those who are fulfilling domestic tasks or looking after family members with a total count of 1679, 642 unemployed participants, 320 in retirement or early retirement, an undefined other category with a count 253, interns with a count of 69, permanently disabled participants with a count of 48, and those in compulsory military or community service with a count of 124.

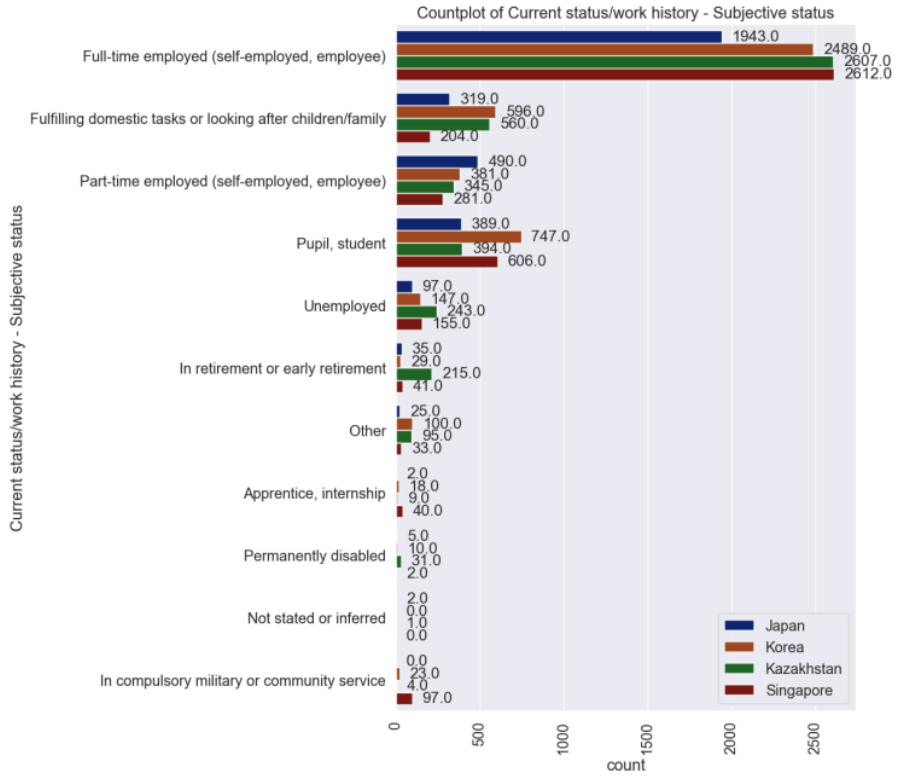


Figure 4.3 Countplot of subjective status for current work status/history

Chapter 4.1.1 Age and Working Hours

To illustrate the normal working hours, we analyse the participants' working hours and categorise them by nation. Boxplots are defined by their quartiles, which are represented by the median, first quartile (q1), and third quartile (q3). In Japan, the

median value is 40, with a q1 of 35 and a q3 of 50. The median in South Korea is 44, with a q1 of 40 and a q3 of 50. Kazakhstan's median is 40, which is equivalent to the q1, while the q3 is 48. The median value in Singapore is 44, with the q1 at 40 and the q3 at 50. Several outliers for working hours can be seen, particularly those that exceed 100 working hours per week can be found in Japan and Singapore.

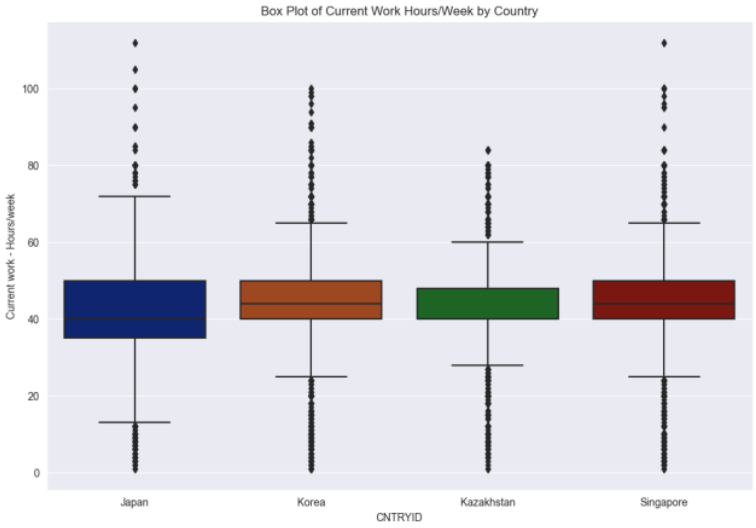


Figure 4.4 Boxplot of current work hours per week by each country

The average working hours in Japan varies across the ages of below 20 up to the age of 65. According to the PIAAC, we can see that the average number of working hours consistently rises until an individual reaches between the age range of 20 to 30. From that point onwards, the average number of hours worked fluctuates between about 50 hours per week and fewer than 40 hours per week until the age of 50. From this point, the trend begins to decline, with a sudden increase in average working hours beyond the age of 60 before ultimately declining.

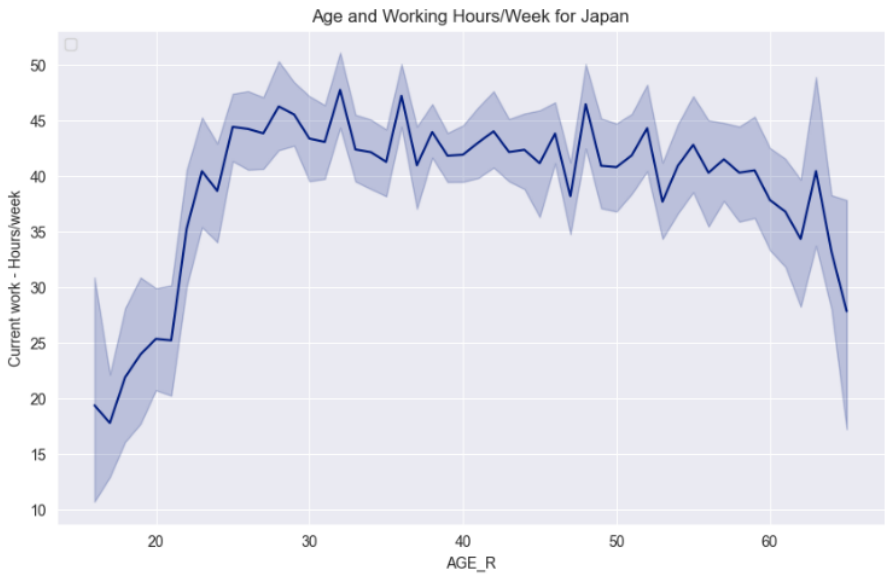


Figure 4.5 Lineplot of age and working hours per week in Japan

In South Korea, the number of working hours gradually rises until an individual reaches the age of 20. Subsequently, the weekly working hours increase before continuously remaining between 40 and 50 hours each week. After reaching the age of 50, there is a noticeable decline in working hours, followed by two periods of increased working hours after the age approaches 60. Those increases are 50 hours and 60 hours. Subsequently, the trend declines to nearly an average of 30 hours before subsequently rebounding.

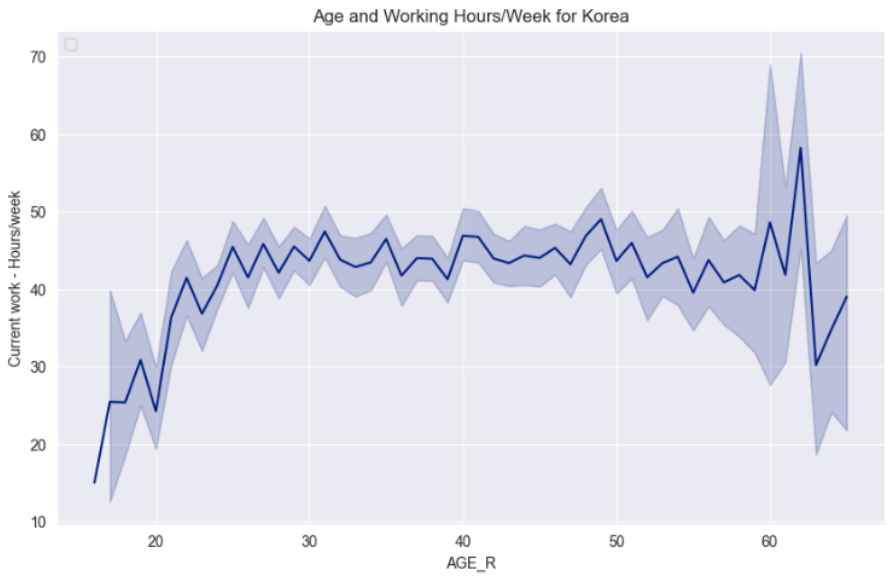


Figure 4.6 Lineplot of age and working hours per week in South Korea

At the age of 20, the working hours in Kazakhstan rise to approximately 40 hours each week. The pattern remains constant within the range of 40 to 50 hours from the ages of 20 to 60. After reaching the age of 60, there is a noticeable and significant decline in the trend, with the minimum number of working hours per week dropping to about 30. Multiple upward spikes are shown before ultimately rising.

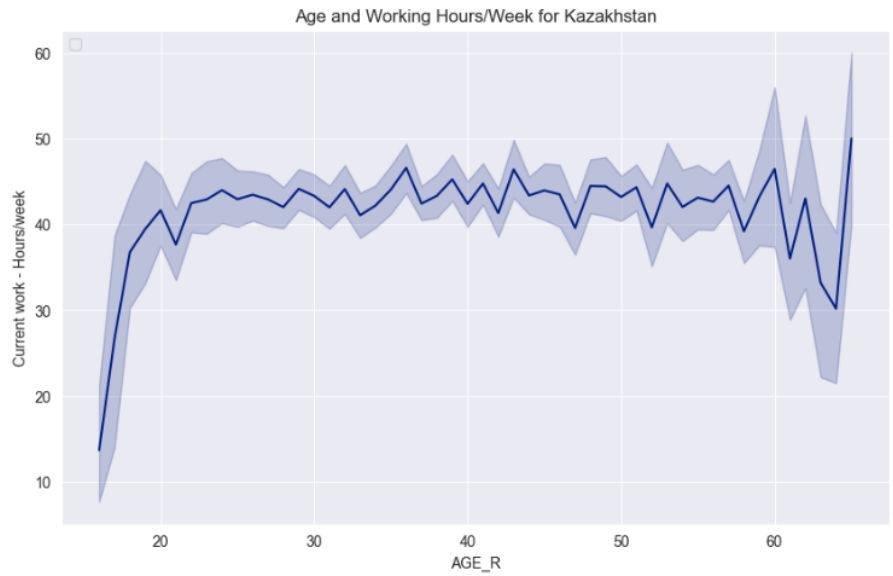


Figure 4.7 Lineplot of age and working hours per week in Kazakhstan

Chapter 4.1.2 Skill Use and PVPSL Performance

The pairplot of index of skill use at home and informal learning for non-job related reasons reveals that the majority of the index of skill use falls within the range of 0 to 6. Informal learning hours in countries typically range from 0 to 500, while the skill use index ranges from 0 to 4. The remaining data remained sparsely distributed up until 1500 hours, with just a small portion falling above that threshold until 2000.

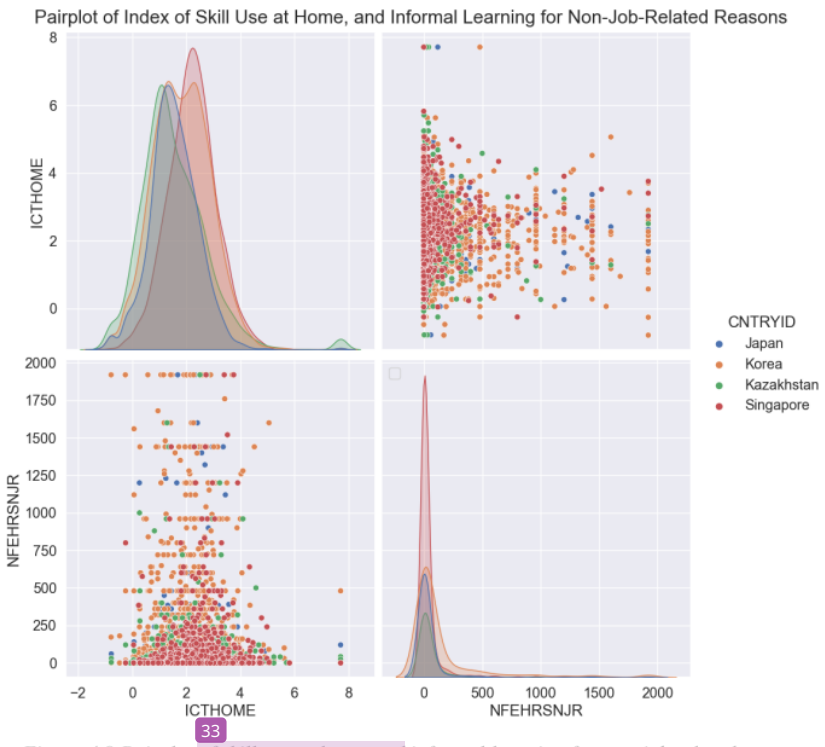
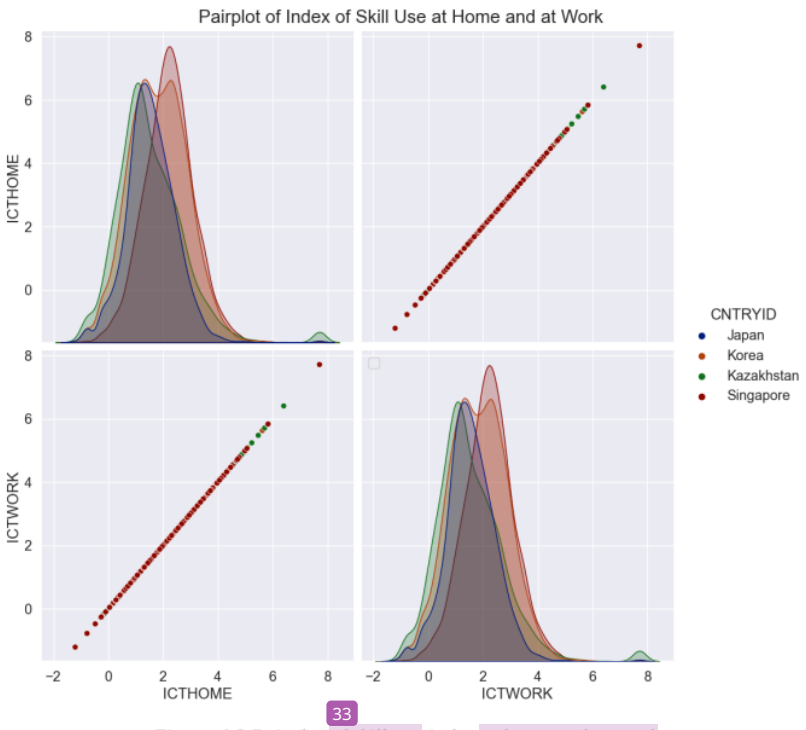


Figure 4.8 Pairplot of skill use at home and informal learning for non-job related reasons

The pairplot of index skill use at home and at work showcases a linear relationship between the two features, displaying a linear line that entails the skill use at home and at work are correlated to one another.



33

Figure 4.9 Pairplot of skill use index at home and at work

The problem solving performance in all countries indicates that the number of individuals with strong performance is below 500 in each country. Japan has the highest count in this category with a total of 469, followed by Singapore with 423. South Korea has a count of 231, while Kazakhstan has the lowest count which is 52. The number of moderate performers individuals from South Korea and Singapore surpasses 1500, with South Korea having a count of 1741 and Singapore with 1648. Japan is currently sporting a count of 1414, while Kazakhstan has the lowest count which is 883. Kazakhstan has the greatest count of weak performers, with a count of 2446, while South Korea has almost 1922. Singapore and Japan fall within the range of 1000 to 1500, with Singapore having 1356 and Japan having 1065. The category labelled as 'at risk' is regarded as the class with the poorest performance. Kazakhstan holds the largest number in this category, surpassing 1000 with 1123 participants. South Korea and Singapore have numbers of 646 and 644 respectively. On the other

hand, Japan has the lowest number of performers considered 'at risk', with a total of 359.

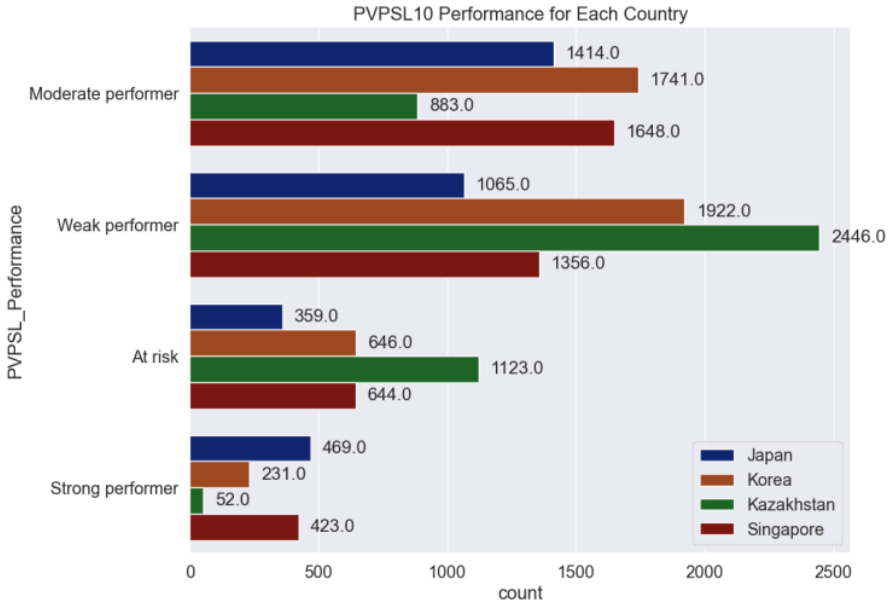


Figure 4.10 Countplot of PVPSL performance for each country

The histogram of performance scores reveals that a significant proportion of individuals with moderate performance from all four Asian countries tend to cluster around the scoring range of 200 to 300. Weak performers on the other hand in Japan and South Korea have a high number of participants scoring between the range of 280 to 290, while Kazakhstan has the highest peak between the scoring range of 260 to 270. Singapore on the other hand, has most of its weak performers scoring at around 270. For those who are at risk, the histogram appears left-skewed, with the participants from this category mostly comprising those scoring between 220 and 240. Strong performers on the contrary exhibit a right-skewed distribution, with the majority of scores falling within the range of 340 to 360.

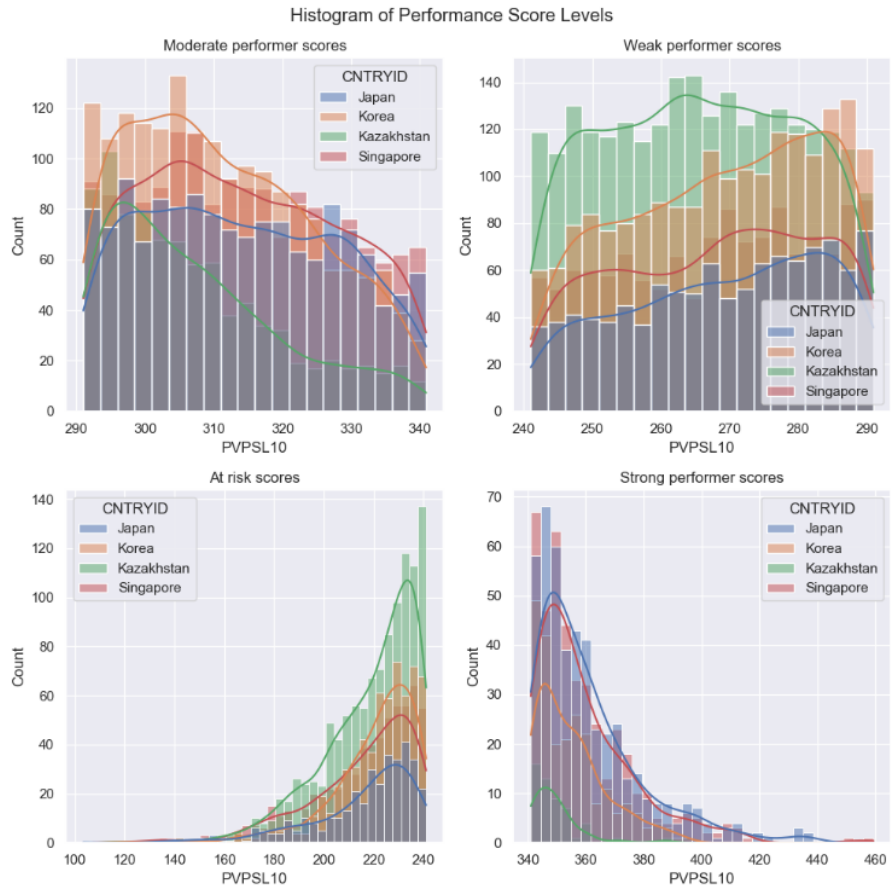


Figure 4.11 Histogram of PVPSL performance scores by levels

Chapter 4.1.3 Education and Qualification Mismatch

A look at the count of highest education levels in the dataset reveals that most of the participants have a level of ISCED 5A, bachelor's degree, with a count of 980 from Japan, 1151 from South Korea, 1077 from Singapore, and the largest count belongs to Kazakhstan with 1616. Besides that, another notable value is ISCED 5B with 696 from Japan, 903 from South Korea, 273 from Kazakhstan, and 952 from Singapore. Another value that has a high count is ISCED 3A-B, where Japan has 813, South Korea has 1101, and Kazakhstan has 645. For this value, there are none from Singapore. A possibility for this reason is that it's between their lower secondary education (ISCED 2) and upper secondary education (ISCED 3 (without distinction

A-B-C, of 2 years or more)) and thus, may not have an equivalent in the ISCED system. ISCED 3 (without distinction A-B-C, of 2 years or more) shows a high count for those in Kazakhstan and Singapore, with counts of 895 and 888 respectively.

Other values include ISCED 2 with a total of 1271, ISCED 3C of 2 years and more with a total of 986, ISCED 3C of shorter than 2 years with a count of 64, ISCED 4 (without distinction A-B-C) with a total of 59, ISCED 5A, master degree, with a total of 655, a total of 24 foreign qualifications, ISCED 6 with a total of 79, ISCED 1 with a total of 130, and ISCED 4A-B with a total of 1099. The lowest level, that is below ISCED 1 or no formal qualification, has a total of 44, with the majority being from Kazakhstan. A total count of 5 participants did not state their highest education level.

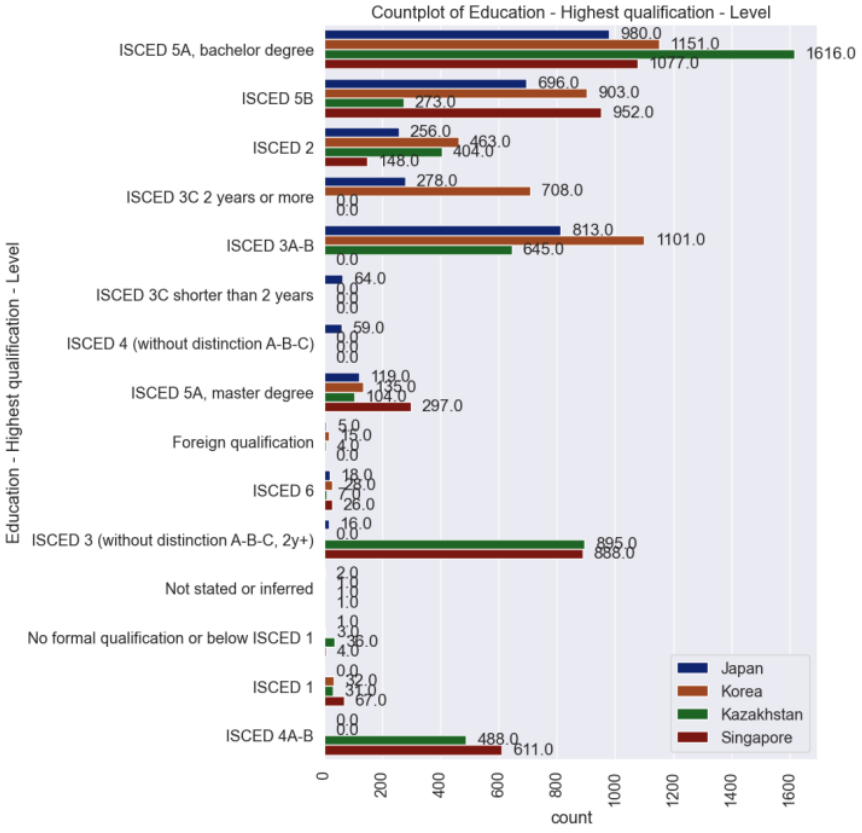


Figure 4.12 Countplot of highest education level

In looking at current work requirements, most of the participants do not state their current work requirements. Most jobs require an education of an ISCED 5A, bachelor degree, with a count of 2971 participants' jobs having this work requirement. Besides that, 1962 of the participants' job requirements is ISCED 5B and 1532 participants' jobs require an education of ISCED 3A-B. The other educational requirements have a count of less than 1000, those being ISCED 3 (without distinction A-B-C, 2 years or more), ISCED 3C for 2 years or more, ISCED 2, ISCED 5A that is a masters' degree, ISCED 1, and ISCED 4A-B. Only three educational requirements are below 100, that is ISCED 4 (without distinction A-B-C), ISCED 3C for shorter than 2 years, and ISCED 6. Jobs that require an education of below ISCED 1 or no formal education has a count of 417.

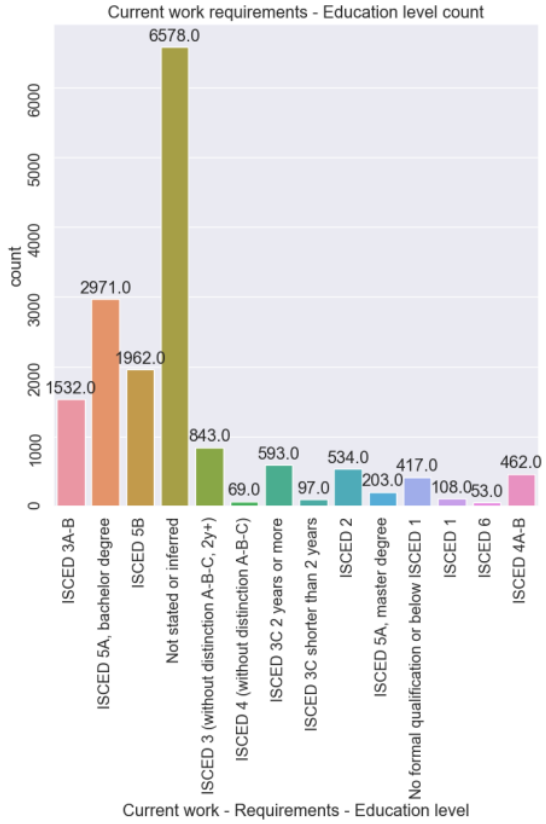


Figure 4.13 Countplot of current work educational requirements

A mismatch in qualification can be seen particularly abundant in those that are over qualified for their job. Typically, this group consists majorly of weak performers with a count of 4139, while moderate performers are the second highest count with 3246. Meanwhile at risk performers have a count of 1767, while the category with the least count, strong performers, has a total of 615.

Those who are of equal status are those under the weak and moderate performers category, with a count of 1950 and 1866 respectively. At risk performers totals up to 704 while strong performers have a count of 461. For under qualified workers, weak performers total up to a count of 700, moderate performers have a count of 574, and at risk performers total up to 301. On the other hand, strong performers have the smallest count in this category, that is a count of 99.

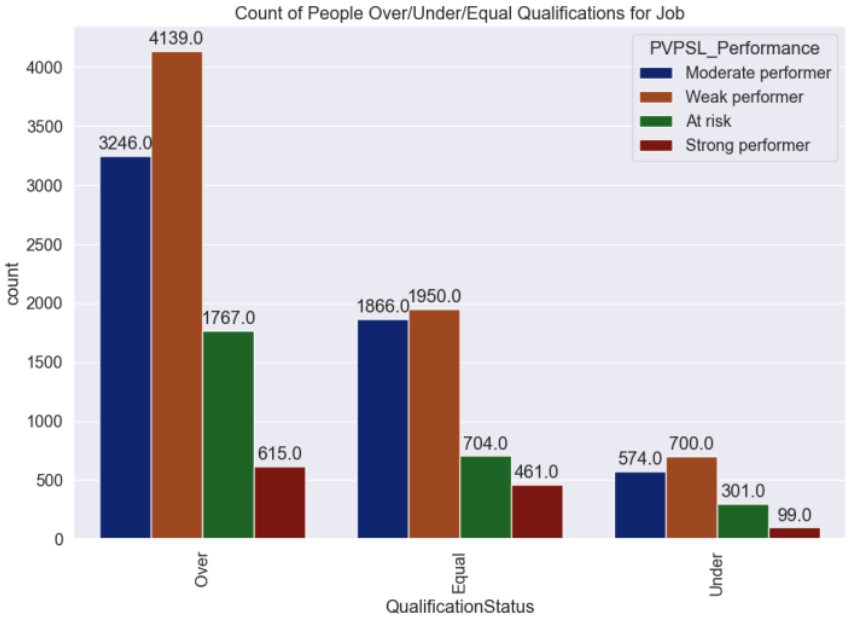


Figure 4.14 Countplot people who are over/under/equal qualified for their job

Chapter 4.2 Investigating the Data

The data was collected from the official PIAAC website. The question of 'why do Asians work so hard?' arises from the initial exploration of the dataset, revealing how the working hours in Asian countries continue to rise even as the workers age. This process is done parallel to conducting the literature review and identifying the research gap. Once these are done, the data will be pre-processed. The next course of action includes exploratory data analysis, modelling the data, and visualisation of the results.

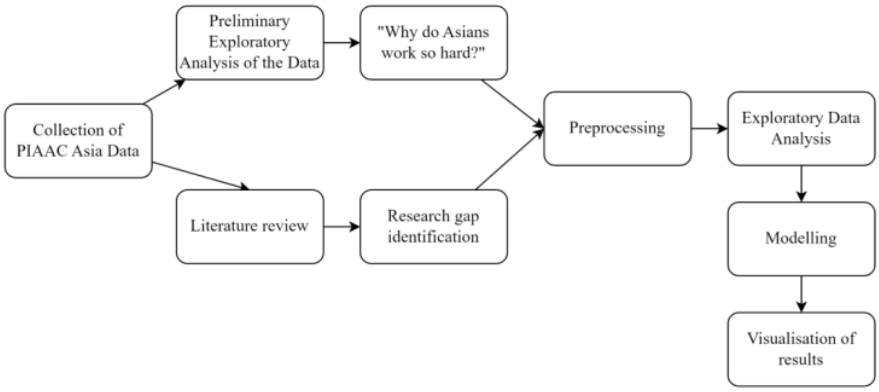


Figure 4.15 FYP1 workflow

The data modelling workflow begins by separating the pre-process data into several smaller datasets. Those who are dictated as working either over and under the average working hours will be separated and fed into three different machine learning models. The subsequent feature importances will then be extracted. Over qualified workers, under qualified workers, and equal qualified workers will be portioned out into their own datasets as well. The association rule mining will then be applied to each dataset.

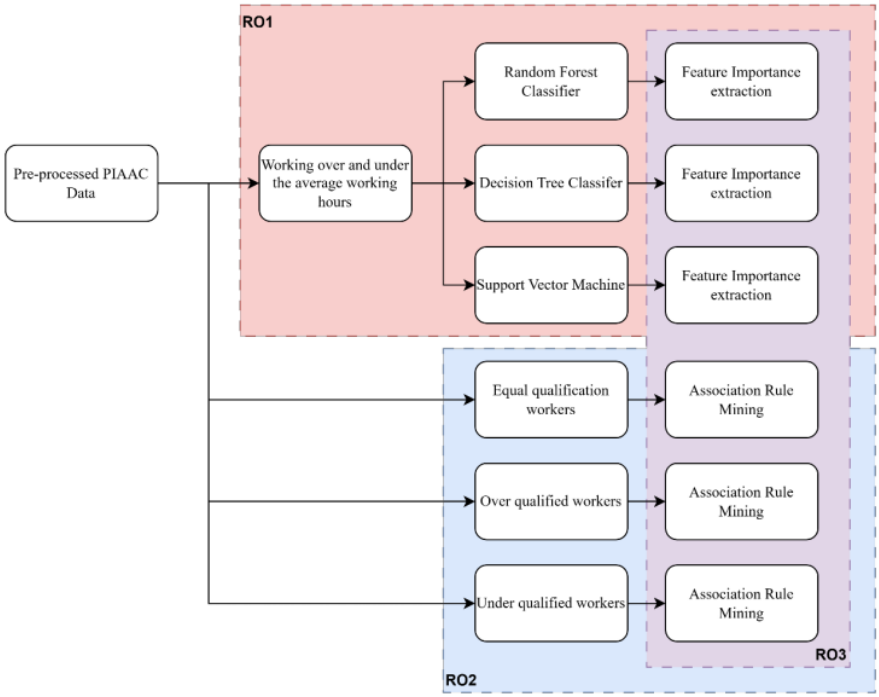


Figure 4.16 Workflow of PIAAC data modelling

Chapter 5 Implementation

Chapter 5.1 Data preparation

PIAAC data from Japan, Kazakhstan, South Korea, and Singapore are undergone preparation for use in this study. Initially, the raw data from each folders combined tallied up to 23,463 rows in total.

The first procedure is cleaning the data by removing all the null values. Initially, the Japan, Kazakhstan, South Korea, and Singapore datasets contained 5278, 6667, 6050, and 5468 rows respectively. If null values were present in any column, except for those concerning the PVPSL, the rows were dropped. This was done using a for loop, iterating through each dataset and using dropna() and drop_duplicates() to clean the dataset of null values and potential duplicates. The only exception were rows that had null values in any of the PVPSL columns, and rows with a valid ID.

```
jpn_drop = jpn_raw.copy()
kor_drop = kor_raw.copy()
kaz_drop = kaz_raw.copy()
sgp_drop = sgp_raw.copy()

# values to keep
tokeep = ['SEQID', 'PVPSL1', 'PVPSL2', 'PVPSL3', 'PVPSL4', 'PVPSL5', 'PVPSL6', 'PVPSL7', 'PVPSL8', 'PVPSL9','PVPSL10']

dfs_drop = {'jpn': jpn_drop, 'kor': kor_drop, 'kaz': kaz_drop, 'sgp': sgp_drop}

for df_name, df in dfs_drop.items():
    df.dropna(subset=tokeep, inplace=True)
    df.drop_duplicates(inplace=True)
```

Figure 5.1 Method of checking null rows

The resulting rows left were 3307, 4540, 4504, and 4071 for Japan, Kazakhstan, South Korea, and Singapore.

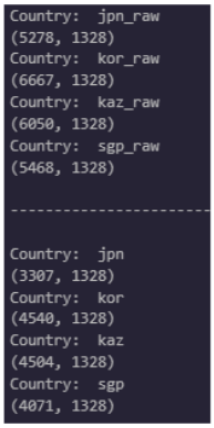


Figure 5.2 Output showing the difference after dropna()

Following that, the null culling continues by tackling any null columns present in the data. Using a for loop to iterate through each column with the functions isna(), fully empty columns were counted and kept in a list. After obtaining a list for all the datasets, the consistently empty column was found via using the set intersection() function and comparing it between all of the lists.

```python
# seeing which  cols are always null
consistent_null_cols = set()
all_null_cols_list = []

for df_name, df in dfs_nulls.items():
    col_all_nan = df.columns[df.isna().all(axis=0)]

    if not col_all_nan.empty:
        # Add the set of columns to the list
        all_null_cols_list.append(set(col_all_nan))

# Find the intersection of sets in the list
if all_null_cols_list:
    consistent_null_cols = set.intersection(*all_null_cols_list)

print(len(consistent_null_cols))
print(consistent_null_cols)


# drop the col
for df_name, df in dfs_nulls.items():
    df.drop(columns=consistent_null_cols, inplace=True)


# finding definition in codebook
for col in consistent_null_cols:
    if col in codebook['Name'].values:
        # Retrieve label and domain
        row = codebook[codebook['Name'] == col].iloc[0]
        print(row)
        print('--------------------------------------------------')
    else:
        print('Column Name: ', col, '(No information in codebook)')
        print('--------------------------------------------------')
```

Figure 5.3 Method of finding null columns

Through this, we find that only 1 column was completely empty for all four countries that is 'D_Q16b_T'. This column was then dropped, resulting in 1327 columns total.

```
1
{'D_Q16b_T'}
Sequence in Dataset                                          162
Name                                                    D_Q16b_T
Label                 Wage or salary [weekly/hourly] before taxes an...
Type                                                    Integer
Level                                                   Ordinal
Width                                                         1
Decimals                                                      0
Domain                                  Background questionnaire (trend)
Value scheme              Trend - Wage, salary or income quintiles (1-5)
Missing Scheme                                          BQ (numeric)
Link to values                                                Values
Name: 161, dtype: object
--------------------------------------------------
```

Figure 5.4 Consistently empty column found

As individual dataset cleaning was completed, the next preparation step involves formatting the data into a more readable format. As the data was encoded to

PIAAC standards, the codebook was used as a guide to decipher the contents of the data.

Chapter 5.1.1 Deciphering the Dataset

To start off, the dataset of each country was combined to create a singular, large dataset. In accordance with the codebook, each column that had a type of 'Integer|Numeric/floating point' was collected and the column was then transformed into a numeric dtype, be it an int64 or a float64, by using Panda's to_numeric() function.

```
# fix the dtype firsst
df_dtype = df_raw.copy()

for col in num_labels:
    if col in df_dtype.columns:
        df_dtype[col] = pd.to_numeric(df_dtype[col], errors='coerce')

display(df_dtype.dtypes)
```

```
CNTRYID              int64
CNTRYID_E            int64
SEQID                int64
AGE_R              float64
GENDER_R             int64
                     ...
TASKDISC_WLE_CA    float64
WRITHOME           float64
WRITHOME_WLE_CA    float64
WRITWORK           float64
WRITWORK_WLE_CA    float64
Length: 1327, dtype: object
```

Figure 5.5 Method of converting columns to numeric

With that done, any value that is not numeric, that is those where participants did not answer or dodged the question in the assessment, were turned into null values. To resolve the null issues created by transforming the numeric columns into its proper dtype format, all the null values were filled with a generic 'N' to denote that it is 'null'. As there were multiple values that are considered 'unanswered', this simplifies the process. The value 0 and -1 were not chosen to fill these roles as it may disrupt the results of the experimentations.

```python
# fill nulls with 'N'
df_fill = df_dtype.fillna('N')

null_columns = df_fill.columns[df_fill.isnull().sum() > 0].tolist()

# Print the null columns
print(len(null_columns))
```

Figure 5.6 Method of filling null columns with a common value

Several columns were found with the ‘N’ value. Each of these ‘N’s were then replaced with ‘Not Available’.

```python
N_dict = {}

for item in columns_with_N:
    if item in codebook_vals_dict:
        for key, value in codebook_vals_dict[item].items():
            if 'Missing' in str(value) or 'Not ' in str(value):
                print(f"Column: {item}, Key: {key}, Value: {value}")
                N_dict[item] = value

len(N_dict)
```

```
Column: PAPER, Key: 0, Value: Missing
Column: CBAMOD1, Key: 0, Value: Missing
Column: CBAMOD2, Key: 0, Value: Missing
Column: CBAMOD2ALT, Key: 0, Value: Missing
Column: CBAMOD1STG1, Key: 0, Value: Missing
Column: CBAMOD2STG1, Key: 0, Value: Missing
Column: CBAMOD1STG2, Key: 0, Value: Missing
Column: CBAMOD2STG2, Key: 0, Value: Missing
Column: ISCO08_C, Key: nan, Value: Not stated or inferred
Column: ISCO08_L, Key: nan, Value: Not stated or inferred
Column: LNG_L1, Key: zxx, Value: No linguistic content; Not applicable
Column: LNG_L1, Key: nan, Value: Not stated or inferred
Column: LNG_L2, Key: zxx, Value: No linguistic content; Not applicable
Column: LNG_L2, Key: nan, Value: Not stated or inferred
Column: LNG_HOME, Key: zxx, Value: No linguistic content; Not applicable
Column: LNG_HOME, Key: nan, Value: Not stated or inferred

13
```

```python
for col, new_val in N_dict.items():
    df_ns[col] = df_ns[col].replace('N', new_val)

df_ns.replace('N', 'Not Available', inplace=True)
```

Figure 5.7 Output of columns that contained the common ‘N’ value

Each column was then transformed into strings to get rid of any trailing '.0' values found when first importing the dataset into the Jupyter notebook. This method also does not disturb any columns that have trailing values that are not 0.

```
df_cb = df_fill.astype(str)

for col in df_cb.columns:
    df_cb[col] = df_cb[col].astype(str).apply(lambda x: x.rstrip('0').rstrip('.') if '.' in x else x)

display(df_cb.head())
```

Figure 5.8 Method of deleting trailing values in string

Once this is done, each column is then iterated twice to map to the decoded value found in the codebook. It involves two iterations because there are two sheets present for decoded values, where one is specific for IBM'S SPSS software.

```
for col, mapping in codebook_vals_dict.items():
    if col in df_cb.columns:
        df_cb[col] = df_cb[col].apply(lambda x: mapping.get(x, x))

for col, mapping in codebook_vals_dict_SPSS.items():
    if col in df_cb.columns:
        df_cb[col] = df_cb[col].apply(lambda x: mapping.get(x, x))

df_cb.head()
```

	CNTRYID	CNTRYID_E	SEQID	AGE_R	GENDER_R	DISP_CIBQ	DISP_MAIN	DISP_MAINWRC
1	Japan	Japan	2	48	Male	Complete	Complete	Not stated or inferred
2	Japan	Japan	3	47	Male	Complete	Complete	Not stated or inferred
3	Japan	Japan	4	58	Male	Complete	Complete	Not stated or inferred
4	Japan	Japan	5	29	Female	Complete	Complete	Not stated or inferred
7	Japan	Japan	8	48	Female	Complete	Complete	Not stated or inferred

5 rows × 1327 columns

Figure 5.9 Output of decoded data

Chapter 5.1.2 Refining the Dataset

Selected features by Hämäläinen, et al. (2017) were kept in the dataset while others were removed, leaving only 88 columns in total.

As the study's primary focus is concerning those who are working, workers were extracted from the dataset based on the column 'Current status/work history - Current - Paid job or family business (DERIVED BY CAPI)'. Responses that had a 'Yes' in their answer were considered workers, while the others will not be included for the study. This totals up to 11540 rows in the dataset.

```
Current status/work history - Current - Paid job or family business (DERIVED BY CAPI)
Yes, paid work one job or business                                          10683
No                                                                           4867
Yes, paid work more than one job or business or number of jobs/businesses missing     682
Yes, unpaid work for family business                                         175
Not known                                                                    14
Not stated or inferred                                                       1
Name: count, dtype: int64
```

Figure 5.10 Output of all current work status value counts

From this, we change those that identify as 'Not stated or inferred' to the question 'Current work - Employee or self-employed' to 'Employee'.

```
df_workers = df_cols.copy()

w = 'Current status/work history - Current - Paid job or family business (DERIVED BY CAPI)'

df_workers = df_workers[(df_workers[w] == 'Yes, paid work one job or business') |
                        (df_workers[w] == 'Yes, paid work more than one job or business or number of jobs/businesses missing') |
                        (df_workers[w] == 'Yes, unpaid work for family business')]

df_workers.loc[df_workers['Current work - Employee or self-employed'] == 'Not stated or inferred', 'Current work - Employee or self-employed'] = 'Employee'

df_workers.shape
```

```
(11540, 88)
```

Figure 5.11 Method of obtaining only workers

Next, the working hours status, that is 'Under Average', 'Average', and 'Over Average' were determined by examining the 'Current work - Hours/week' column and labelling the working hours in another column titled 'WorkHours'. This was done by comparing each of the values against the 'Current work - Hours/week' column and labelling it appropriately according to the country stated in the 'CNTRYID' column.

This was done by iterating through the entire dataset using iterrows() and using if-else statements.

```
df_hours['Current work - Hours/week'] = pd.to_numeric(df_hours['Current work - Hours/week'], errors='coerce')
df_hours['WorkHours'] = 'Average'
df_hours['WorkHours_Difference'] = 0

# SEEING BY EACH COUNTRY
# jpn 40-42
# kor - 40-52
# kaz - 40-52
# sgp - 44-48

for index, row in df_hours.iterrows():
    cntry = row['CNTRYID']
    hours = row['Current work - Hours/week']

    if cntry != 'Singapore':
        if hours < 40:
            df_hours.at[index, 'WorkHours'] = 'Under Average'

            diff = hours - 40
            df_hours.at[index, 'WorkHours_Difference'] = diff

        elif hours > 42:
            if cntry == 'Japan':
                df_hours.at[index, 'WorkHours'] = 'Over Average'

                diff = hours - 42
                df_hours.at[index, 'WorkHours_Difference'] = diff

            elif hours > 52:
                df_hours.at[index, 'WorkHours'] = 'Over Average'

                diff = hours - 52
                df_hours.at[index, 'WorkHours_Difference'] = diff

    elif cntry == 'Singapore':
        if hours > 48:
            df_hours.at[index, 'WorkHours'] = 'Over Average'

            diff = hours - 48
            df_hours.at[index, 'WorkHours_Difference'] = diff

        elif hours < 44:
            df_hours.at[index, 'WorkHours'] = 'Under Average'

            diff = hours - 44
            df_hours.at[index, 'WorkHours_Difference'] = diff
```

Figure 5.12 Method of classifying working hours

Following that, the qualifications mismatch is then considered. The mismatch is done by comparing the qualifications between the columns ‘Education - Highest qualification – Level’ and ‘Current work - Requirements - Education level’. Should the initial column be higher than the former, it is considered ‘Over’ while the opposite is considered ‘Under’. Those with the same qualification levels are denoted as ‘Equal’.

This was performed by first establishing a dictionary with the each of the ISCED levels and assigning them ordinally. A function will check the education column, for foreign and non-foreign, and will return each of the ISCED levels mapped. This function will then iterate through the entire working class dataset.

```python
iscad_level_mapping = {
    'No formal qualification or below ISCED 1': 1,
    'ISCED 1': 2,
    'ISCED 2': 3,
    'ISCED 3C shorter than 2 years': 4,
    'ISCED 3A-B': 5,
    'ISCED 3C 2 years or more': 6,
    'ISCED 3 (without distinction A-B-C, 2y+)': 7,
    'ISCED 4 (without distinction A-B-C)': 8,
    'ISCED 4A-B': 9,
    'ISCED 5A, bachelor degree': 10,
    'ISCED 5A, master degree': 11,
    'ISCED 5B': 12,
    'ISCED 6': 13,
}

isced3 = 4,5,6,7
isced4 = 8,9
isced5 = 10,11,12

def isced_mapping(row, checkcol, foreigncol = 'Education - Highest qualification - Level of foreign qualification'):
    if 'Foreign qualification' in row[checkcol]:
        return iscad_level_mapping.get(row[foreigncol], 0)
    else:
        return iscad_level_mapping.get(row[checkcol], 0)
```

Figure 5.13 Method of obtaining ISCED mapped levels

PV10 was then used as a measure to determine the problem solving performance level of a respondent. A simple function was created that would examine the value present in the PV10 column and return the appropriate performer level based on the value received.

```python
def pvpsl_score(x):
    if x >= 341:
        return 'Strong performer'
    elif x >= 291:
        return 'Moderate performer'
    elif x >= 241:
        return 'Weak performer'
    else:
        return 'At risk'
```

```python
df_pvpsl = df_sorted.copy()

df_pvpsl['PVPSL_Performance'] = 'Moderate performer'

for index, row in df_pvpsl.iterrows():
    score = row[pvpsl[-1]]
    result = pvpsl_score(score)

    df_pvpsl.at[index, 'PVPSL_Performance'] = result

df_pvpsl['PVPSL_Performance'].unique()
```

```
array(['Moderate performer', 'Weak performer', 'At risk',
       'Strong performer'], dtype=object)
```

Figure 5.14 Method of classifying PVPSL performers

Our research objectives require a closer look at the informal learning hours by each participant. To enhance the quality of the data, a look into the relevant columns informs us that there are several accounts of 'Not stated or inferred'.

```
Not stated or inferred count
Column 'NFEHRS': 5224 occurrences
Column 'NFEHRSJR': 6704 occurrences
Column 'AGE_R': 3076 occurrences
Column 'NFEHRSNJR': 6704 occurrences

All zero response count
Column 'ICTHOME': 1334 occurrences
Column 'ICTWORK': 2747 occurrences

Did not participate count
```

Figure 5.15 Output of difference after replacing the informal learning hours

To address this, we take the values from the 'NFEHRS' column, that is the general informal learning hours column, and transfer the values towards the same rows

with empty ‘NFEHRSJR’ and ‘NFEHRSNJR’ which are informal learning hours for job-related purposes and non-job related purposes. Meanwhile, participants that had an ‘All zero response’ for ‘ICTHOME’ and ‘ICTWORK’ were replaced by the value 0.

```
df_num.loc[df_num['ICTWORK'] == 'All zero response', 'ICTWORK'] = 0
df_num.loc[df_num['ICTHOME'] == 'All zero response', 'ICTHOME'] = 0

for index, row in df_num.iterrows():
    hrs = row['NFEHRS']

    if hrs != 'Not stated or inferred':
        if row['NFEHRSJR'] == 'Not stated or inferred' and row['NFEHRSNJR'] == 'Not stated or inferred':
            df_num.at[index, 'NFEHRSJR'] = hrs
            df_num.at[index, 'NFEHRSNJR'] = hrs
        elif row['NFEHRSJR'] == 'Not stated or inferred':
            df_num.at[index, 'NFEHRSJR'] = hrs
        elif row['NFEHRSNJR'] == 'Not stated or inferred':
            df_num.at[index, 'NFEHRSNJR'] = hrs

for index, row in df_num.iterrows():
    hrs = row['NFEHRSJR']

    if hrs != 'Not stated or inferred':
        if row['NFEHRS'] == 'Not stated or inferred':
            df_num.at[index, 'NFEHRS'] = hrs

for index, row in df_num.iterrows():
    hrs = row['NFEHRSNJR']

    if hrs != 'Not stated or inferred':
        if row['NFEHRS'] == 'Not stated or inferred':
            df_num.at[index, 'NFEHRS'] = hrs

df_num['ICTWORK'] = df_num['ICTWORK'].astype(float)
df_num['ICTHOME'] = df_num['ICTHOME'].astype(float)
```

Figure 5.16 Method of filling the informal learning hours

Afterwards, we find that there are no more ‘All zero response’ counts for the ‘ICTHOME’ and ‘ICTWORK’ columns. The number of ‘Not stated or inferred’ count for ‘NFEHRSJR’ and ‘NFEHRSNJR’ have also been reduced from 6704 to 5224.

Following that, the dtype of the dataset is then fixed according to the codebook. This leaves only the columns related to the PVPSL, the working hours in a week, ‘ICTHOME’, and ‘ICTWORK’ to be all numeric values.

For the case of association rule mining, additional steps were implemented to enhance the results of the model. This includes imputing any values that were labelled as ‘Not stated or inferred’ using the most frequent value according to each country.

```
jp = df[df['CNTRYID'] == 'Japan'].copy()
sg = df[df['CNTRYID'] == 'Singapore'].copy()
kz = df[df['CNTRYID'] == 'Kazakhstan'].copy()
kr = df[df['CNTRYID'] == 'Korea'].copy()

countries = [jp, sg, kz, kr]
```

0.0s

```
imputer = SimpleImputer(strategy='most_frequent')

for current in countries:
    count_per_column = df.apply(lambda col: col[col == 'Not stated or inferred'].count())
    notstated_cols = []

    for col, count in count_per_column.items():
        if (current[col] != 'Not stated or inferred').any():
            notstated_cols.append(col)

    current.replace('Not stated or inferred', np.nan, inplace=True)
    current[notstated_cols] = imputer.fit_transform(current[notstated_cols])
```

0.4s

Figure 5.17 Method of imputing null cells

Once done, the dataset was merged again. Additional imputing was also added for two columns, 'LNG_HOME' and 'AGE_R', as they were found to still be null. The same method as before was used on these columns.

Chapter 5.2 Implementing Machine Learning Models

For the first research objective, few prerequisites were required before feeding the data into the machine learning models. First, columns for age, ISCO2C, current work hours, the difference between the work hours with standing working hours, and all the PVPSL columns. Age is dropped because of insufficient data, ISCO2C, and working hours columns were dropped as it hindered the models' performance, and the individual PVPSL columns were dropped as they will be represented by the column denoting the PVPSL performer level instead. The target for this is 'WorkHours' where they are denoted as working 'over average' and 'under average'.

The experiment will be performed on three models, namely a Random Forest Classifier, Decision Tree Classifier, and a SVM. Each of the models' parameters will be fine-tuned to ensure the model is performing at the best possible capacity to achieve the highest scoring. Each of the models feature importance will be extracted and compared with one another.

Before the data was fed into the models, the categorical columns were first encoded using an ordinal encoder to maximise the models' efficiency. In the case of train-test-split, any unknown values found in the test dataset after fitting and encoding it into the training dataset were dealt with by implementing a simple imputer with the strategy set as 'most frequent'. Meanwhile, the target variable goes through a .ravel() function which returns the target array in a contiguous flattened array form. The experiment is performed twice, once with train-test-split using varying degrees of test sizes, that is 5% to 30% in intervals of 5, and another using cross validation with a range of 2 to 10 folds.

```
1  sizes = [0.05, 0.1, 0.15, 0.2, 0.25, 0.3]
2  enc = OrdinalEncoder(handle_unknown='use_encoded_value', unknown_value=np.nan)
3  imputer = SimpleImputer(strategy='most_frequent')
4
5  cat = X.select_dtypes(include = 'object').columns
```

Figure 5.18 Test sizes, method of encoding, imputation method, and categorical columns selection

To prepare the train and test sets for the train-test-split experiment, the train and test sets were split beforehand by running a function that would generate sets based on the sizes given. The results of the splits were appended to a list to be used during the experimentation.

```
X_train_enc_list = []
X_test_enc_list = []
y_train_list = []
y_test_list = []

for size in sizes:
    # encoding
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=size, random_state=5)

    y_train = y_train.values.ravel()
    y_test = y_test.values.ravel()

    y_train_list.append(y_train)
    y_test_list.append(y_test)

    X_train_enc = X_train.copy()
    X_test_enc = X_test.copy()

    X_train_enc[cat] = enc.fit_transform(X_train_enc[cat])
    X_test_enc[cat] = enc.transform(X_test_enc[cat])

    X_train_enc[cat] = imputer.fit_transform(X_train_enc[cat])
    X_test_enc[cat] = imputer.transform(X_test_enc[cat])

    X_train_enc_list.append(X_train_enc)
    X_test_enc_list.append(X_test_enc)
```

Figure 5.19 Train-test-split function

The case is different in using the cross validation method. A function is defined to be able to take the model and the current fold as an input. Using the 'test_score' column from the cross validation output, the mean of the column is taken as the score for the models' output for the current fold.

```
folds = [i for i in range(2, 11)] # 2-10 folds

# encoding
X_encoded = X.copy()

X_encoded[cat] = enc.fit_transform(X_encoded[cat])
X_encoded[cat] = imputer.fit_transform(X_encoded[cat])

y_encoded = y.values.ravel()

X_enc, y_enc = shuffle(X_encoded, y_encoded, random_state=5)

def crossValidate(model, fold):
    cv_res = cross_validate(model, X_enc, y_enc, cv=fold, return_train_score=False)

    mean = cv_res['test_score'].mean()

    print('Fold: ', fold)
    # print(cv_res.keys())
    print('Scores: ', cv_res['test_score'], '-'*25, 'Mean score: ', mean)
    print()

    return mean
```

Figure 5.20 Cross validation method

To evaluate each of the models, a function is defined that would take the target test array, the predicted output, and the probability of the predicted output, in order to evaluate the accuracy score, create a confusion matrix, generate a classification report and obtain the ROC AUC score.

```
report_items = ['precision', 'recall', 'f1-score', 'support']

# evaluation function
def getEvaluationMetrics(y_test, y_pred, y_proba):
    evals = []
    evals_score = [] # for tables

    # accuracy score
    score = accuracy_score(y_test, y_pred)
    evals.append(score)
    evals_score.append(score)

    # confusion matrix
    cm = confusion_matrix(y_test, y_pred)
    evals.append(cm)

    # classification report
    report = classification_report(y_test, y_pred, target_names=class_labels)
    evals.append(report)

    report_dict = classification_report(y_test, y_pred, target_names=class_labels, output_dict=True)
    for i in class_labels:
        for item in report_items:
            evals_score.append(report_dict[i][item])

    # roc auc
    auc = roc_auc_score(y_test, y_proba)
    evals.append(auc)
    evals_score.append(auc)

    return evals, evals_score
```

Figure 5.21 Function to obtain evaluation metrics

Another function is defined to output the evaluation metrics within the Jupyter notebook.

```python
def showEvaluationMetrics(evals, text=''):
    acc = 'Accuracy score'
    cm =  'Confusion matrix'
    rep = 'Classification report'
    roc = 'AUC ROC score'

    if text != '':
        acc = acc + ' for ' + text
        cm = cm + ' for ' + text
        rep = rep + ' for ' + text
        roc = roc + ' for ' + text

        print(acc, ' : ', evals[0])

        sns.heatmap(evals[1], annot=True, fmt='g',
                xticklabels=class_labels,
                yticklabels=class_labels,)
        plt.xlabel('Prediction')
        plt.ylabel('Actual')
        plt.title(cm)
        plt.show()

        print(rep, ' : ')
        print(evals[2])

        print(roc, ':', evals[3])
    else:
        print(acc, ' : ', evals[0])

        sns.heatmap(evals[1], annot=True, fmt='g',
                xticklabels=class_labels,
                yticklabels=class_labels,)
        plt.xlabel('Prediction')
        plt.ylabel('Actual')
        plt.title(cm)
        plt.show()

        print(rep, ' : ')
        print(evals[2])

        print(roc, ':', evals[3])
```

Figure 5.22 Function to print evaluation metrics

For the cross validation method, another function was defined to output the predictions with the model and fold number as the input. This function returns the predicted target values and predicted probabilities of the target values using the .cross_val_predict() function.

```python
def crossValidate_pred(model, fold):
    cv_res = cross_validate(model, X_enc, y_enc, cv=fold, return_estimator =True)

    y_pred = cross_val_predict(model, X_enc, y_enc, cv=fold)
    y_proba = cross_val_predict(model, X_enc, y_enc, cv=fold, method='predict_proba')[:, 1]

    return cv_res, y_pred, y_proba
```

Figure 5.23 Function to obtain probability from cross validation

Another cross validation specific function, that is to generate the feature importances from the model, is also defined. Here, it takes the cross validation output and obtains the importance score of each estimator. Once the importance scores are gathered from the estimators, the mean of the scores will be returned.

```python
def get_importance(cv_res, criterion):
    features = []
    col_cri = 'Features_' + str(criterion)
    col_imp = 'Importance_' + str(criterion)

    for estimator in cv_res['estimator']:
        features_df = pd.DataFrame({col_cri: feature_names,
                                    col_imp: estimator.feature_importances_,}).sort_values(by=col_cri).reset_index(drop=True)
        # sort for consistency
        features.append(features_df[col_imp].values)

    mean_features = np.mean(features, axis=0)

    mean_features_df = pd.DataFrame({col_cri: feature_names,
                                    col_imp: mean_features}).sort_values(by=col_imp, ascending=False).reset_index(drop=True)

    return mean_features_df
```

Figure 5.24 Function to obtain feature importance for cross validation

Once all the data preprocessing is finished, it is ready to be inputted into the models. To find the best accuracy each of the models will go through a for loop where they will iterate over possible criterions in the case of the Random Forest Classifier and Decision Tree Classifier and possible kernels for the SVM. They will also iterate through all the possible train-test splits and folds. In the case of the Random Forest Classifier, an additional factor, that is the n value, will also be considered. For this research, the n values tested are within the range of 5 to 100 with intervals of 5.

```
n_vals = [i for i in range(5, 105, 5)]
criterions = ['gini', 'entropy', 'log_loss']

scores_dict = {}

for cri in criterions:
    scores_dict[cri] = {}

    for fold in folds:
        scores_dict[cri][fold] = []
        index = folds.index(fold)

        scores = []

        for n in n_vals:
            rf = RandomForestClassifier(n_estimators=n, criterion=cri, random_state=5)
            score = crossValidate(rf, fold)
            scores_dict[cri][fold].append(score)
```

Figure 5.25 Random Forest Classifier implementation

The accuracy scores for each of these is gathered into a dictionary and from there, the highest accuracy model from each criterion and kernel is selected. The scorings for these will then be stored into a csv file for an easy means of referencing.

```
all_scores = []
top_scores = {}

for cri, inner in scores_dict.items():
    max_size = 0
    max_score = 0

    print(cri)

    for size, acc in inner.items():

        print(size, '-', acc)
        all_scores.append((str(cri), str(size), acc))

        if acc > max_score:
            max_score = acc
            max_size = size
    top_scores[cri] = {'Fold': max_size, 'Accuracy': max_score}
    print()

print('Top Scores')
dt_top_scores = top_scores.copy()

print(top_scores)
```

Figure 5.26 SVM implementation

From the highest accuracy scores of each criterion and kernel, the feature importance will be extracted. The top scorers will also go through the classification evaluation functions to examine the performance of each of the models.

```
#  feature_importances with top scores

feature_df_list = []
eval_scores_list = []

for cri, inner in top_scores.items():
    fold = inner['Fold']

    model = DecisionTreeClassifier(criterion=cri, random_state=5)

    cv_res, y_pred, y_proba = crossValidate_pred(model, fold)

    eval, eval_scores = getEvaluationMetrics(y_enc, y_pred, y_proba)
    showEvaluationMetrics(eval, ('Decision Tree Classifier - ' + cri))
    print('-' * 100)
    eval_scores_list.append(eval_scores)

    features_df = get_importance(cv_res, cri)
    feature_df_list.append(features_df)
```

Figure 5.27 Decision Tree Classifier implementation

The classification report results alongside the ROC AUC scores gathered will then be stored in a csv file for an easier means of referencing.

```
eval_cols = ['Accuracy',
             'OverAverage_precision', 'OverAverage_recall', 'OverAverage_f1-score', 'OverAverage_support',
             'UnderAverage_precision', 'UnderAverage_recall', 'UnderAverage_f1-score', 'UnderAverage_support',
             'AUC_ROC_score']
```

```
rfc_top_scores_eval_df = pd.DataFrame(eval_scores_list, columns=eval_cols, index=criterions)
save = folder_path + 'RandomForestClassifier_TopAccuracy_Evaluation.csv'
rfc_top_scores_eval_df.to_csv(save, index=False)
display(rfc_top_scores_eval_df)
```

Figure 5.28 Method of saving evaluation metrics to a csv file

Similarly, the feature importances gathered from the models will also be stored into a csv file for easier access later.

```
rfc_importance_df = pd.DataFrame()
for i in feature_df_list:
    rfc_importance_df = pd.concat([rfc_importance_df, i], axis=1)

save = folder_path + 'RandomForestClassifier_TopAccuracy_Importance.csv'
rfc_importance_df.to_csv(save, index=False)
display(rfc_importance_df)
```

Figure 5.29 Method of saving feature importance to a csv file

Chapter 5.3 Association Rule Mining Implementation

The target for the association rule mining is the qualification status of an individual. This includes over qualified, those who are equally qualified, and under qualified, denoted as 'Equal', 'Over', and 'Under' respectively. The dataset is split by these qualifications, generating three datasets.

Association rule mining will be conducted using the apriori and mlextend libraries. A function is defined to make the association rules, taking the data, the support value, and threshold value as input. Note that the 'type' input is simply a string input to aid in naming the csv file appropriately.

```
def makeAssocRules(type, data, support, threshold):
    # filename_growth = results_path + type + '/ARM_Growth_' + str(support) + '.csv'

    fp_apriori = fp.apriori(data,min_support=support,use_colnames=True)
    # fp_fpgrowth = fp.fpgrowth(data, min_support=support, use_colnames=True)

    assoc_rules = fp.association_rules(fp_apriori,metric='confidence', min_threshold=threshold,support_only=False)

    # fp_fpgrowth.to_csv(filename_growth, index=False)

    refineRules(type, assoc_rules, support, threshold)
```

Figure 5.30 Method of generating association rules

Once the rules are generated, they will be further refined by only keeping rules that content the target that is 'Qualification_Status'. Two csv files will be generated, wherein one contains rules with consequents with at least the target, and another where it contains only the target.

```python
target = 'Qualification_Status'

# get only the 1 cons

def refineRules(type, assoc_rules, support, threshold):
    filename_con_all = results_path + type + '/ARM_consequents_' + str(support) + '_' + str(threshold) + '.csv'
    filename_con = results_path + type + '/ARM_consequents_ONE_' + str(support) + '_' + str(threshold) + '.csv'

    con_all_condition = assoc_rules['consequents'].apply(lambda x: any(target in str(item) for item in x))
    con_one_condition = assoc_rules['consequents'].apply(lambda x: any(target in str(item) for item in x) and len(x) == 1)

    consequents_all = assoc_rules[con_all_condition]
    consequents_one = assoc_rules[con_one_condition]

    consequents_all.to_csv(filename_con_all, index=False)
    consequents_one.to_csv(filename_con, index=False)
```

Figure 5.31 Method of refining association rules and saving to csv file

To generate quality rules, a support value of 0.5 will be used and a confidence threshold of 0.9. The rules will only include those with a support above 0.9 should the results allow it.

Investigating the rules requires additional functions to aid the process. Such functions include one that would prune the rules and obtain rules that are considered the most relevant. That is, having a high value of lift, support, and confidence.

```python
def pruneRules(data, lift = 1, confidence = 0.9, support = 0.9, sorting = 'lift'):
    newData = data.copy()

    newData = newData[newData['lift'] > lift]
    newData = newData[newData['confidence'] > confidence]
    newData = newData[newData['support'] > support]

    newData.sort_values(by=sorting, ascending=False, inplace=True)
    newData.reset_index(drop = True, inplace = True)

    newData.drop(columns = ['antecedent support', 'consequent support', 'leverage', 'conviction', 'zhangs_metric'], inplace=True)

    print(newData.shape)
    print('Range for LIFT: ', np.min(newData['lift']), ' - ', np.max(newData['lift']))
    print('Range for CONFIDENCE: ', np.min(newData['confidence']), ' - ', np.max(newData['confidence']))
    print('Range for SUPPORT: ', np.min(newData['support']), ' - ', np.max(newData['support']))


    return newData
```

Figure 5.32 Method of pruning the association rules

Additionally, in finding commonalities between the rules, a function was made for doing as such.

```python
def getCommon(compare):
    common = set(compare[0])

    for i in range(1,len(compare)):
        common &= set(compare[i])

    print(len(common))

    for i in common:
        print(i)

    return common
```

Figure 5.33 Method of obtaining the common values for association rules

Another function for finding the except, that is to find a unique value not present in any other lists given, was also defined. This was done by simply a combination of every other list before comparing it against the current list.

```python
naming = ['Over', 'Under', 'Equal']

def getExcept(compare):
    things = []

    for i in range(len(compare)):
        union_others = set()

        for j in range(len(compare)):
            if i != j:
                union_others |= set(compare[j])

        print(naming[i], '-' * 150)

        not_common = set(compare[i]) - union_others

        print(len(not_common))
        for item in not_common:
            print(item)
        print()
        things.append(not_common)

    return things
```

Figure 5.34 Method of obtaining the unique values for association rules

Chapter 6 Evaluation of Findings

Chapter 6.1 Research Objective 1

The first research objective requires investigation towards those working over the required working hours and under the required working hour. To recap, the objective is "examining factors that influence working over the required working hours and under the required working hours."

Chapter 6.1.1 Train-Test-Split

Results from the train-test-split showcased that in terms of accuracy scores alone, the best parameters for the Random Forest Classifier were using the gini criterion with an n_val of 60 using a 15% test size to achieve a 0.757 accuracy. The decision tree classifier on the other hand was tied between using entropy and log loss criterion, both achieved the same score of 0.676 using a 5% test size. Meanwhile, the SVM showed best results when using the linear kernel with a 15% test size, achieving an accuracy of 0.730.

Table 6.1 Train-test-split accuracy results

Model	Criterion/Kernel	Test size	Accuracy (rounded up to 3 decimals)	N_val (for Random Forest Classifier)
Random Forest Classifier	Gini	0.15	0.757	60
	Entropy	0.15	0.753	70
	Log_loss	0.15	0.753	70
Decision Tree Classifier	Gini	0.05	0.667	-
	Entropy	0.05	0.676	-
	Log_loss	0.05	0.676	-
Support Vector Machine (SVM)	Linear	0.15	0.730	-
	Poly	0.2	0.626	-
	RBF	0.1	0.627	-
	Sigmoid	0.1	0.547	-

Delving deeper into the accuracy scores, classification matrices were applied to each of the models' best performers. The Random Forest Classifier with gini criterion showcases the best overall performance, scoring the highest precision, recall, and f1-score for both the 'Over Average' and 'Under Average' class labels with scores

within the range of 0.73 and 0.78, with a support of 483 for 'Over Average' and 507 for 'Under Average'. It also achieved the highest AUC ROC score that is 0.828529. The second best model goes to SVM with linear kernel. Precision, recall, and f1-scores for both classes were within the range of 0.70 and 0.76, while the support scores were similar to the Random Forest Classifier. The AUC ROC score achieved was 0.804195. Both Decision Tree Classifiers present similar results, resulting in them both being the least performing model in the experiment. Scores for precision, recall, and f1-score were between the range of 0.64 and 0.70, while supports for both were below 156 and 174 for the 'Over Average' and 'Under Average' classes respectively. Both models achieved low scores for AUC ROC that is 0.676945.

Table 6.2 Train-test-split classification metrics result

Model		**Random Forest Classifier (Gini)**	**Decision Tree Classifier (Entropy)**	**Decision Tree Classifier (Log_loss)**	**SVM (Linear)**
Over Average	**Precision**	0.736328	0.644970	0.644970	0.708494
	Recall	0.780538	0.698718	0.698718	0.759834
	F1-Score	0.757789	0.670769	0.670769	0.733267
	Support	483	156	156	483
Under Average	**Precision**	0.778243	0.708075	0.708075	0.754237
	Recall	0.733728	0.655172	0.655172	0.702170
	F1-Score	0.755330	0.680597	0.680597	0.727273
	Support	507	174	174	507
AUC ROC Score		0.828529	0.676945	0.676945	0.804195

Feature importances were then extracted from the models with the highest accuracy scorings. Our findings reveal that all the models agree that 'Current status/work history - Subjective status' takes the most priority in determining whether an individual is classed as working over or under the average working hours. The Random Forest Classifier, and both Decision Tree Classifiers agree up to the first 6 features, that is the gender, ICT skill at home and at work, the area of study for an individual's highest education, and the language spoken at home. In the case of the SVM linear kernel model, the findings show that being a native speaker, country of origin, the level of computer use at work, how often an individual conducts manual labour for longs at work, and the ISCOSKIL 4, are determining factors in the working hours of an individual.

Table 6.3 Train-test-split top feature importance

Random Forest Classifier (Gini)		**Decision Tree Classifier (Entropy)**		**Decision Tree Classifier (Log_loss)**		**SVM (Linear)**	
Feature	**Importance**	**Feature**	**Importance**	**Feature**	**Importance**	**Feature**	**Importance**
Current status/work history - Subjective status	0.104349	Current status/work history - Subjective status	0.093441	Current status/work history - Subjective status	0.093441	Current status/work history - Subjective status	0.388185
GENDER_R	0.039295	GENDER_R	0.031380	GENDER_R	0.031380	NATIVESPEAKER	0.346316
ICTHOME	0.030366	ICTHOME	0.031114	ICTHOME	0.031114	CNTRYID	0.165858
ICTWORK	0.028993	ICTWORK	0.028774	ICTWORK	0.028774	Skill use work - ICT - Computer - Level of computer use	0.150191
Education - Highest qualification - Area of study	0.024812	Education - Highest qualification - Area of study	0.024597	Education - Highest qualification - Area of study	0.024597	Skill use work - How often - Working physically for long	0.113012
LNG_HOME	0.021547	LNG_HOME	0.021143	LNG_HOME	0.021143	ISCOSKIL4	0.108504

For features ranked at the lowest importance, similarities were only present for the bottom two features. The Random Forest Classifier with gini criterion and both the entropy and log loss criterion Decision Tree Classifiers agree that experience with a computer in the job at work and highest level of foreign qualification were not strong factors to determining the working hours. On the other hand, SVM with linear kernel finds that features that were not important to predicting working hours were whether the individual was born in the country they took their assessment and their gender. It should be noted that the SVM kernel was the only model that had features with an importance scoring of below 0.

Table 6.4 Train-test-split bottom feature importance

Random Forest Classifier (Gini)		Decision Tree Classifier (Entropy)		Decision Tree Classifier (Log_loss)		SVM (Linear)	
Feature	Importance	Feature	Importance	Feature	Importance	Feature	Importance
Skill use work - ICT - Experience with computer in job	0.001432	Skill use work - ICT - Experience with computer in job	0.001196	Skill use work - ICT - Experience with computer in job	0.001196	Background - Born in country	-0.550901
Education - Highest qualification - Level of foreign qualification	0.000059	Education - Highest qualification - Level of foreign qualification	0.000063	Education - Highest qualification - Level of foreign qualification	0.000063	GENDER_R	-0.925268

Chapter 6.1.2 Cross-Validation Implementation

The best accuracy results from the cross validations showed that the best number of folds were 2, 3, 4, 7, 9, and 10. The Random Forest Classifier performed its best with an n_val of 100 for each criterion, with the gini criterion obtaining the highest accuracy compared to the remaining two that is an accuracy of 0.760. As for the Decision Tree Classifier, the gini criterion performed best, obtaining an accuracy of 0.678. Finally, the SVM model showcased that the linear kernel outperformed the other kernels wit an accuracy score of 0.742.

Table 6.5 Cross validation accuracy results

Model	Criterion/Kernel	Folds	Accuracy (rounded up to 3 decimals)	N_val (for Random Forest Classifier)
Random Forest Classifier	Gini	7	0.760	100
	Entropy	4	0.759	100
	Log_loss	4	0.759	100
	Gini	3	0.678	-

Decision Tree Classifier	Entropy	3	0.672	-
	Log_loss	3	0.672	-
Support Vector Machine (SVM)	Linear	9	0.742	-
	Poly	9	0.627	-
	RBF	10	0.614	-
	Sigmoid	2	0.514	-

Classification matrices were then applied to the models with the highest accuracies. Here, we observe that both the Random Forest Classifier with gini criterion and the SVM with linear kernel were the best suited models for classifying individuals with over and under working hours. Each of their precision, recall, and f1-scores are above 0.7, while their AUC ROC score is above 0.8. Meanwhile, the decision tree classifier scored above 0.6 but below 0.7 in each category, denoting it as the lesser effective model. Each model however, scored 3250 for over average support, and 3347 for under average support.

Table 6.6 Cross validation classification results

Model		**Random Forest Classifier (Gini)**	**Decision Tree Classifier (Gini)**	**SVM (Linear)**
Over Average	**Precision**	0.741664	0.673331	0.721508
	Recall	0.787077	0.673538	0.777231
	F1-Score	0.763696	0.673435	0.748334
	Support	3250	3250	3250
Under Average	**Precision**	0.780178	0.682905	0.766150
	Recall	0.733791	0.682701	0.708694
	F1-Score	0.756274	0.682803	0.736303
	Support	3347	3347	3347
AUC ROC Score		0.839633	0.678120	0.815173

The features were then extracted from the models with the highest accuracies. Random Forest Classifier with gini criterion and Decision Tree Classifier with gini criterion both shared their top 3 features, those being the educational level requirements of an individual's current workplace, the gender of an individual, and how often teaching was involved in the workplace. Meanwhile, the SVM with linear kernel had a different set of features for the first top 3 attributes. Those being whether the individual was a native speaker of the country they took the PIAAC assessment in, their current work status, and the country they took the assessment in.

Table 6.7 Cross validation top feature importance

Random Forest Classifier (Gini)		Decision Tree Classifier (Gini)		SVM (Linear)	
Feature	**Importance**	**Feature**	**Importance**	**Feature**	**Importance**
Current work - Requirements - Education level	0.100609	Current work - Requirements - Education level	0.189814	NATIVESPEAKER	0.395082
GENDER_R	0.038942	GENDER_R	0.057833	Current status/work history - Subjective status	0.394827
Skill use work - How often - Teaching people	0.030630	Skill use work - How often - Teaching people	0.035822	CNTRYID	0.179803

As for features that ranked lowest, both the Random Forest Classifier and Decision Tree Classifier with gini criterion state that 'ICTWORK' were of low importance. Meanwhile, the SVM model with linear kernel states that the individual's gender was of low importance.

Table 6.8 Cross validation bottom feature importance

Random Forest Classifier (Gini)		Decision Tree Classifier (Gini)		SVM (Linear)	
Feature	**Importance**	**Feature**	**Importance**	**Feature**	**Importance**
ICTWORK	0.000061	ICTWORK	0.000000	GENDER_R	-0.887238

Chapter 6.2 Research Objective 2

To answer the second research objective, that is, "conducting association rule analysis to identify factors influencing the mismatch between highest qualifications and employment qualifications", requires diving into the rules generated by the association rule mining model. This involves a thorough analysis of the rules and observing the disparities between an individual's highest educational qualification and the qualifications required for their employment.

The rule analysis will cover consequents that have at least the qualification status, and rules where the consequents contain only the qualification status. The qualification status here being the type of disparity between the highest education and the qualification required by their employer.

Chapter 6.2.1 Over qualified

The model obtained 403204 rules where the consequents contained an occurrence of the qualification status alongside other rules. Of those rules, 30 were found that showcased a dependant relationship between the antecedents and the consequent and has a support and confidence of over 0.9. Those rules can be seen in appendix C.5 and appendix C.6.

The rules found show a lift between the range of 1.0001865341172709 to 1.051924159103957. The range of confidence for those rules is from 0.9339764201500536 to 0.9829419583517944. Meanwhile, the support showcased a range of 0.901323955316508 to 0.917873396772859.

8 unique antecedents were found, and from those were 5 unique attributes. Those being a native speaker of the country where the induvial took the assessment, being born in the country where the individual took the assessment, having used a computer in their everyday life, having a count of 1 to 10 employees working under you, formal education being related to the job,

Meanwhile, 9 unique consequents were found alongside the qualification status. From these were 5 unique attributes, those are being born in the country where the assessment was taken, being a native speaker of the country where the assessment was taken, having used a computer in everyday life, having a count between 1 to 10 employees working for you, and having a formal education being for job related reasons.

For rules where there was only one consequent, 39419 rules were found. 25 rules were found that had a confidence and support of 0.9. Those rules had a lift and confidence of 1, while the support ranges from 0.901323955316508 to 1.

25 unique antecedents were found and from these were 7 unique attributes. These include the individual having used a computer in everyday life, having a count between 1 to 10 employees working for the individual, formal education qualification being related to the job, being a native speaker of the country where the assessment was taken, being born in the country where the assessment was taken, managing a count between 1 to 5 employees, and current work status being paid work be it a job or a business.

The 5 common antecedents attributes found between the two are formal education being job related, having used a computer in everyday life, being born in the country where the assessment was taken place, having a count between 1 to 10 employees working for you, and being a native speaker of the country where the assessment was taken. A table of the unique antecedents for over qualified workers can be found in appendix C.4.

Chapter 6.2.2 Under qualified

For consequents that include attributes besides the qualification status, 638332 rules were generated. These rules were further refined to only include those with a lift over 1, a support and confidence of over 0.9. 54 rules fit these criteria, ranging a lift between 1.08224924657898 to 1.08224924657898. Meanwhile, the confidence had a range between 0.983745123537061 to 0.9967061923583662, while the support had a solid value of 0.9059880239520958. These rules can be seen in appendix C.8 and appendix C.8.

16 unique antecedents were gathered. The unique attributes include being a native speaker in the country where the assessment was taken, having used a computer in everyday life, being born in the country where the assessment was taken, having a count of 1 to 10 employees working for you, and they are an employee at work.

Meanwhile, 16 unique consequents found. The 5 unique attributes for these are being born in the country where the assessment was taken place, being a native speaker of the country where the assessment was taken place, being an employee, having used a computer in everyday life, and having a count of 1 to 10 employees working for you.

For rules where the only consequent is the qualification status, 50167 rules were generated. Further refining the rules to have a support of above 0.9 leaves us with 55 rules. These rules have a lift and confidence of 1, while the support ranges from 0.9005988023952096 to 1.

From these rules, 55 unique antecedents were extracted. The 7 unique attributes from these include being an employee, having used a computer in everyday life, having a count 1 to 10 employees working for you, having a formal educational qualification for job related reasons, current work status is paid work for a job or business, being a native speaker of the country where the assessment is taken place, and having experience with a computer in everyday life.

The common antecedents' unique attributes found were having used a computer in everyday life, being a native speaker of the country where the assessment was taken, being an employee, having a count of 1 to 10 employees working for you, and being born in the country where the assessment was taken. A table of unique antecedents for under qualified workers can be found in appendix C.7.

Chapter 6.2.3 Equal qualifications

For those with equal qualifications, a total of 797219 rules were generated. Further refining the rules to only include those with a confidence and support of over 0.9 and lift of over 1 gives us 108 rules. The lift from these rules ranges from 1.0023796923457946 to 1.0746374096900175. Meanwhile, the confidence for these ranges from 0.9503027771977448 to 0.9863466196872936 while the support ranges from 0.9019331453886428 to 0.9164317358034636. These rules can be seen in appendix C.11 and appendix C.12.

There were 34 unique antecedents found, and the unique attributes for these include being a native speaker of the country where the assessment was taken, being an employee, being born in the country where the assessment was taken, having 1 to 10 employees working for you, had used a computer in everyday life, formal education qualification being for work related reasons, and current paid work being a job or a business. A table of unique antecedents for under qualified workers can be found in appendix C.10.

Chapter 6.2.4 Comparison of Association Rule Mining Results

An intersect, that is the common elements, of these rules reveal that 3 most common attributes for antecedents that have several consequent values are being born in the country where the assessment was taken place, having used a computer in everyday life, and being a native speaker of the country where the assessment took place.

Meanwhile, there are 6 unique attributes for antecedents where the qualification status was the only consequent. Those are being born in the country where the assessment was taken place, formal education being related to the job, having used a computer in everyday life, count of employees working for you is between 1 to 10, and being a native speaker of the country where the assessment is taken place.

The common attributes present in both these results are being born in the country where the assessment was taken, having used a computer in everyday life, and being a native speaker of the country where the assessment was taken place.

The intersects between these rules provide an idea of what commonly occurs between each of these classes, which may indicate that it is simply a common trait and does not influence how an individual addresses their skill gap.

Rules and attributes that are unique from other classes could reveal insight for what leads to being over qualified, under qualified, and equal qualified. A breakdown of the common antecedents found can be seen in appendix C.13.

A look into the antecedents where there are other values in the consequents besides the qualification status reveals several pieces of information. For those who are over qualified for their current careers, the attribute found is that they have used a computer in everyday life and their current work involves having a count between 1 to 10 people working for the individual. Meanwhile, for those who are of equal qualification, the attributes found are current paid work being a paid job or business, formal education qualification being job related, having used a computer in everyday life, being an employee, and having a count between 1 to 10 people working for the individual. The under qualified rules were not able to obtain rules that were not present in the over qualified and equal qualifications rules.

As for the antecedents where the consequents is only the qualification status, several attributes were found for all 3 categories. The over qualified attributes include those who were born in the country where the assessment was taken, having a formal education related to their job, having used a computer in everyday life, managing a count between 1 to 5 other employees at work, and being a native speaker of the country where the assessment was done. Meanwhile, the under qualified attributes were having experience with a computer in everyday life, being born in the country where the assessment was taken, having paid work that is a job or a business, formal education being job related, having used a computer in everyday life, being an employee, having a count between 1 to 10 people working for the individual, and being a native speaker of the country where the assessment was taken. Finally, the equal qualifications attributes include having experience with a computer in everyday life, being born in the country where the assessment was taken, having paid work be it a job or a business, formal education qualification being related to the job, having used a computer in everyday life, being an employee, having a count of 1 to 10 people working for you, and being a native speaker of the country where the assessment was taken. A table view of this can be found in appendix C.14.

Chapter 6.3 Research Objective 3

The third research objective involves investigating how individuals address the skill gap when they lack the necessary qualifications in their current profession. The attributes taken for the association rule mining will be based on the unique attributes found in both the antecedents with multiple consequent values and a singular consequent value, instead of the unique antecedents. The common attributes found in all three classes is removed, leaving only those unique to the class.

In addressing the skill gap, we look over at the results from each of our models. For both the train-test-split and cross validation, they each share the common attribute of gender, but they share no other attributes with the association rule mining results. This would indicate that these factors are highly associated with working hours alone as opposed to association with education and work requirement gap.

The association rule mining results showcase that for the over qualified workers, the common attributes are those born in the country of the where the assessment was taken, those having a formal education related to their job, having used a computer in their everyday life, having a count between 1 to 10 employees working for the individual, and being a native speaker of the country where the assessment was taken place. These attributes entail an over qualified individual is most often a local citizen, manage a small team at work, use a computer in their daily lives, and took an education for their current career.

Meanwhile, attributes that are associated with those who are under qualified include those born in the country where the assessment was taken place, having used a computer in their everyday life, having a count between 1 to 10 employees working for the individual, is an employee themselves, and is a native speaker of the country where the assessment was taken place. The attributes lead to a local who manages a team under their superior, and most likely uses a computer at home.

As for those with equal qualifications, we gather the attributes being born in the country where the assessment was taken place, being a native speaker of the

country where the assessment was taken place, having a formal education due to job related reasons, receiving paid work be it a job or a business, having used a computer in their everyday life, having a count between 1 to 10 employees working under the individual, and being an employee. The description of a local citizen with a qualification suited for their career comes to mind. This individual is also an employee managing a small team, while working a paid job that may or may not be family business.

Table 6.9 Unique attributes from feature importance and association rule mining

	Unique attributes
Train-test-split	• Current status/work history - Subjective status • GENDER_R • ICTHOME • ICTWORK • Education - Highest qualification - Area of study • LNG_HOME
Cross validation	• Current work - Requirements - Education level • GENDER_R • Skill use work - How often - Teaching people
Association Rule Mining – Over qualified (Intersect)	• Background - Born in country_Yes • Education - Formal qualification - Reason job related_Yes • Skill use everyday life - ICT - Ever used computer_Yes • Current work - Employees working for you - Count_1 to 10 people • NATIVESPEAKER_Yes
Association Rule Mining – Under qualified (Intersect)	• Background - Born in country_Yes • Skill use everyday life - ICT - Ever used computer_Yes • Current work - Employees working for you - Count_1 to 10 people • Current work - Employee or self-employed_Employee • NATIVESPEAKER_Yes
Association Rule Mining –Equal qualification (Intersect)	• Background - Born in country_Yes • Education - Formal qualification - Reason job related_Yes • Current status/work history - Current - Paid job or family business (DERIVED BY CAPI)_Yes, paid work one job or business • Skill use everyday life - ICT - Ever used computer_Yes • Current work - Employees working for you - Count_1 to 10 people • Current work - Employee or self-employed_Employee • NATIVESPEAKER_Yes

Chapter 7 Conclusion

The first research objective, to examine factors that influence working over or under the required working hours, was conducted with the implementation of 6 machine learning models. Of these models, the Random Forest Classifier with gini criterion showed itself as the best fit model thus far. The common features extracted from all the models used indicated that gender was a primary factor in determining working hours. Meanwhile, other factors such as an individuals' ICT skills at work and at home, the area of study for their highest educational qualification, language spoken at home, their occupation's educational requirement level, and how often they teach people at work, are additional factors to consider.

The second objective was achieved by conducting an association rule analysis to identify factors influencing the mismatch between highest qualification and employment qualification requirement. The output highlights that those who are not under qualified are associated with having an educational qualification that is related to their career. Meanwhile, equal qualification workers showcase that their current paid work is either a job or a business. Both these indicate that those who are under qualified tend to lack the educational requirements for their career. They also may not be recipients of paid work.

The final objective, that is to investigate how individuals address the skill gap when they lack the necessary qualifications in their current profession, was done by comparing the outputs obtained from the experiments conducted for this project. What is concluded is that in addressing the issue of how individuals address the skill gap when they lack the necessary qualifications in their current profession, the findings show that most often the difference between those who are under qualified and those who have equal qualifications are working a paid work be it a job or a business. On the other hand, the difference between those who are over qualified and those who are under qualified is that over qualified individuals tend to take a leadership role as opposed to a passive one.

To summarize, the project reveals that gender is a primary factor influencing individuals' working hours. The interplay between highest qualification, educational attainment, and professional role at work is crucial in addressing career skill gaps. It highlights the importance of aligning educational qualifications with career requirements and understanding the behaviours of those who exceed these skill gaps. These findings can aid in the development of policies and support mechanisms to enhance individuals' professional careers.

Chapter 7.1 Limitations

Since the majority of researchers utilised software programmes like STATA, the financial need for utilising these capabilities was not able to be fulfilled. Therefore, additional investigation was necessary to discover other approaches that were open source or freely available for use.

The initial framework set for FYP1 had to be remade for the 2nd phase of FYP. Due to this, additional time was taken to reconfigure a new framework for this project.

The size and complexity of the dataset proved itself robust as the association rule mining parameters were limited to ensure it would run smoothly. This is due to the hardware equipment used being unable to support the amount of memory required to run the model.

Chapter 7.2 Future Plans

Additional improvements, specifically, the concept of categorising workers according to the ISCOSKIL4 classification, would further refine the scope of the project and help gain more insights towards each of the specific classification details. This will provide a comprehensive study to compare and identify the specific categories of workers that are particularly prone to experiencing skill mismatch and exceeding the normal working hours. Here, the distinct demands and wants of various worker groups can be identified to determine the most effective approach to reducing the skill gap they possess.

FYP2-Report-01072026

ORIGINALITY REPORT

9%	8%	6%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.oecd.org Internet Source	2%
2	nces.ed.gov Internet Source	1%
3	academicworks.cuny.edu Internet Source	<1%
4	docplayer.net Internet Source	<1%
5	www.tandfonline.com Internet Source	<1%
6	doi.org Internet Source	<1%
7	www.abs.gov.au Internet Source	<1%
8	madoc.bib.uni-mannheim.de Internet Source	<1%
9	Franziska Hampf, Simon Wiederhold, Ludger Woessmann. "Skills, earnings, and employment: exploring causality in the	<1%

estimation of returns to skills", Large-scale Assessments in Education, 2017
Publication

10	sda.chass.utoronto.ca Internet Source	<1 %
11	Kaul, Corina R.. "Using Structural Equation Modeling to Examine the Relationships Between Environmental Characteristics, Intrapersonal Characteristics, and Adult Numeracy Achievement.", Baylor University, 2019 Publication	<1 %
12	www.researchgate.net Internet Source	<1 %
13	www.ssoar.info Internet Source	<1 %
14	ideas.repec.org Internet Source	<1 %
15	iuslaboris.com Internet Source	<1 %
16	www.piaacgateway.com Internet Source	<1 %
17	Umurzakova, Tolganay. "Effects of Skills Mismatches on Job Satisfaction in Kazakhstan: Evidence from PIAAC Data.", M.	<1 %

	Narikbayev KAZGUU University (Kazakhstan), 2023 Publication	
18	www.scilit.net Internet Source	<1 %
19	Anke Grotlüschen, Christopher Stammer, Thomas J Sork. "People who teach regularly: What do we know from PIAAC about their professionalization?", Journal of Adult and Continuing Education, 2020 Publication	<1 %
20	econpapers.repec.org Internet Source	<1 %
21	"Large-Scale Cognitive Assessment", Springer Science and Business Media LLC, 2020 Publication	<1 %
22	d-nb.info Internet Source	<1 %
23	issuu.com Internet Source	<1 %
24	pdfs.semanticscholar.org Internet Source	<1 %
25	www.mdpi.com Internet Source	<1 %
26	9pdf.net Internet Source	<1 %

27	Raija Hämäläinen, Kari Nissinen, Joonas Mannonen, Joni Lämsä, Kaisa Leino, Matti Taajamo. "Understanding teaching professionals' digital competence: What do PIAAC and TALIS reveal about technology-related skills, attitudes, and knowledge?", Computers in Human Behavior, 2021 Publication	<1 %
28	hanushek.stanford.edu Internet Source	<1 %
29	labourmarketresearch.springeropen.com Internet Source	<1 %
30	Submitted to Teaching and Learning with Technology Student Paper	<1 %
31	open.metu.edu.tr Internet Source	<1 %
32	David J Maume, Orlaith Heymann, Leah Ruppanner. "National Board Quotas and the Gender Pay Gap among European Managers", Work, Employment and Society, 2019 Publication	<1 %
33	Margaret Becker Patterson. "Assessed Numeracy Skills and Skill Use of Adults With Learning Disabilities in PIAAC", Learning Disability Quarterly, 2022	<1 %

Publication

34	community.lincs.ed.gov Internet Source	<1%
35	hdl.handle.net Internet Source	<1%
36	Takatoshi Ito, Kazumasa Iwata, Jong-Wha Lee, Colin McKenzie, Shujiro Urata. "Labor, Health and Education in Asia-Analysis of Micro Data: Editors' Overview", Asian Economic Policy Review, 2017 Publication	<1%
37	Submitted to University of Illinois at Urbana-Champaign Student Paper	<1%
38	"Education and Mobilities", Springer Science and Business Media LLC, 2020 Publication	<1%
39	Submitted to KAZGUU University Student Paper	<1%
40	Katie B. Biello, Pablo K. Valente, Daniel Teixeira da Silva, Willey Lin et al. "Who prefers what? Correlates of preferences for next-generation HIV prevention products among a national U.S. sample of young men who have sex with men", Journal of the International AIDS Society, 2023 Publication	<1%

41	repository.library.carleton.ca Internet Source	<1%
42	www1.oecd.org Internet Source	<1%
43	sciendo-parsed-data-feed.s3.eu-central-1.amazonaws.com Internet Source	<1%
44	Hill, Askia. "An examination of factors of engineering epistemology development in electrical and computer engineering students.", Proquest, 2016. Publication	<1%
45	stats.oecd.org Internet Source	<1%
46	www.statcan.gc.ca Internet Source	<1%
47	Daiji Kawaguchi, Takahiro Toriyabe. "Measurements of skill and skill-use using PIAAC", Labour Economics, 2022 Publication	<1%
48	www.nesa.com.au Internet Source	<1%
49	www.pedocs.de Internet Source	<1%

50	Julia Gorges, Débora B. Maehler, Tobias Koch, Judith Offerhaus. "Who likes to learn new things: measuring adult motivation to learn with PIAAC data from 21 countries", Large-scale Assessments in Education, 2016 Publication	<1 %
51	ebin.pub Internet Source	<1 %
52	econstor.eu Internet Source	<1 %
53	moam.info Internet Source	<1 %
54	tcg.uis.unesco.org Internet Source	<1 %
55	vdocuments.mx Internet Source	<1 %
56	www10.iadb.org Internet Source	<1 %
57	Bram De Wever, Raija Hämäläinen, Kari Nissinen, Joonas Mannonen, Lisse Van Nieuwenhove. "Teachers' problem-solving skills in technology-rich environments: a call for workplace learning and opportunities to develop professionally", Studies in Continuing Education, 2021 Publication	<1 %

58	Ding, Qi. "Influence of social cognitive variables on the career exploratory behaviors of African American undergraduate STEM-intensive agricultural sciences majors at Historically Black Land-Grant Institutions.", Proquest, 2015. Publication	<1 %
59	Hikaru Komatsu, Jeremy Rappleye. "Refuting the OECD-World Bank development narrative: was East Asia's 'Economic Miracle' primarily driven by education quality and cognitive skills?", Globalisation, Societies and Education, 2019 Publication	<1 %
60	ecommons.udayton.edu Internet Source	<1 %
61	repositorio.minedu.gob.pe Internet Source	<1 %
62	www.zora.uzh.ch Internet Source	<1 %
63	Jong-Wha Lee, Do Won Kwak, Eunbi Song. "Can older workers stay productive? The role of ICT skills and training", Journal of Asian Economics, 2022 Publication	<1 %
64	Simon Ellis. "Current International Data for Tvet and their Limitations", PROSPECTS, 2005	<1 %

Publication

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		