College of Business and Economics **African Center of Excellence in data science**

Analysis of job vacancies and skills mismatch in Rwanda using data mining techniques, (2019-2021)

BY

MUKUNDABANTU Jean Marc

Registration Number: 221000700

A Thesis submitted in partial fulfilment of the requirements for the degree of Master of Data Science in Demography

University of Rwanda, College of Business and Economics

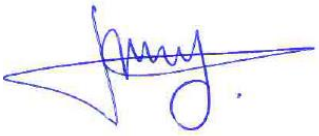
Supervisor: Dr.Innocent NGARUYE

Kigali, June 2023

Declaration

I declare that this thesis entitled **Analysis of job vacancies and skills mismatch in Rwanda using data mining techniques** is the result of my original work and has not been presented for any award at the University of Rwanda or in any other institution.

Names



Signature

Approval sheet

This thesis entitled **Analysis of job vacancies and skills mismatch in Rwanda using data mining techniques** written and submitted by MUKUNDABANTU Jean Marc in partial fulfilment of the requirements for the degree of Master of Science in Data Science majoring in Demography is hereby accepted and approved. The rate of plagiarism tested using Turnitin is 3 %, less than 20% accepted by the African Centre of Excellence in Data Science (ACE-DS).

Dr. Innocent NGARUYE

_____

Supervisor

_____

Head of Training

Dedication

This thesis is dedicated to:

My beloved wife and children for their immeasurable love, sacrifices, their tenderness, their support, and their prayers throughout my studies,
All the workmates for their encouragement,

This work is dedicated.

Acknowledgements

I would like to thank the National Institute of Statistics of Rwanda (NISR) for allowing me to study for a Master of Science in data science while working. This thesis could not have been completed successfully except after a combination of efforts from different key players through their encouragement and financial support.

I thank Dr Innocent NGARUYE, my supervisor, for his guidance, day to day follow-up for the accomplishment of this thesis.

Many thanks go to the Department of Demographic and Social Statistics at the National Institute of Statistics of Rwanda for their kind assistance in providing needed data and information especially the Labour Force Survey to accomplish this thesis.

Special thanks go to the University of Rwanda especially the African Centre of Data Science for the valuable and professional knowledge I gained from them. Also, to the classmates for their kindness while studying.

I extend my gratitude to my beloved family for their encouragement and their patience during the study.

To all who contributed in one way or the other to this; May God bless them.

Abstract

The current study aims at identifying skills mismatch and skills needed in the labour market in Rwanda in the year 2021-2022 which is a result of declining labour productivity.

The researcher used both qualitative and quantitative data to describe skills mismatch in Rwanda. Primary data were used using web-scraping to obtain advertised jobs in Job in Rwanda or on MIFOTRA Portal.

Machine learning algorithms were used to predict skills mismatch and the best models to estimate skills mismatch were both Decision Tree and Random Forest with an overall precision of 79 percent, according to the accuracy and degree of precision of the machine learning technique that was used to make this choice.

The place of residence (Urban/Rural), economic activity, the age of the employee, and the type of education pursued by the employee were found to be factors associated with skills mismatch. Employed persons with high education are more inclined to do any job or occupation for low-skill employment, which could impede output and employee happiness in a particular organization. There was a sizeable fraction of the employed population part-time in low-skill jobs even if they would prefer to work full-time in an occupation that matched their skill set.

Forty-two per cent of the employed population declared themselves as having occupations that match with the level of education required (self-declaration). The incidence of skills mismatch by occupation level has shown that persons employed in Service and sales workers are relatively under-skilled (60 percent) as

compared with other main occupation groups. The incidence of skills mismatch was 37 percent in the agriculture sector and 29 percent craft and related.

The mismatch among unemployed was 7.9% meaning that they lack the skills necessary for work given the amount of education they have acquired and the positions they are qualified for. The population of the Labour force was another part that is match affected by skills mismatch and this has an impact on the production of the country.

When workers lack the necessary abilities to do their jobs more effectively and profitably, it has been discovered that the skills mismatch affects output among them as well.

Key words: ***Job, Employment, Labour Force, Skills miss-match, under-empoloyment***

Table of contents

Acknowledgements v

Abstract vi

Table of contents viii

Lis of Acronyms xiii

1.1 Definition of key concepts xiv

Chapter 1 Introduction ..... - 1 -

1.1 Background ..... - 1 -

1.2. Problem formulation ..... - 2 -

1.3. Study objectives ..... - 4 -

1.3.1. General objective ..... - 4 -

1.3.2. Specific objectives ..... - 4 -

1.3.3. Research Hypothesis ..... - 4 -

1.4 Research Questions ..... - 4 -

1.4. Significance of the study ..... - 4 -

1.6 Scope and Limitation of the Study ..... - 5 -

1.7 Ethical consideration ..... - 5 -

1.8 Organization of the Study ..... - 5 -

Chap 2 Literature review and conceptual framework ..... - 7 -

2.1 Conceptual frame for Estimating the number of advertised jobs and Skills mismatch analysis. ..... - 10 -
Chapter three: Research Methodology ..... - 12 -
3.1 Study approach ..... - 12 -
3.2 Research instruments and data-gathering procedures ..... - 12 -
3.3 Materials ..... - 13 -
3.4 Study design ..... - 14 -
3.4.1 Specific objective achievements ..... - 14 -
3.4.2 Study variables ..... - 15 -
3.4.3 Outcome variable ..... - 16 -
3.5 Statistical analysis ..... - 16 -
3.5.1 Sample Size calculation ..... - 17 -
3.5.2 Sampling techniques ..... - 17 -
3.5.3 Analytical Framework ..... - 18 -
3.6. Evaluation of Performance for Predictive Models ..... - 20 -
3.6.1. Confusion Matrix ..... - 20 -
3.6.2. Accuracy ..... - 21 -
3.6.3 Precision ..... - 22 -
3.6.4 Recall ..... - 22 -
3.6.5 F1 score ..... - 23 -
Chapter 4 Results discussion ..... - 23 -

4.1 Introduction ..... - 24 -
4.2 Explanatory analysis ..... - 24 -
4.2.1 Main occupation and level of education required ..... - 24 -
4.2.2 Incidence of Mismatch by Major Occupation Group ..... - 26 -
4.2.3 Dissimilarity between employed and unemployed/outside labour force ..... - 27 -
4.2.4 Analysis of skills mismatch ..... - 29 -
4.3 Analysis of Skills miss match using data mining techniques ..... - 31 -
4.3.1 Job advertised ..... - 31 -
4.3.2 Main significant factor from multinomial logistic regression ..... - 32 -
4.3.3 Predictive models ..... - 33 -
4.4 Discussion ..... - 43 -
Chapter 5 Conclusion and Recommendations ..... - 45 -
5.1 Conclusion ..... - 45 -
5.2 Recommendations ..... - 46 -
REFERENCES ..... - 47 -

Lists of figures

Figure 1: Conceptual Framework ..... - 11 -
Figure 2:Confusion matrix ..... - 21 -
Figure 3: Main occupation and level of education required ..... - 25 -
Figure 4: Occupation group by a Skills mismatch ..... - 27 -
Figure 5: General Skills mismatch ..... - 29 -
Figure 6: Skill mismatch with respect to sex of employed person ..... - 30 -
Figure 7: Skills according to Province ..... - 30 -
Figure 8: Confusion matrix for decision Tree ..... - 35 -
Figure 9: Area Under the Curve and the Receiver Operating Characteristics (AUC-ROC) for Decision tree ..... - 36 -
Figure 10: Important Futures from random forest ..... - 37 -
Figure 11: Confusion Matric for the Random Forest ..... - 39 -
Figure 12: Area Under the Curve and the Receiver Operating Characteristics (AUC-ROC) for Random Forest ..... - 40 -
Figure 13: Confusion Matrix for he KKN ..... - 40 -
Figure 14: Confusion Matrix for Naïve Bayes ..... - 42 -

List of Tables

Table 1:The selected independent variables used for analysis .....- 15 -
Table 2: Level of education attained by the required level of education.....- 26 -
Table 3: Mismatch rate of unemployment).....- 28 -
Table 4: Mismatch rate of employed population .....- 28 -
Table 5: Advertised job through job in Rwanda and MIFOTRA Portal 2021 2022.....- 31 -
Table 6: Main Significant factors (multinomial approach).....- 32 -
Table 7: Confusion Matric for general prediction .....- 33 -
Table 8: Level of precision prediction for Random forest.....- 38 -
Table 9: Precisions and recall for KNN.....- 40 -
Table 10: Precision and recall for Naïve Bayes Model .....- 41 -
Table 11: Summary of the predicted model.....- 42 -

Lis of Acronyms

GoR: Government of Rwanda

ILO: International Labour Occupation

ISCO: Internationa Standards for Classification of Occupation

ISIC: International Statandars of Inductry Classification

LFS: Labour Force Survey

NISR: National Institute of Statistics of Rwanda

RDB: Rwanda Government Board

ACE-DS: African Centre of Excellence in Data Science

1.1 Definition of key concepts

Job: According to the International Labour Organization office Job is defined as a set of tasks and duties performed, or meant to be performed, by one person for a single economic unit

Employment: Recent international standards define people of working age as being employed if, during a short reference period (usually seven days or one week); (i) they performed any work (even for just one hour) for pay or profit, in cash or in kind; or (ii) they were connected to a job or had an enterprise from which they were "temporarily" absent. In contrast to the previous definition, the new one excludes people who were working for subsistence through activities like agriculture (ILO, Resolution concerning statistics of the economically active population, employment, unemployment and 1982).

The difference with the previous definition is in that the current one no longer covers people who were working for subsistence purposes, such as subsistence farming. As a result, the phrase "work for family gain" is no longer used in the definition. In accordance with the System of National Accounts (SNA) definition of economic activity performance, the prior standards defined the working age population and employment (AfDB 2012).

Labour Force Survey: A survey that is intended to measure the three main labour market categories "employed," "unemployed," and "not in the labour force "primarily by classifying people according to their activities, based on priority rules and the "activity principle" what they were actually doing during a given reference period (ILO, 19th ICLS Guideline 2013).

Skills: The aptitude, either innate or learned, to put information into practice that has been learnt through experience, study, practice, or teaching and to carry out the activities and duties necessary for a particular employment. This can be in terms of technical skills, Transferable skills or Basic skills. (ILO, International Standards of Classification of Occupation 2008).

Chapter 1 Introduction

1.1 Background

Developing the country's plan for Rwanda with the support of Vision 2020, the nation was to transit from an agrarian to a middle-class, knowledge-based civilization. The annual creation of off-farm jobs is a crucial part of this approach. The economy produced 206,000 additional employees between 2017 and 2018. In the next step, the National Strategy for Transformation (NST1) has set the target of creating 1.5 million decent and productive jobs by 2024, which amounts to 214,00 jobs per annum (Rwanda 2017). To reach these figures, employment needs to be regarded as a priority policy issue, which can only be addressed through re-orientation and consolidation of different policy interventions, such as support for the business environment, active labour market programmes or investment strategies. Furthermore, it has to be considered to what extent the newly created jobs qualify as productive jobs.

The Rwandan state has assumed a significant role in directing the nation's economic growth processes in recent years. Agro-processing, Construction, Light Manufacturing, Meat and Dairy, Leather, Textiles and Garments, Horticulture, Tourism, Knowledge-Based Services, Value Addition and Processing of Mining Products, Creative Arts, Aviation, Logistics, and Transportation have all been identified as priority sectors with high potential for economic growth and employment. According to indicators collected from job adverts or job vacancies in social media, Rwanda has a variety of potential areas for job development, as seen by the variety displayed by the prioritized sectors (AfDB 2012).

However, this path of economic development faces several challenges, especially as private sector enterprises are still very weak. At the same time, there is an over-reliance on the economic activities of large domestic investment-holding groups or large foreign corporations, which shapes also the development of sectoral/regional labour markets and associated skills needs. Despite extensive investments in education, health and infrastructure, skills gaps or shortages are reported for several sectors (RDB, 2012).

Rwanda through the Rwanda Development Board’s National has developed a strategy for skills Development and promotion of Employment (2019 – 2024) to highlight available skills for both general and TVET education that do not fully respond to labour market needs. As employment promotion is among key government priorities, public and private institutions involved in skills and

talent development must ensure market relevance of training programs, demand-led education initiatives, and job placement approaches. Hence, there is a need for a partnership between the Government, private sector, and universities in profiling required skills by industries and skilling the Rwandan labour force to exploit untapped employment opportunities in the labour market.

1.2. Problem formulation

Globally, millions of jobs are advertised on different channels. In recent years, academic and policy discussions have shifted from an emphasis on overeducation or skill shortages to a more nuanced review of imbalances that incorporates skill gaps and skill underutilization. Those positions might be linked with specific skills that are needed in the labour market. Sometimes people consider talent shortages to be more essential than skill gaps. In fact, certain predictions tend to imply that over the next ten years, the supply of skills may exceed the need, leading to increased rates of over skilling (Amador 2006). Nevertheless, policy concerns today place more of a focus on the need to address the negative effects of underusing skills than on the requirement to reduce the supply of educated persons (Cappelli 2014). Thoughts that high levels of education are required to meet the long-term needs of the labour market and protect against rapid, technically biased change and competitiveness in the 21st century counter concerns about overeducation (Allen J., van der Velden R. 2021). Making sure that work practices develop in a way that effectively utilizes the skills of more educated people is important to prevent skill atrophy and lost opportunities to boost productivity. Unless work practices alter to promote more effective utilization of talents, an increase in highly educated people may eventually push out lower educated individuals and worsen the problem of skill underutilization (Arsenau, D. and Epstein, B. 2014).

In developing countries, High rates of workforce participation are typically associated with insufficiently high-quality jobs. An income impact can be used to explain this, where households have to put in more labour to reach a certain level of resources. This shows that developed countries may be forcing more people to participate in the labour market than is normal for its level of development due to low average earnings, a large concentration of occupations in low-productivity sectors, and a lack of appropriate wage employment alternatives (Adalet McGowan M. and Andrews 2015). However, it is crucial to distinguish between structural effects that result in higher labour participation for a clear interpretation of this labour market symptom. In contrast to the substantial growth in demand for skilled labour during the prosperous years prior to 2014, Rwanda's need for low-skilled labour has decreased

over the past ten years. This shows that Rwandan businesses tended to hire more competent individuals during periods of robust employment growth and were more likely to fire low-skilled workers during periods of stagnant employment growth. In the past three years, the majority of advertised positions have been for highly skilled workers who are not required to have both technical and soft abilities (Autor, D., Levy, F., & Murnane, R., 2003,).

Employers need cooperation from education systems in order to develop talent pools with skills relevant to current business needs because they are unable to address the skills gap problem on their own. While some businesses have worked closely with educational partners to construct talent pipelines, the majority rely on educational systems to independently identify and develop the capabilities that the future workforce will need.

To that extent, the research question is: ***"What are the estimated jobs advertised and skills required in the last three years in Rwanda"?***

1.3. Study objectives

1.3.1. General objective

The aim of this study is to conduct a deep statistical analysis of skills mismatch in Rwanda at all levels.

1.3.2. Specific objectives

The specific objectives of the study are:

- To estimate the number of jobs and skills needed in Rwanda's Labour Market.
- To analyze skills mismatch in Rwanda.

1.3.3. Research Hypothesis

i. Null Hypothesis

- Advertised jobs are not lacking any skills in Rwanda
- There is no skills match in employment in Rwanda.

ii. Alternative Hypothesis

- There is a skills gap in advertised jobs in Rwanda.
- Skills mismatches are in different economic sectors in Rwanda.

1.4 Research Questions

This study was designed to answer different questions, which included the following questions:

- What are the estimated jobs advertised and skills required in the last three years in Rwanda?
- What is the rate of skills mismatch in Rwanda in 2021?

1.4. Significance of the study

Theoretically, our study was vital and of importance, since the study enriched the existing literature on skills mismatch and brought out contradictions in the theories and in the process exemplify the gaps in the theories in Rwanda about skills gap and skills mismatch. Academically, our study has been

important since the findings of the study enhanced the understanding of the study and open up new areas for further academic research to the academic fraternity. The policymakers were informed on the knowledge of skills mismatch, existing and needed skills in the labour market, and furthermore, provided guidance for sustainable curriculum development, and jobs that may exist or are already in the process.

Furthermore, this study is useful to ACE-DS students, line Ministries, agencies, and other researchers as it has been added to existing literature at the university.

1.6 Scope and Limitation of the Study

The study on skill mismatches was based on normative categorization, whereby classification was based on the level of education that corresponds to the occupation category where employed workers are. The study, therefore, focuses on the studying education mismatch according to the occupation but is limited to the skills needed to perform any task or duty in each employment. The study was also limited to the data collected from the Labour force survey and the data published online o jobs in Rwanda and MIFOTRA Portal(e-recruitment) there for describing used may be influenced by some duplicates where jobs have been published in two different formats.

1.7 Ethical consideration

This study was conducted exclusively using primary and secondary and existing datasets which have been authorized by the concerned parties and therefore, do not require an institutional review board.

However, as the present study used a database from LFS 2019-2021 an online registration was observed, and prior approval has been secured before accessing the database, as shown in the NISR Program, Job in Rwanda and MIFOTRA authorization letter attached.

1.8 Organization of the Study

The study is divided into five chapters, which are the introduction, which includes a statement of the study's history, the problem statement, its aims, its research questions, and its justification. The second chapter is a review of previous studies that have been done on skills mismatch, including but not limited to those that have looked at theoretical reviews, critical research gaps, and conceptual frameworks.

The third chapter discusses the study's approach. It goes into detail about the study's design, the data source, data collection methods, model specification, the variables of interest chosen for analysis to meet the research's goal, the prediction model, as well as performance evaluation for predictive models. The whole explanation of research findings and their interpretation is found in chapter four. The study closes with research-based recommendations in the final section, which is about the conclusion.

Chap 2 Literature review and conceptual framework

The goal of analysis of skill mismatching is to identify the issue of a mismatch between skill demand and skill supply, or overeducation. This book uses overeducation to provide knowledge about the notion of skills. Methods for Addressing Skills Mismatch in the Workplace: Job analysis, worker self-assessment, and skills matching are the three primary different measures of job mismatch that have been used to quantify the degree of skills-mismatch.

According to Brun-Schammé, A. and Rey, M. (2021), it is crucial to keep in mind that a disparity in skill levels within a profession is not always indicative of a skills mismatch but might also be related to an employee's performance. Additional training is crucial when an individual's performance is subpar. Age is a crucial impact in situations when there is an apparent skills mismatch, as we have been able to demonstrate. As people get older, it becomes less likely that they will be able to change careers, which emphasizes the importance of continuing education in these situations. (Brun-Schammé 2021).

In their working paper titled "A new approach to skills mismatch," Brun-Schammé et al. (2021) noted that because the labour market has not been able to determine an individual's ability levels, the traditional mechanisms for matching labour supply and demand have not been effective. Many people who work in the observed occupations appear to have different abilities from their occupation proxy, but not from their training profile. In their working paper titled "A new approach to skills mismatch," Brun-Schammé et al. (2021) noted that because the labour market has not been able to determine an individual's ability levels, the traditional mechanisms for matching labour supply and demand have not been effective. Many people who work in the observed occupations appear to have different abilities from their occupation proxy, but not from their training profile (Brun-Schammé, A. and Rey, M. 2021).

According to several studies from the European Community (2018), skills mismatch refers to the sub-optimal application of a person's skills in their line of work and can cause employee dissatisfaction as well as a hindrance to productivity growth. In our opinion, a disparity in skill level within a profession is insufficient to conclude that a skills mismatch exists. The quantification of skills mismatch must therefore be based on the factors causing this imbalance, as skills mismatch is caused by a discrepancy between the supply and demand of labour (European Commission 2018).

Adalet (2015) in their article on labour productivity and labour market mismatch from 2015, McGowan and Andrews came to the conclusion that when an individual's performance is subpar, additional

training is crucial. Age is a crucial impact in situations when there is an apparent skills mismatch, as we have been able to demonstrate. Age reduces the likelihood of a professional shift, which emphasizes the importance of continuous education in these situations (Adalet McGowan M. and Andrews 2015).

Francois Rycx (2012) conducted research on the direct effects of educational mismatch on firm productivity using an employer-employee panel, and he concluded that overeducated workers are more productive throughout their careers due to additional skills and capabilities acquired through education, whereas undereducated workers either succeed in making up for their lack of productivity by accumulating more work experience and training or end up in less demanding jobs as they age. (Stephan Kampelmann , François Rycx 2012).

Wiczer (2015) discovered that younger, more educated cohorts have less access to high-skill employment than older workers. This could simply be due to the older, more seasoned individuals entering the workforce at a time when there were more high-skilled positions available. This implies that a failure to take into consideration cohort factors weakens the explanatory power of existing studies looking at the connection between overeducation and informal human capital accumulation (Wiczer 2015).

Mc Guinned (2011) emphasized the methods for boosting output by hiring staff who are more educated. When analyzing the skills gap in the UK, for instance, he noted that: this strategy implicitly assumes that there is an excess demand for tertiary education or that businesses hiring more educated workers will improve their production techniques to take advantage of those additional skills. (S. a. McGuinness 2011)

Gullem (Sala 2011), In his book on approaches to a skills mismatch in the labour market, wrote that job analysis (JA) is the systematic assessment of the necessary level of education for job titles by qualified job analysts through an occupational classification of the level of qualifications needed to carry out a specific job.

According to Lourdes B. A.'s (2006) study, "Education and competence mismatches: Job satisfaction consequences for workers," workers' levels of professional competence which may be lower or higher than those necessary for their current jobs are determined by their abilities, skills, attitudes, and knowledge. When this occurs, there is a mismatch between the competence levels required for the job and the workers: workers are under-competent when their competence levels are lower than those in their jobs, and workers are over-competent when their competence levels are higher than those in their

jobs. Mismatches in professional competence have economic ramifications because they may affect compensation and the quality of the fit between workers' capacities and the demands of the tasks, they do (Amador 2006).

Groot et al. (Green, 1996) discovered that younger, more educated cohorts have less access to high-skill employment than older workers. This could simply be due to the older, more seasoned individuals entering the workforce at a time when there were more high-skilled positions available. This implies that a failure to take into consideration cohort factors weakens the explanatory power of existing studies looking at the connection between overeducation and informal human capital accumulation.

For instance, Becker(1987), who adhered to this theory, stated that i) education, formal training, and unpaid work experience all help employees gain abilities that increase production, and ii) salary differences are a reflection of productivity differences. As a result, from how the latter affects salaries, it may be possible to determine how over and under-skilling affect productivity. (Becker 1964).

A three-step methodology was used by Tsang (1987) to demonstrate the detrimental effects of overeducation on business production over two years of observation in the years 1981 to 1983 using employee data from 22 US facilities owned by Bell corporations. He estimated the impact of overeducation on job happiness using data from specific employees and concluded that job satisfaction served as a stand-in for employee work effort (Tsang 1987).

Freeman (1976) stated in his essay on "the Overeducated American" that skills mismatch in the labour market (SMLM) is characterized by the fact that an individual's levels or types of talents are insufficient in light of a certain job demand. Richard Freeman introduced this issue to researchers by using the idea of overeducation in a novel way (Freeman, R. B. 1976)

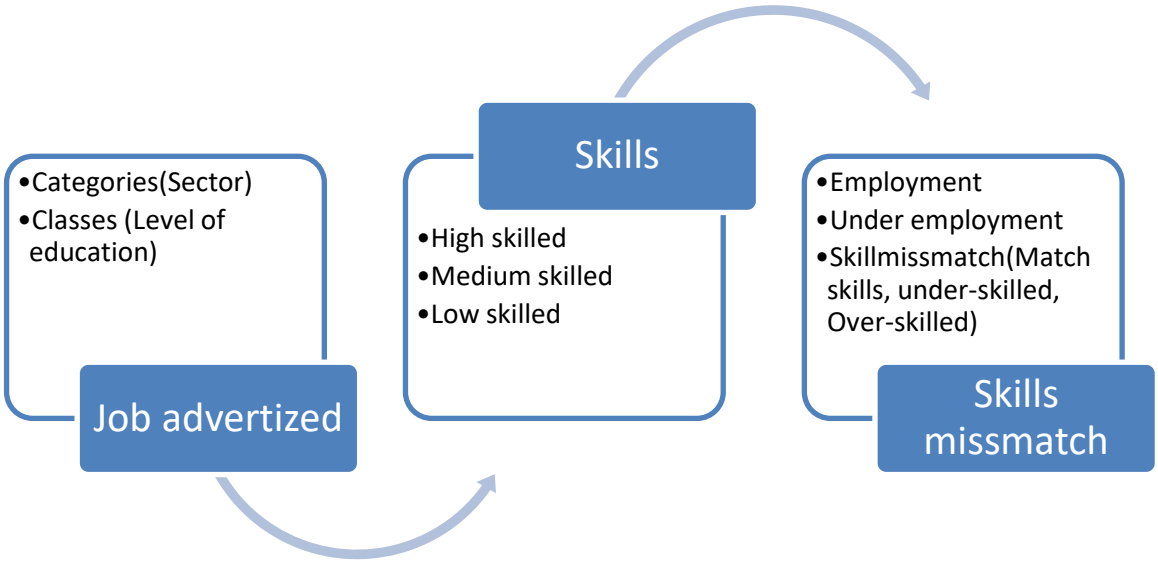
Thurow (1976) published his theory of job competition. He is of the opinion that employment prospects are closely tied to educational attainment on the job market, and that a higher education level leads to more employment opportunities. In his book, he proposes that all employees be sorted according to a particular queue based on the costs of training for businesses. He discovered a chance with a smaller budget that benefits people with training because they need less work-related training. Additionally, he thinks that social backgrounds rather than advanced education are more important in employment battles than schooling. Employees from privileged backgrounds, for instance, frequently have access to superior work environments. According to these presumptions, the salary is solely determined by the

needs of the labour market and the presence of overeducation, which is a result of labour market competitiveness. As a result, the financial benefits of excess and deficit schooling, such as premiums and fines, should be negligible (Thurow 1975).

2.1 Conceptual frame for Estimating the number of advertised jobs and Skills mismatch analysis.

The study looked at various job listings and highlighted industries that needed workers or employed more people than other industries. As mentioned in the literature review, the specific abilities needed have also been examined to spot skill mismatches and gaps based on their theoretical and practical significance.

Figure 1: Conceptual Framework



A conceptual framework is a theoretical structure that helps researchers understand, analyze, and explain complex phenomena. In our case, we want to explore the relationship between skills mismatch (dependent variable) and several independent variables: occupation, education level, age, and sex.

The analysis of skills mismatch begins with an assessment of the skills demanded by the labor market, often advertised in job postings. These demanded skills are then compared with the skill levels required for a specific occupation or job, typically categorized as high, medium, or low skill levels. This comparison can reveal whether there was a skills mismatch among employed individuals, categorizing them into different groups, including the underemployed, those in positions matching their educational background, or those lacking the skills necessary for the labor market.

Chapter three: Research Methodology

This chapter provides an in-depth overview of the approaches and strategies employed to examine the research concerns within the respective field. It encompasses the study area, source of data, research design, target population, data collection procedures, and statistical analysis methods employed in this study.

3.1 Study approach

The research methodology employed in this study was quantitative in nature. According to Grove and colleagues (Daniel 2017), quantitative research is a structured and unbiased approach that systematically analyzes numerical data to generate novel insights about the world. Quantitative methods emphasize understanding and breaking down phenomena into constituent elements to assess the resulting outcomes (Lehti and Lehtinen 2005).

Quantitative research can encompass various types, such as descriptive, exploratory, correlational, quasi-experimental, or experimental. In this study, a descriptive approach was adopted to present the variables by illustrating their frequencies. Additionally, the study had an exploratory nature, as it aimed to investigate the variables and their potential associations. Furthermore, this study was supervised learning as different classification used in data analysis was based on readymade assumption.

3.2 Research instruments and data-gathering procedures

The study used different data collection techniques including:

- Web scraping

In the field of data science, web scraping refers to the procedure of gathering and analyzing unprocessed information from the Internet. The Python community has developed highly robust tools for web scraping that enable the collection and examination of data from various websites.

The steps involved in web scraping:

Identify the target website: Determine the website that wanted to extract data from after reading the website's terms of service or any legal restrictions on scraping before proceeding. After we selected a scraping tool/library suitable for study. We Inspected the website structure and identified the specific

elements (tags, classes, or IDs) that contain the data needed. After this, the researcher wrote the scraping code using the chosen web scraping library to extract the desired data and process and store the data in a CSV file to perform further analysis.

- Literature review

A thorough literature review was undertaken on existing studies, documents and articles on skills mismatch in general, paying particular attention to the skills mismatch and labour underutilisation in Rwanda

- Dataset from the NISR website

The annual dataset of the labour force survey has been extracted and used for the analysis of job mismatch. An algorithm has been developed in order to classify a person into categories of Skills that they possessed.

Primary data were collected using data mining techniques where data on advertised jobs that have been derived from jobs in the Rwanda Portal using web scraping techniques and Secondary data were obtained from the National Institute of Statistics of Rwanda from a series of LFS conducted by NISR in 2021.

The analysis therefore highlighted the estimated number of jobs advertised in Rwanda via the target of job creation in Rwanda and later highlighted the skills needed according to those jobs in Rwanda and the e-recruitment platform managed by MIFOTRA. Primary data have been processed, categorized and classified in order to collect indicators on the estimated advertised job from July 2021 up to June 2022.

3.3 Materials

Both Primary and secondary data from the LFS 2019 to 2021 datasets have been used to reflect relevant issues in employment and skills in Rwanda. The technique that has been used to collect these data have adopted to match the subject understudy.

The formula provided below is utilized to compute the educational mismatch rate EM_U for unemployment:

$$EM\_U = \frac{1}{2} * \sum_{k=1}^{n} \left| \left(\frac{Employed\ with\ education\ K}{Total\ employed} - \frac{Unemployed\ with\ education\ K}{Total\ unemployed} \right) \right| \qquad (1)$$

The following formula is employed to determine the educational mismatch rate for individuals who are out of the Labour Force (POLF) noted by EM_POLF:

$$EM\_POLF = \frac{1}{2} * \sum_{k=1}^{n} \left| \left(\frac{Employed\ with\ education\ K}{Total\ employed} - \frac{OLF\ with\ education\ K}{Total\ OLF} \right) \right| \qquad (2)$$

3.4 Study design

The present study was based on both primary and secondary cross-sectional data that were analysed to inform us about the job advertised and the skills required in Rwanda from 2019 to 2021.

Some of these data are in the public domain and are available on the NISR, e-procurement portal and Job in Rwanda website from which we have an authorizing letter to explore these data. This entailed the consolidation, compilation, and/or integration of existing research findings, whereas secondary research utilizes. Primary data as a source of data for analysis was used to estimate and analyzed job vacancies and to check if they are statistical correlations between some variables that determine skill mismatch in Rwanda.

3.4.1 Specific objective achievements

To achieve our study objective which is to identify skills mismatch and estimate the number of jobs and skills needed in Rwanda. The algorithm has been designed to estimate the number of jobs, existing skills and skills mismatch in Rwanda among the working-age population. The cutoff point for statistical significance has been set to be at least 0.05.

3.4.2 Study variables

In the LFS data sets, variables of personnel skills, education and employment characteristics were created reflecting the person's age, gender, residential area, as well as economic sector; to explore the evidence of correlation depending on the level of statistical significance for all analyses through comparisons. Using data sets of the LFS 2017-2021 that are available on the NISR website program site and the primary data obtained from the Job in Rwanda Portal, we proceeded with a secondary analysis on skill mismatch and potential sectors to provide employment through jobs advertised in Rwanda. A person whose level of education is not matching with Occupation, the person whose employment requires more skills than they have, and the person whose skills are not fully absorbed by the labour market have been analyzed. We used the education level and International Standard of Classification of Occupation with the related skills to analyse their association with skills. Furthermore, we employed web scraping techniques to make an estimation of the number of jobs in Rwanda that require specific skills

Table 1:The selected independent variables used for analysis

Variables	**Labels in the LFS dataset**	**Categories**
Education level	attained	0=None 1= Primary 2= Lower Secondary 3= Upper Secondary 4=University
General or TVT	TVT2	1= No level completed 2= Completed General 3= Completed TVET
Age	Age5	1= 15-19 2= 20-24 3=25-29 4= 30-34., etc
Sex	A01	1= Male 2=Female

Variables	Labels in the LFS dataset	Categories
Place of residence	Code_UR	Urban =1 Rural =2
Economic activity	Indd03	
Main_sect		1= Agriculture 2=Industry 3=Service
Field of education	indb05	1=Education 2=Humanity and art 3= Social Science business and art 4=Sciences 5=Engineering manufacturing and construction 6=Agriculture 7=Health and welfare
Province		Kigali=1 south=2 west=3 North=4 East =5
Skills		0= Skills match 1= Under-skilled 2= Over-skilled

3.4.3 Outcome variable

From the present study, the analysis of job mismatch was conducted in order to identify occupations with related skills requirements and to determine potential employment visa skills required in Rwanda. Furthermore, the study highlighted skills mismatch among employed persons in Rwanda.

3.5 Statistical analysis

The data from the four rounds of LFS were organized, and recorded according to the variable of interest, then statistical analysis was performed through python and R software. Determinants of skill mismatch were examined using the test of statistical significance of independence to determine the

possible relationship with employment. The dependent variable is skill mismatched while the independent variables were: Education level, gender, economic sector, and residential area (rural/urban) as well as any other variable that could explain our model (see table 1 above).

To describe the information that was used in the analysis, the indicators were calculated as recommended.

Number of jobs advertised: Number of Potential jobs.

The potential sector of employment: Sector that advertised/needed a lot of Jobs

***Potential skills (Skills in high demand)*:** Highly demanded skills.

***Skills mismatch*:** The percentage of the difference between the skills desired by employers and the skills possessed by individuals.

3.5.1 Sample Size calculation

Since this study was mainly secondary data analysis from LFS data 2017- 2021, the initial sample calculations for the above surveys were considered, and particularly, we were interested in persons in employment among those reported. From LFS 32017-2021; all persons in the labour force are reported.

3.5.2 Sampling techniques

In all households, the eligibility criterion for participation in the LF surveys was being a person of working age (16 years and above). The datasets from sample distributions of the labour force who responded were grouped according to the objective of the study (available at the DHS Program and NISR website).

The sampling techniques considered for this research were based on the sample size measured in the labour force survey series, which is nationally representative.
For the Labour Force Survey (LFS), a sampling frame was employed, which included a list of enumeration areas. A two-stage sampling design was implemented with the aim of enabling estimations of important indicators at the national level.

3.5.3 Analytical Framework

The categorization or the delivery of variables was based on the proposed three main categories that were used to classify skills mismatch among employed persons.

The research based on the level of skills (Skill level) according to the occupation category has decides to use 3 categories based on the minimum education level in each category obtained by computing the mode as a central tendency parameter of each occupation group and then classifying the person into Under-skilled, match Skills or Over-skilled.

Table 2: Analytica frame Classification Table

ISCO High level	None	Primary	Lower_secondary	Upper_secondary	University
Managers					
Professionals					
Technicians and associate professionals					
Clerical support workers					
Service and sales workers					
Skilled agricultural, forestry and fishing					
Craft and related trades workers					
Plant and machine operators and assemble					
Elementary occupations					

Under Skilled

Match Skills

Over Skilled

On the other hand, machine learning techniques were employed to determine the most suitable model for predicting skills mismatch in Rwanda.

Predictive modelling is the machine learning technique that is used to create, process, and validates a model that can be used to predict future events or outcomes in a given set of input datasets. The study was identified as a classification problem and algorithms were used in predictive outcomes like “K-

nearest neighbour, Random Forest, Decision tree, and Naïve Bayes" machine algorithms were applied to construct a predictive model for Skills mismatch in Rwanda.

3.5.3.1 Decision Trees Model

In most cases, the decision tree algorithm is employed to calculate a metric for each feature in the dataset. It then selects the feature that provides the highest improvement in the metric as the splitting feature at each node. This iterative process continues on each resulting subset until a stopping criterion is met, such as reaching a minimum number of samples in a leaf or reaching the maximum depth of the tree (Rokach and Maimon 2005).

3.5.3.2 Random Forest Model

Random forest is a compilation of decision trees that are typically trained using the bagging technique, where the maximum sample size is set to the size of the training set. Unlike the traditional approach of searching for the best feature when splitting a node, the random forest method selects the optimal feature from a randomly selected subset of features (Belle Fille M., J. Felicien I , Idrissa K 2023) .

3.5.3.3 K-Nearest Neighbor Model

The K-Nearest Neighbor Model is a supervised learning approach within machine learning. It is capable of addressing both classification and regression problems, with a stronger emphasis on classification tasks. By comparing the similarity between the new data and the existing categories, the K-NN model assigns the new data to the most relevant category. This model maintains all the available data and classifies a new data point based on their similarities. It works better and is more resistant to noisy training data when the training dataset is too big.

3.5.3.3 Naïve Bayes Model

Naive Bayes is a classification algorithm used for both multiclass and binary classification problems. It acquired the name "Naive" or "idiot" Bayes because it simplifies the calculation of probabilities for each class to make them more manageable. Instead of computing the probabilities of individual attribute values, it assumes that the attributes are conditionally independent given the class value. Although this assumption of attribute independence is often not realistic in real-world data, Naive Bayes still demonstrates surprisingly effective performance even when the assumption is violated (Brownlee 2019).

3.6. Evaluation of Performance for Predictive Models

The evaluation of the model was based on the performance of the model evaluation which plays a crucial role in machine learning techniques for predictive modelling, it is the process of applying different evaluation metrics to understand the performance of the model as well as its weakness and strengths so it is very important to select the right metrics for a particular predictive algorithm to get the best accurate outcome. In this study, various evaluation metrics including confusion matrix, recall, Area Under the Curve (AUC), accuracy, precision, and F1 score has been employed to assess the performance of the predictive model for skill mismatch and identify the best-performing model.

3.6.1. Confusion Matrix

In a classification problem, the confusion matrix is a descriptive term used to depict the prediction outcomes of the optimal solution. It is structured as a table, with one column representing the actual class and another column representing the predicted class. It can only be utilized when the true values for the testing data are known. This matrix is divided into two dimensions: actual values and predicted values, where actual values are the true values for the provided dataset and predicted values are those predicted by the model. The table below depicts the confusion matrix.

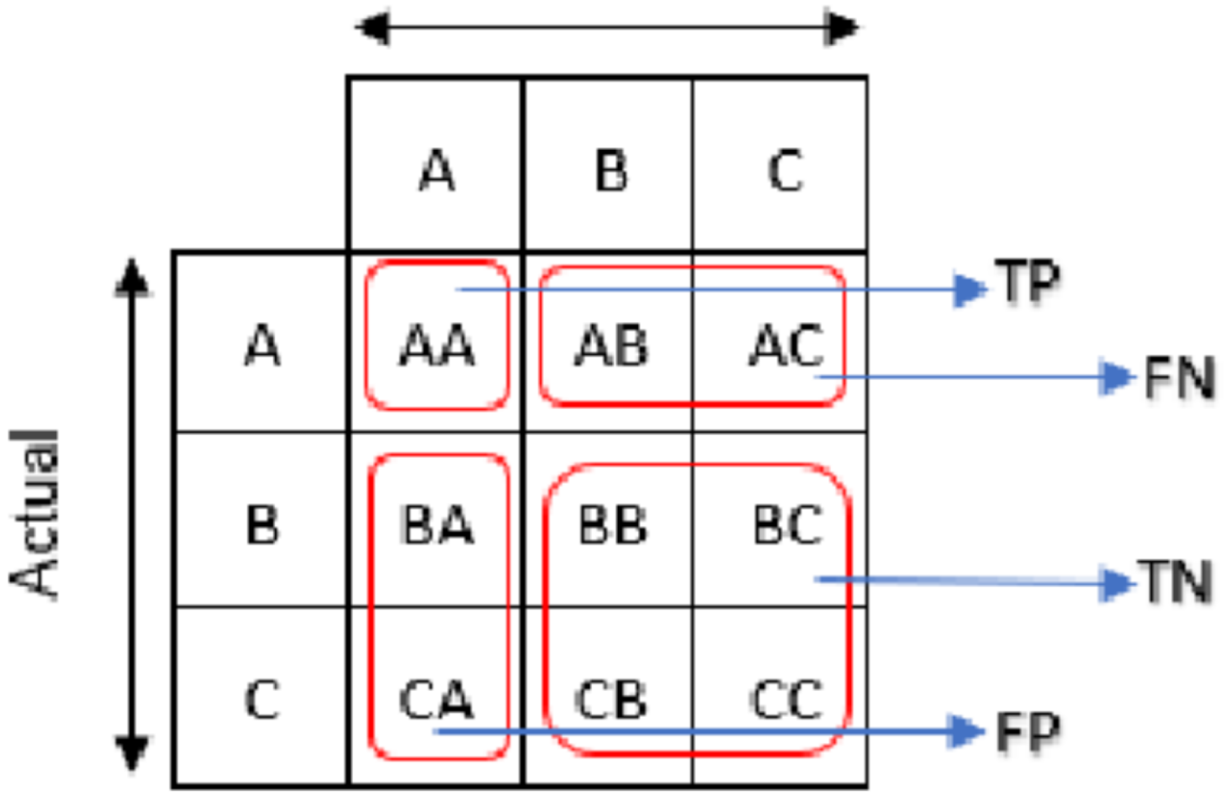


Figure 2:Confusion matrix

- **True Positive (TP):** The model correctly predicted a positive outcome, and the actual value confirmed this prediction.
- **True Negative (TN):** The model accurately predicted a negative outcome, and the actual value aligns with this prediction.
- **False Negative (FN), also known as Type-II error**: The model predicted a negative outcome, but the actual value is positive.
- **False Positive (FP), also known as Type-I error**: The model predicted a positive outcome, but the actual value is negative.

3.6.2. Accuracy

The accuracy metric is employed as one of the performance evaluation measures for a classifier model. It evaluates the overall effectiveness of the model by estimating the ratio of correctly predicted positive

and negative values in relation to the total number of cases assessed (Hossin, M.; Sulaiman, M.N. 2005):

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP} + \boldsymbol{TN}}{\boldsymbol{TP} + \boldsymbol{TN} + \boldsymbol{FP} + \boldsymbol{FN}} \tag{3}$$

3.6.3 Precision

Precision, in the context of classification, is a performance metric that assesses the accuracy of positive predictions made by a classifier. It measures the proportion of true positive predictions (correctly predicted positive instances) out of the total predicted positive instances. In other words, precision calculates the ratio of true positives to the sum of true positives and false positives. A higher precision value indicates a lower rate of false positives and reflects the classifier's ability to make precise and accurate positive predictions. Precision is particularly useful in situations where false positives need to be minimized or when the focus is on the accuracy of positive predictions (Vujovic, Ž.D. Classification 2021):

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} \tag{4}$$

3.6.4 Recall

Recall, also known as sensitivity or true positive rate, is a performance metric used in classification tasks. It measures the ability of a classifier to correctly identify all positive instances out of the total actual positive instances in a dataset. In other words, recall calculates the ratio of true positive predictions to the sum of true positive predictions and false negative predictions. A higher recall indicates a higher ability of the classifier to capture and identify positive instances correctly, making it an important measure in evaluating the completeness or comprehensiveness of a classifier's performance (Hossin, M.; Sulaiman, M.N. 2005) :

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}} \tag{5}$$

3.6.5 F1 score

The F1 score is a commonly used performance metric in classification tasks. It is a harmonic mean of precision and recall, combining both measures to provide a balanced evaluation of a classifier's accuracy. The F1 score considers both the precision (the ratio of true positive predictions to the total predicted positives) and the recall (the ratio of true positive predictions to the total actual positives). By considering both precision and recall, the F1 score provides a comprehensive assessment of a classifier's effectiveness in achieving both accurate positive predictions and capturing all positive instances (Vujovic, Ž.D. Classification 2021):

$$\boldsymbol{F_1} = \frac{\boldsymbol{2TP}}{\boldsymbol{2TP + FN + FP}} \tag{6}$$

Chapter 4 Results discussion

4.1 Introduction

The following section discussed the findings of the study conducted on the skills mismatch by showing the level of skills in Rwanda. Deeply has explained jobs advertised and their needed skills, existing skill and the skill mismatch in Rwanda.

In order to analyze skills mismatch, one must look at how skills earned through training systems are marketed. Depending on the availability of training, these skills are varied. Both a degree of education, which establishes the length of training time and expense, and an occupation level or group, which links the information obtained with a particular area of activity, can be used to map this skills mismatch. Jobs were also examined by looking at their level (the social position they achieve) and education. Thus, a variety of occupations are defined as those that are related to various social classes and educational levels. The normative method considers the degree of education (level and field of speciality) as well as the variety of occupations (social category) to analyze the relationship between education and occupation and to assess the effectiveness of training programs.

4.2 Explanatory analysis

4.2.1 Main occupation and level of education required

The finding from the analysis of skills mismatch in Rwanda using Labour Force Survey data 2021 has shown some discrepancies in the consequences of excessive education on how important people think employment was compared to how important people think success was in the workplace. Since it affects 32% of all workers, the skills mismatch is large, as was predicted. Those with high are more inclined to do any job or occupation for low-skill employment, which could impede output and employee happiness in a particular organization. One can see that a sizeable fraction of the employed population works part-time in low-skill jobs even if they would prefer to work full-time in an occupation that matched their skill set.

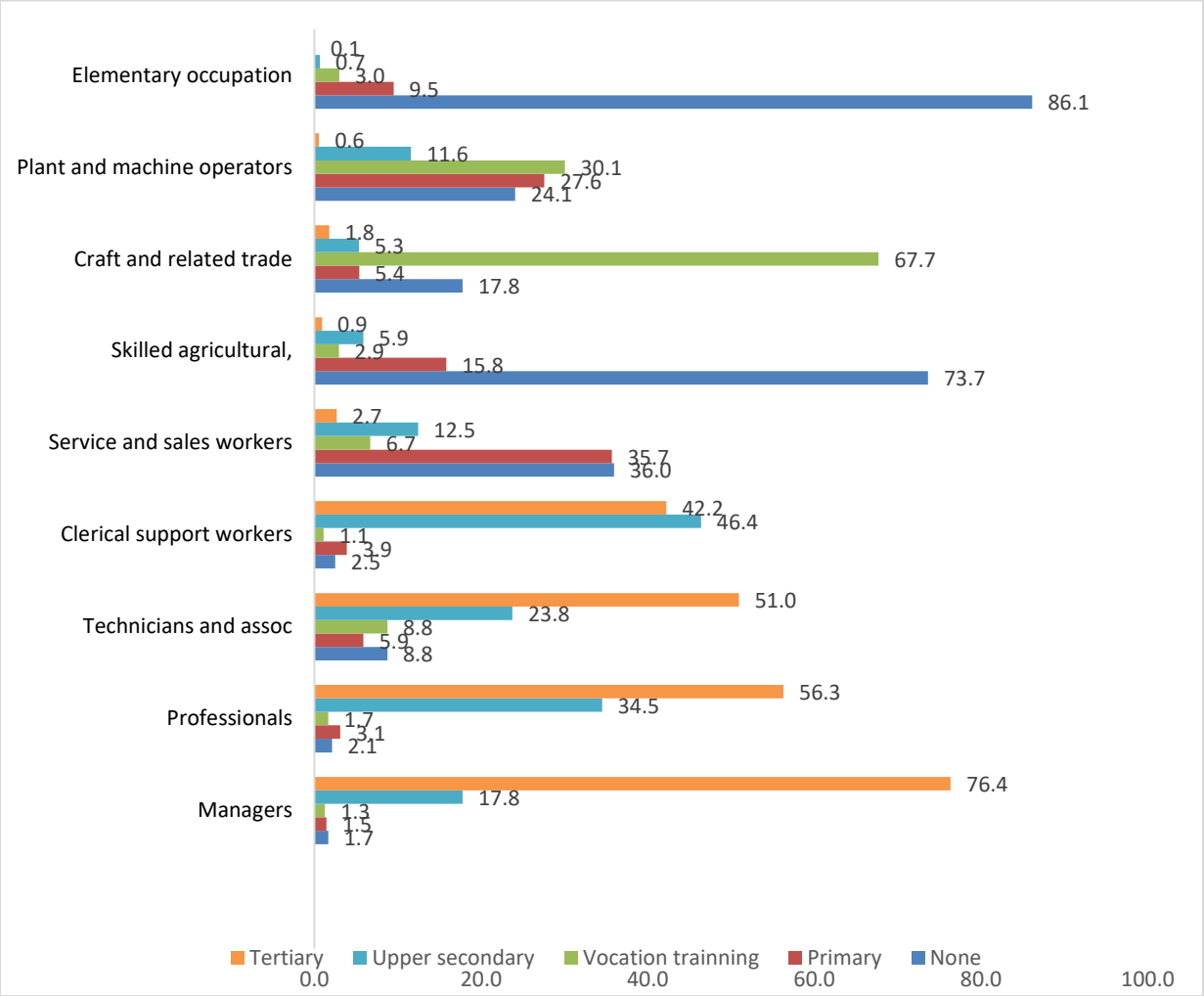


Figure 3: Main occupation and level of education required

Source: Labour Force Survey FS 2021

The Labour Force Survey included the question on self-evaluation about skills where the respondent was to answer the level of education needed to fill the post that he/she is occupying, and the findings indicate that on the post of managers; most of the participants think that the skills should be linked with university education (86.1 percent of the person in employment).

The results also showed that there are inconsistencies between the educational levels they possess and those needed for a certain position or occupation.

Table 2: Level of education attained by the required level of education

Required / Attained	None	Primary education	Lower Secondary	Upper secondary	University	Total
None	1,264,010	287,918	9,630	15,883	757	1,578,197
Primary	600,025	324,116	16,469	27,176	1,647	969,433
Lower Secondary	74,121	83,257	13,577	17,009	1,180	189,144
Upper secondary	52,293	84,124	18,894	121,403	21,768	298,483
University	9,719	13,971	4,573	48,112	162,291	238,665
Total	2,000,168	793,385	63,142	229,583	187,643	3,273,921

Source: Labour Force Survey 2021

Table 2 illustrates the employed population according to the attained level of education and required education level according to the position that they are occupying. The data has shown that 1,264,010 persons with no level of education declared that their occupation position matches their level of study. But an estimated 287,918 who were No level of education declared that their occupation post requires at least a Primary level of education, and another 26,286 (The sum of 9630+15883+757) persons with no level of education attained occupying a post that is requiring at least Lower secondary school.

The diagonal elements of the matrix, Table 1, are therefore generally matching skills and the off-diagonal elements are generally over/under skilled persons. The degree of skills-match based on the off-diagonal elements of Table 3, (287,918+9,630+16,469+15883+…600025 +…. + 48,112)/3,273,921 = 42.4%, is, therefore, an overestimate of the true skills mismatch among the employed population in Rwanda. Correspondingly, the degree of skill match estimated on the basis of the diagonal elements of Table 1, (1,264,010+324,116+ ….162,291)/3,273,921 = 57.6% is an underestimate of the true skills match of the employed population.

4.2.2 Incidence of Mismatch by Major Occupation Group

Studying the skills mismatch by occupation level has shown that persons employed in Service and sales workers are relatively under-skilled (60 percent) as compared with other main occupation groups,

followed by Skilled agriculture (37 percent), craft and related trade (29 percent) and by plant and machine operators (19 Percent).

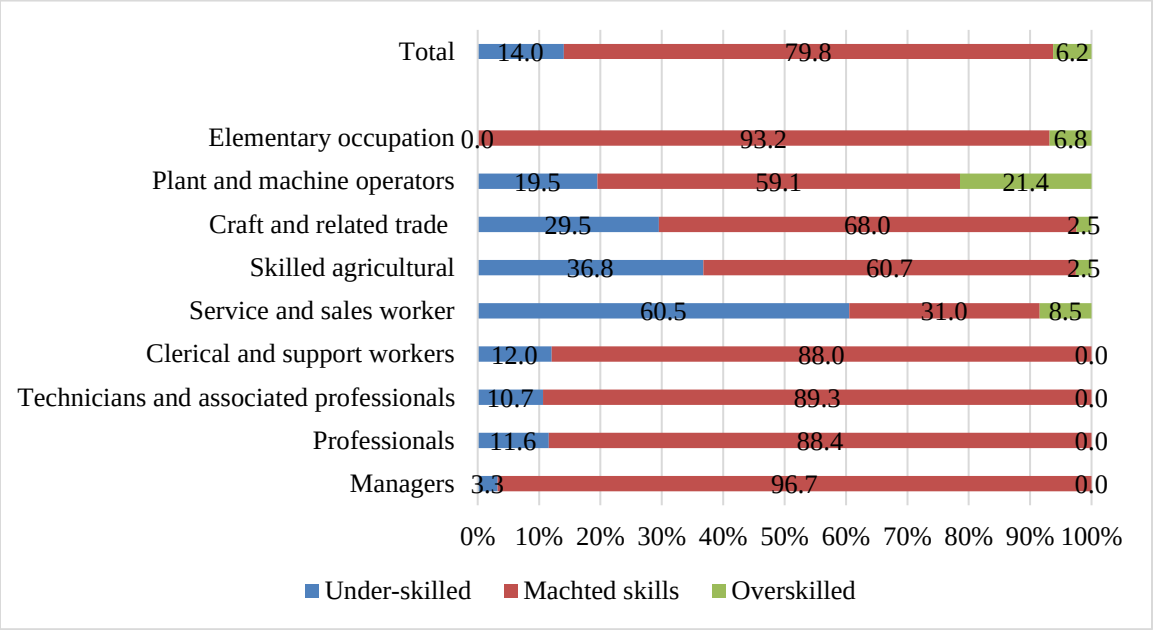


Figure 4: Occupation group by a Skills mismatch
Source: Labour Force Survey 2021

4.2.3 Dissimilarity between employed and unemployed/outside labour force

The analysis of the Labour force survey conducted in Rwanda in the year 2021 calculated the Mismatch rate of the level of education which is described as the disparities in the distribution of educational attainment between various categories showing how much more education the unemployed or those who are not in the labour force need to have in order to have the same distribution of educational attainment as employed people. According to this study's findings, 7.9% of the unemployed population lacks the skills necessary for work given the amount of education they have acquired and the positions they are qualified for (see Table 3)

Table 3: Mismatch rate of unemployment).

Level of education	(1)	(2)	|(1)-(2)|
	% Employed	% unemployed	Abs_Diff
None	48.2	41.6	6.6
Primary	29.6	28.7	0.9
Lower Secondary	5.8	6.9	1.1
Upper secondary	9.1	15.9	6.8
University	7.3	6.9	0.4
Total			**15.8**
Total/2			**7.9**

Source: Labour Force Survey 2021

On the other hand, the analysis has considered the important part of the working-age population considered as Out of the labour force. Population of the labour force is mainly composed of the so-called Potential Labour force (those willing to work but not available and those who are not willing but available work employment).

The working-age population who is not in the labour force, as shown in Table 4, has an inadequate degree of education compared to what is needed for their jobs. The same rate was observed while studying this rate among the working-age population out of the Labour force.

Table 4: Mismatch rate of employed population

Level of education	(1)	(2)	|(1)-(2)|
	% Employed	% OLF	Abs_Diff
None	48.2	49.6	1.4
Primary	29.6	30.6	0.9
Lower Secondary	5.8	11.3	5.6
Upper secondary	9.1	7.4	1.7
University	7.3	1.1	6.2
Total			15.8
Sum/2			7.9

4.2.4 Analysis of skills mismatch

The analysis of skills mismatch using the normative approach generally has shown that 73.3 percent of the employed population in Rwanda have matched skills in their job positions whereas 21.8 were classified as under-skilled meaning that their level of education does not match their occupations and only 4.8 were classified as over-skilled meaning that their education level was found higher than the required education level.

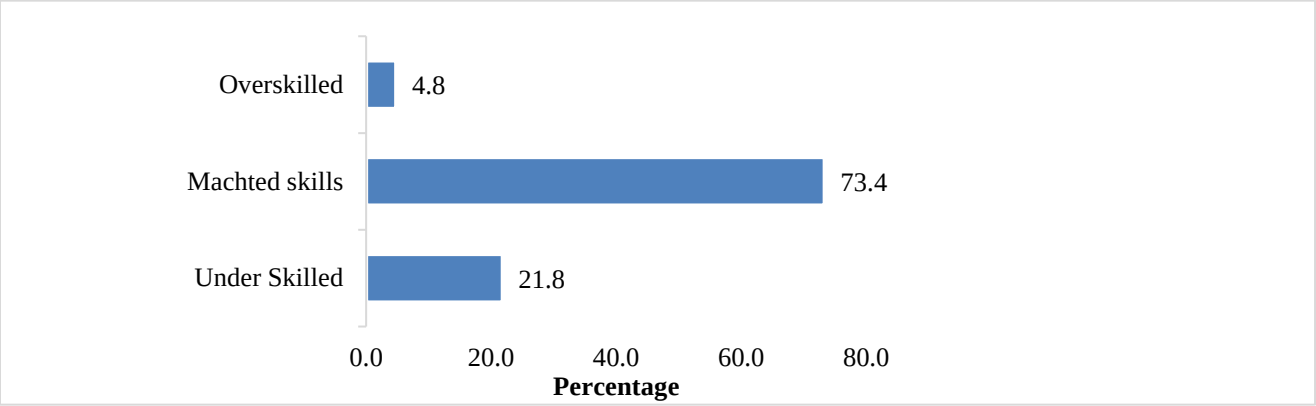


Figure 5: General Skills mismatch

Source: Labour Force Survey 2021

Several pieces of information can be derived from the analysis of skills mismatch. The current section analyzed skills mismatch concerning Demographic characteristics. It can be noted that skills were matching among Female employees (75 percent) as compared to Male (72 percent) employees. The rate of under-skilled employment was lower among females and slightly higher among Male employees.

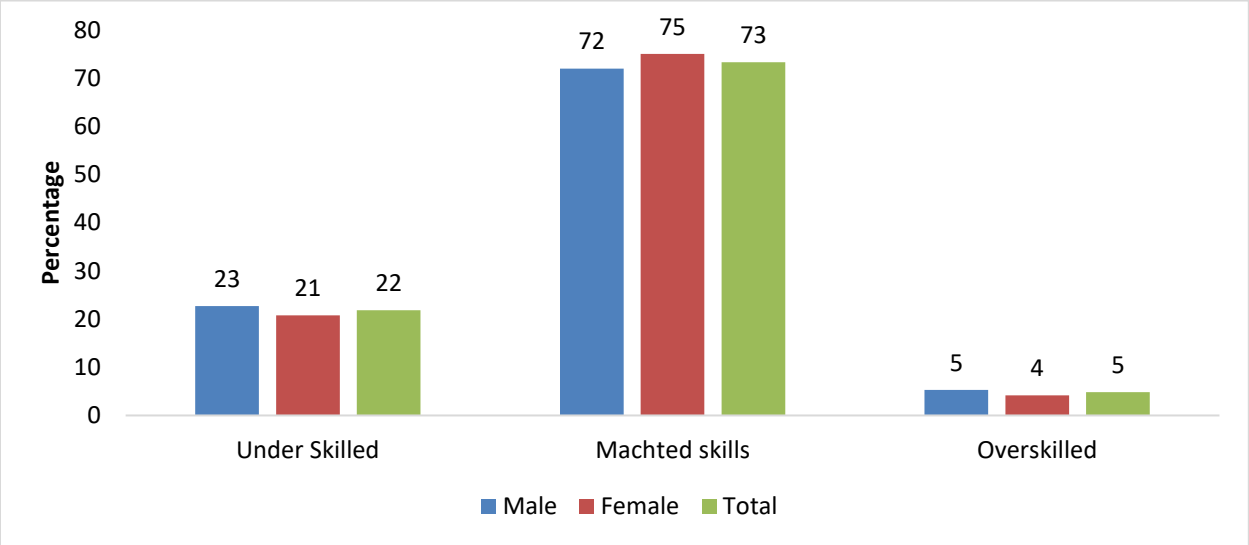


Figure 6: Skill mismatch with respect to sex of employed person

Source: Labour Force Survey 2021

According to the regions, the results showed that except for Kigali City, the rest of the provinces have rates close to the national average.

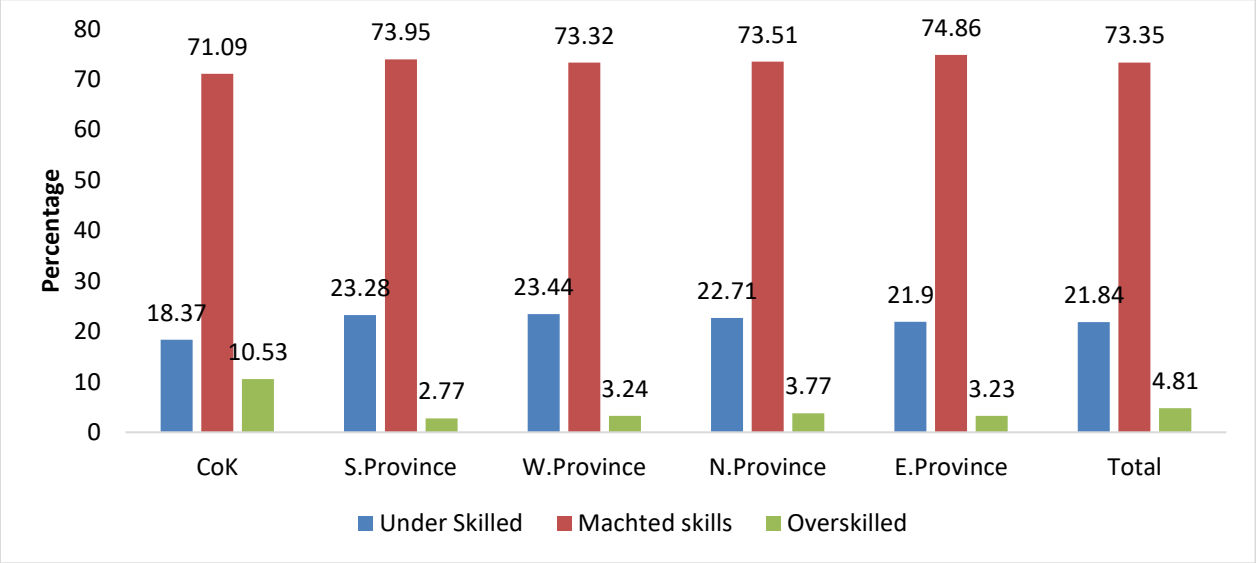


Figure 7: Skills according to Province

Source: Labour Force Survey 2021

4.3 Analysis of Skills miss match using data mining techniques

4.3.1 Job advertised

To achieve the main objective of the study, this research through web scraping has estimated the number of advertised jobs according to the level of the occupation.

Table 5: Advertised job through job in Rwanda and MIFOTRA Portal 2021 2022

Level of Skills		2021	2022	Total
3 and 4	Managers	48	32	80
4	Professionals	978	978	1956
2 and 3	Technicians and associate professionals	2,121	2,121	4,242
	Clerical support workers	2,954	2,954	5,908
	Service and sales workers	132	236	368
	Skilled agricultural, forestry and fishery workers	8	11	19
	Craft and related trades workers	0	0	0
2	Plant and machine operators and assemblers	2,445	2,645	5,090
1	Elementary occupations	0	0	0
	Total	8,686	8,977	17,663

The findings indicated that the labour market in Rwanda needs more technicians and Cleric support workers. Among the technicians and associate professionals, we can mention the teachers in Primary, secondary and university and among the Cleric and support workers we can mention the administrative assistants, cashier and other persons in charge of clerical services. Workers needed among Plant and machine operators are most dominated by drivers and machine operators, especially for road maintenance and construction machines.

4.3.2 Main significant factor from multinomial logistic regression

Table 6: Main Significant factors (multinomial approach)

Skills=2	**coef**	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-0.0152	0.012	-1.272	0.203	-0.039	0.008
province	0.2332	0.124	1.878	0.06	0.01	0.477
code_dis	-0.0152	0.012	-1.272	-1.272	-0.039	0.008
Code_UR	0.665	0.054	12.428	0	0.56	0.77
Sex	0.053	0.042	1.878	0.06	0.01	0.477
indd03	-0.1592	0.006	-28.727	0	-0.17	-0.148
main_sect	2.0964	0.041	50.553	0	2.015	2.178
age5	0.1737	0.008	20.754	0	0.157	0.19
TVT2	-0.5833	0.034	-17.394	0	-0.649	-0.518
Skills=2	**coef**	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-1.9546	0.294	-6.654	0	-2.53	-1.379
province	0.2486	0.241	1.033	0.301	-0.223	0.72
code_dis	-0.0349	0.023	-1.512	0.131	-0.08	0.01
Code_UR	-0.7511	0.091	-8.258	0	-0.929	-0.573
Sex	-0.084	0.065	-1.294	0.196	-0.211	0.043
indd03	-0.0293	0.007	-4.075	0	-0.043	-0.015
main_sect	0.3291	0.07	4.688	0	0.192	0.467
age5	-0.154	0.016	-9.426	0	-0.186	-0.122
TVT2	0.7706	0.05	15.342	0	0.672	0.869

Several variables were identified to represent the model when considering matched skills as the reference category with a p-value less than 0.05. Among those factors we can mention; The area of residence (Code_UR), the Economic activity(Indd03), the Main Economic Sector (main_sect), the age group(age5) and the completeness of the General of the TVET Program (TVT2) while explaining between skills miss-match in terms of under-skilled (Skills=1).

Furthermore, the same factor/ variables were found as the best feature in explaining skills mismatch in terms of over-skilled (Skills=2)

Table 7: Confusion Matric for general prediction

	predicted_Under_skilled	***predicted_match_skilled***	***predicted_Over_skilled***
match skilled	3244	158	0
under skilled	669	335	0
Over skilled	298	6	

According to the confusion matrix above, the True positive number that can be predicted by the decision tree model are about 3244 persons in skills matched, while 157 were predicted as False Positives (158+0) as they were predicted to be matched skilled yet they are either Under-skilled or Over-skilled, etc.

4.3.3 Predictive models

There are several classification models that can be used for multiple classes (multi-class classification) in Python. We used Naive Bayes: Naive Bayes classifiers, such as Gaussian Naive Bayes or Multinomial Naive Bayes to handle multi-class classification by estimating the probability of each class given the input features; Decision Trees; Random Forest and K-Nearest Neighbours (KNN)

4.3.3.1: Decision tree

The study has found that the consists of a structured tree classifier where the internal nodes show the dataset's features and have an accuracy of 76 percent, and the results are presented in the table below.

Table 8:Precision and recall for Decision Tree

	precision	recall	f1 score	Support
match skilled	0.82	0.91	0.86	3,462
Under-skilled	0.64	0.51	0.57	967
Over skilled	0.3	0.1	0.15	281
accuracy			0.78	4,710
macro avg	0.59	0.5	0.52	4,710
weighted avg	0.75	0.78	0.76	4,710

The actual value and expected value of the confusion matrix obtained from the decision tree model sho uld be the same. The True Positive value for the "Skills match" class is 3156, whereas the Sum of value s for the relevant rows with the exception of the TP value are False Positive. The True Negative values are the sum of all columns and rows except for the values of that class that we are computing the values for True Negative Value (528+15+15+30), and the True Positive values are shown as the sum of values of the relevant column except for the True Positive value (441+235).

Similar calculations are made for the Under-skilled class, considering the interaction class. 528 employ ed people provided the True Positive value, 441+15 provided the False Negative, 218+18 provided the False Positive, and 3156+70+235+30 provided the True Negative.

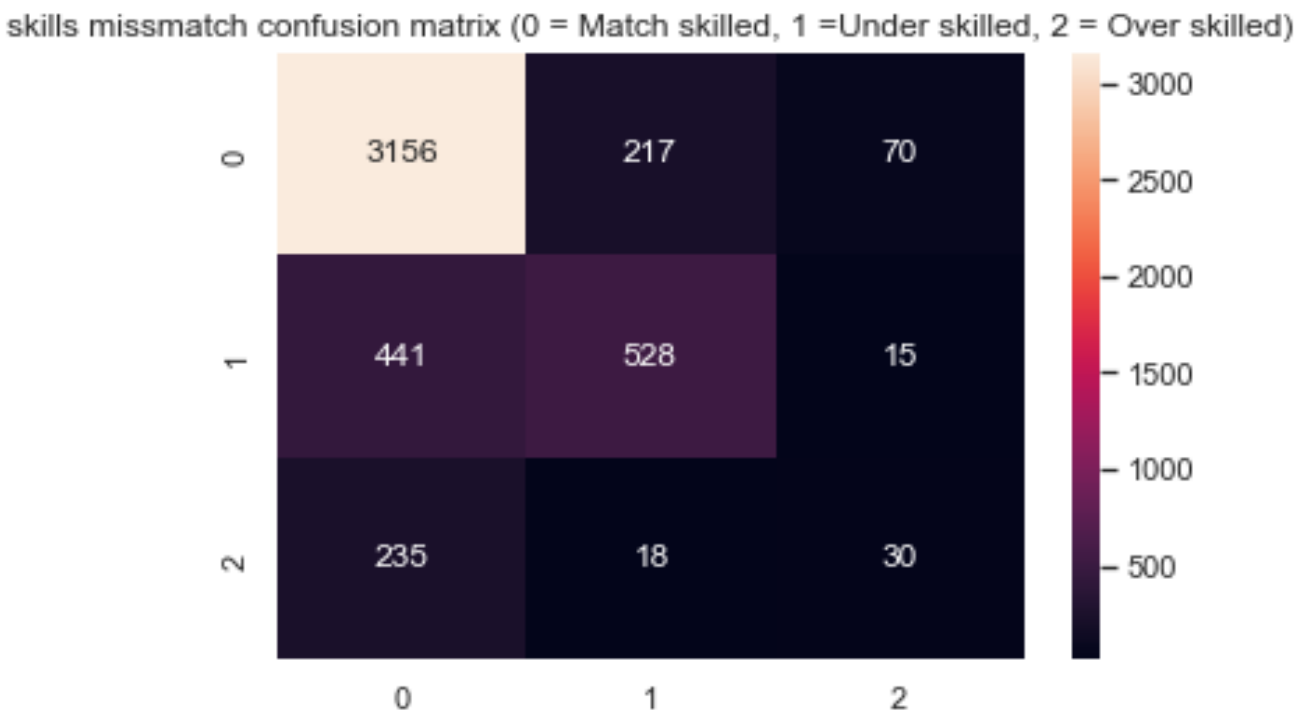


Figure 8: Confusion matrix for decision Tree

The overall accuracy model was found to be 78 percent; this means that 78 per cent of the data can be explained using this model.

The computation of the area under the curve (AUC) for a Decision Tree classifier revealed the following results:

Class 0, representing "Skills-matched," had an AUC of 0.75.

Class 1, representing "under-skilled," achieved an AUC of 0.8.

Class 3, representing "Over-skilled," demonstrated an AUC of 0.78.

As it can be represented in the figure 8 in this study.

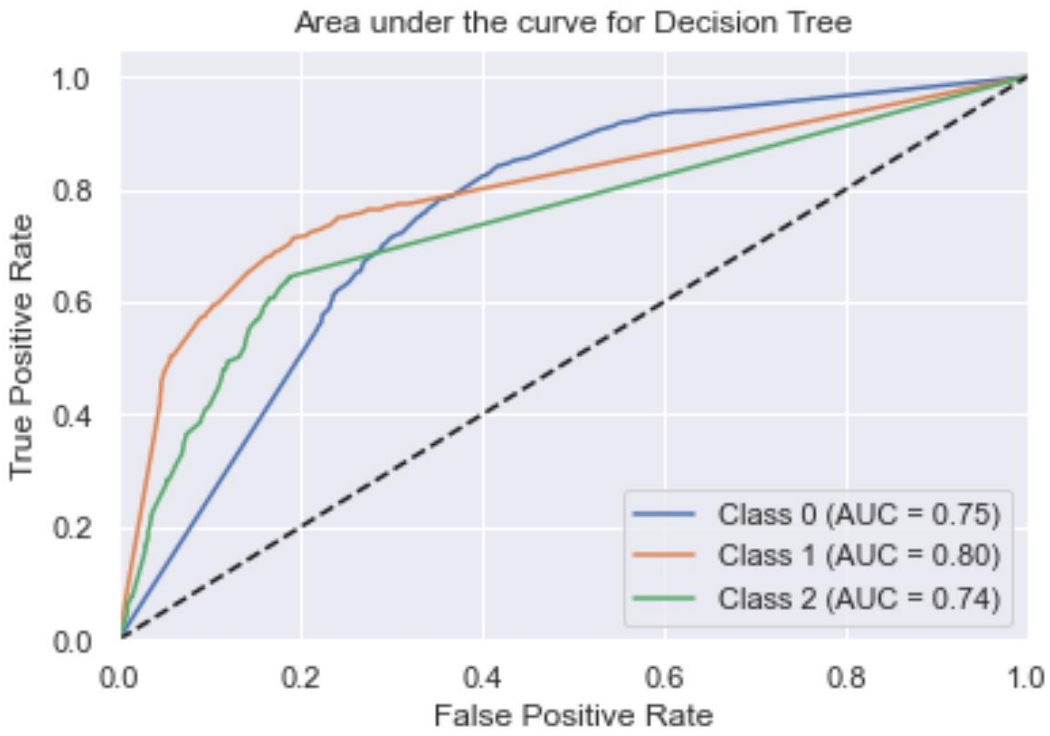


Figure 9: Area Under the Curve and the Receiver Operating Characteristics (AUC-ROC) for Decision tree

4.3.3.2: Random Forest

Among different variables that were in the model, the research intended to highlight the important feature, whereby the most important feature to predict skills miss-match was the economic activity (indd03), followed by age (grouped in five years), followed by the district where the employed person is located, again by main economic activity, province, residential area and by sex. The main four important features (probability more than 5 percent) identified to predict skills miss-match were the main economic activity, the age, the district location and the Education status (Either completed general education or TVT)

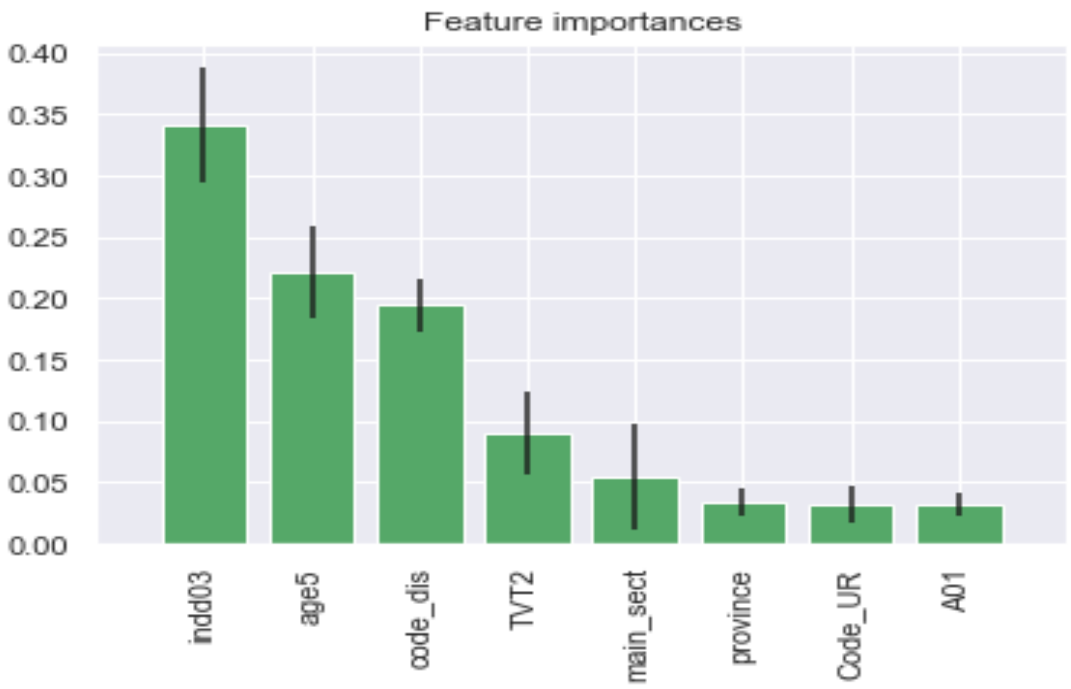


Figure 10: Important Futures from random forest

The overall precision of this model was found to be 79 percent meaning that 79 percent of the observation can be explained using this model. The Recall which can be explained as how well the model identifies true positives specifies that the model can predict 54 percent of under-skilled and 11 percent of Over-skilled. Out of all the respondents who were classified as a skilled mismatch, only 11 percent can be truly classified as Over skilled. In other words; the prediction of being over-skilled is low and this may be due to few numbers of observations in this class.

The f1-score which is also the level of combined precision of predictions illustrated that the model can predict positive (True Positive) 87 percent for skills match, 59 percent for under-skilled and 17 percent for Over skilled.

Table 9: Level of precision prediction for Random Forest

	precision	recall	f1 score	Support
Match skilled	0.83	0.91	0.87	3,462
Under-skilled	0.65	0.54	0.59	967
Over skilled	0.36	0.11	0.17	281
accuracy			0.79	4,710
macro avg	0.61	0.52	0.54	4,710
weighted avg	0.76	0.79	0.77	4,710

Actual and expected values for the confusion matrix produced by the Random Forest model should coi ncide. The matrix showed that the value 3155 is the True Positive value in the "Skills match" class, whi le all other values in the corresponding rows (apart from the TP value) are False Positive. The True Pos itive values are displayed as the total of all values in the relevant column, with the exception of the Tru e Positive value (259+48), and the True Negative values are displayed as the sum of all values in all col umns and rows, except for the values for the class for which we are computing the True Negative Value (520+10+18+32).

Similar calculations are made for the Under-skilled class, considering the interaction class. Where 520 employed individuals provided the True Positive value, (437+10), (259+18) and (3155+48+231-23) for the False Negative and True Negative, respectively.

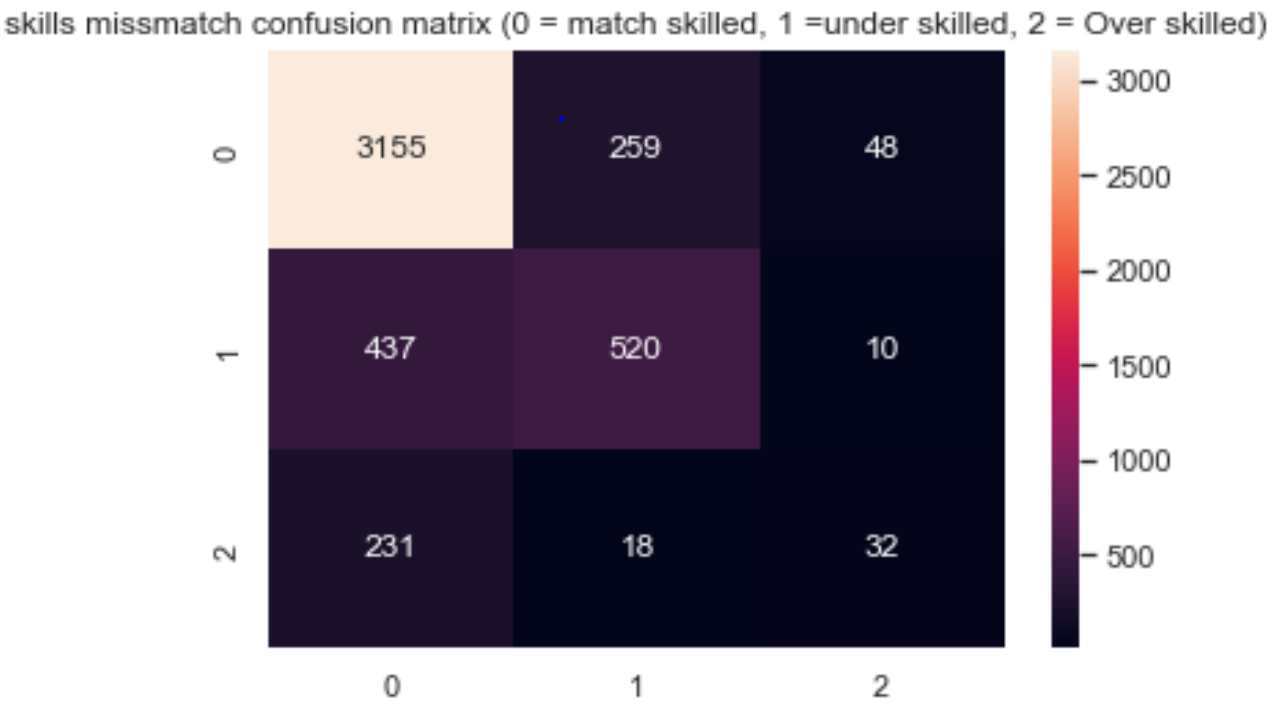


Figure 11: Confusion Matric for the Random Forest

The computation of the area under the curve (AUC) for a Decision Tree classifier revealed the following results:

Class 0, representing "Skills-matched," had an AUC of 0.75.

Class 1, representing "under-skilled," achieved an AUC of 0.8.

Class 3, representing "Over-skilled," demonstrated an AUC of 0.78.

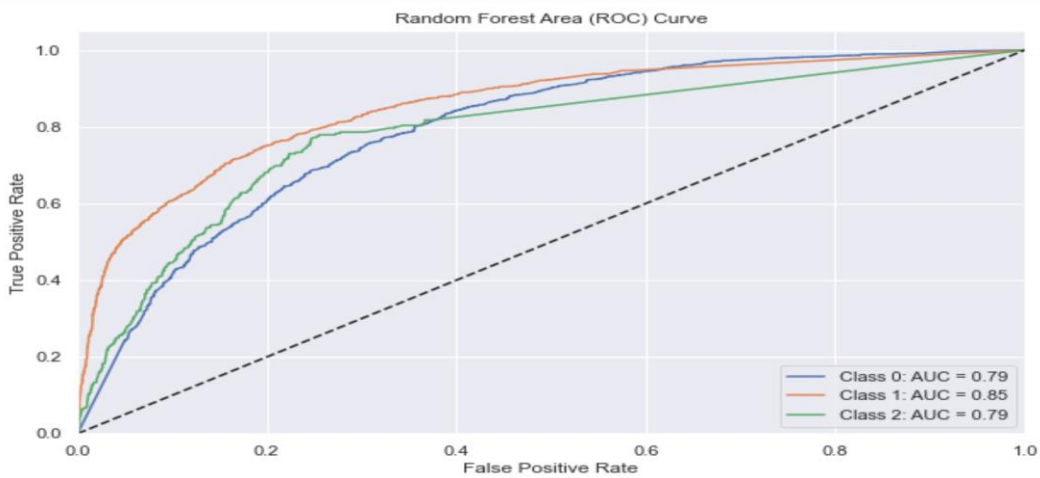


Figure 12: Area Under the Curve and the Receiver Operating Characteristics (AUC-ROC) for Random Forest

4.3.3.3: K Nearest Neighbour

The KKN model was also used to test the best-fit model, and the total precision was determined to be 77%, indicating that this model may be used to estimate 77% of the observation. Recall and F1-score are probably the same as those seen with the prior model.

Table 10: Precisions and recall for KNN

	precision	recall	f1 score	Support
Match skilled	0.81	0.9	0.87	3,462
Under-skilled	0.63	0.48	0.54	967
Over skilled	0.2	0.08	0.12	281
accuracy			0.77	4,710
macro avg	0.55	0.49	0.5	4,710
weighted avg	0.74	0.77	0.75	4,710

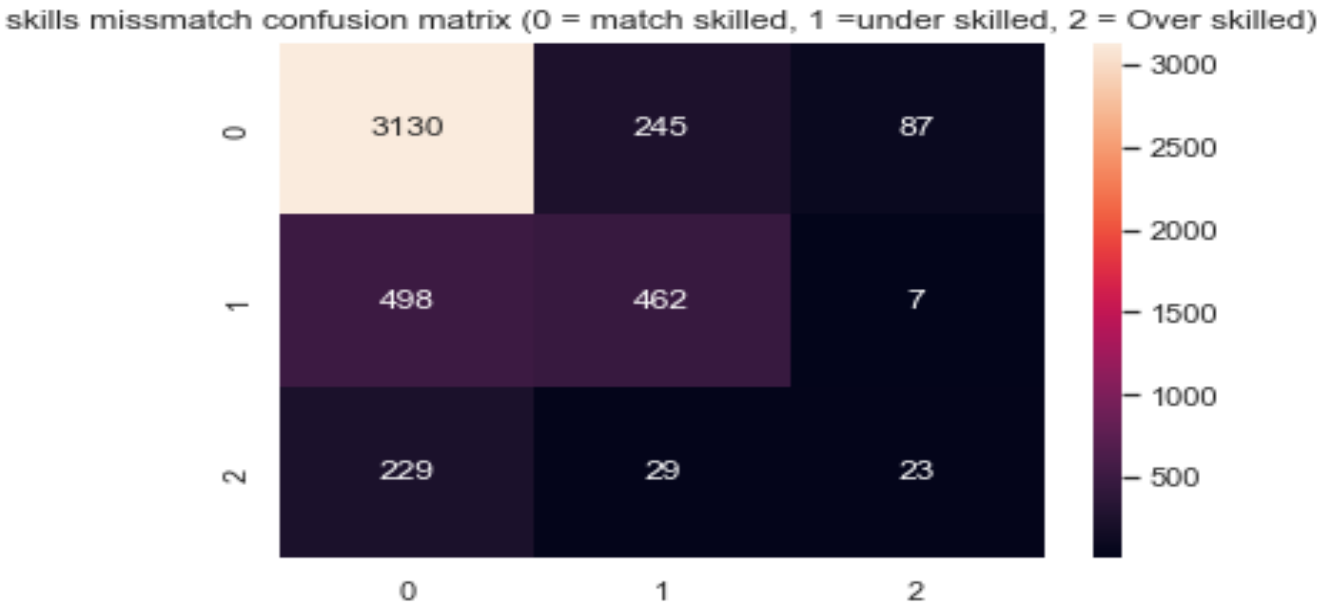


Figure 13: Confusion Matrix for he KKN

The confusion matrix derived from the KNN model confirmed the previous finding, showing that for th e "Skills match" class, the value 3130 is the True Positive value, while the Sum of values of correspond ing rows other than the TP value is False Positive were (245+87), and the True Negative values are the sum of values of all columns and rows other than the values of that class that we are calculating the val ues for True Negative Value (462+7+29+23).

In contrast, the values/metrics for the Under-skilled class are derived while taking the interaction class into account. Employed individuals provide 462 of the True Positive value, 498 of the False Negative value, 218 of the False Positive value, and 3130 of the True Negative value.

4.3.3.4: Naïve Bayes Model

Table 11: Precision and recall for Naïve Bayes Model

	precision	recall	f1 score	Support
Match skilled	0.76	0.79	0.77	3,462
Under-skilled	0.70	0.11	0.19	967
Over skilled	0.15	0.49	0.22	281
accuracy			0.63	4,710
macro avg	0.54	0.46	0.40	4,710
weighted avg	0.71	0.63	0.62	4,710

The confusion matrix for the actual value and predicted value derived from the Naive Bayes model sho wed that in the "Skills match" class, the value 3313 is the True Positive value, while the FP is shown as the sum of values of the corresponding column except for the True Positive value (149+0), and the True Negative values are the sum of values of all columns and rows except for the values of that class that w e are calculating the values for True Negative Value (470+0+0+4).

The metrics are calculated similarly for the Under-skilled class while taking the interaction class into ac count. The False-Negative value is provided by 497+0, the False Positive by 218+18, and the True Neg ative value is provided by (3313+0+277+0) while 470 employed persons person were classified as true positive.

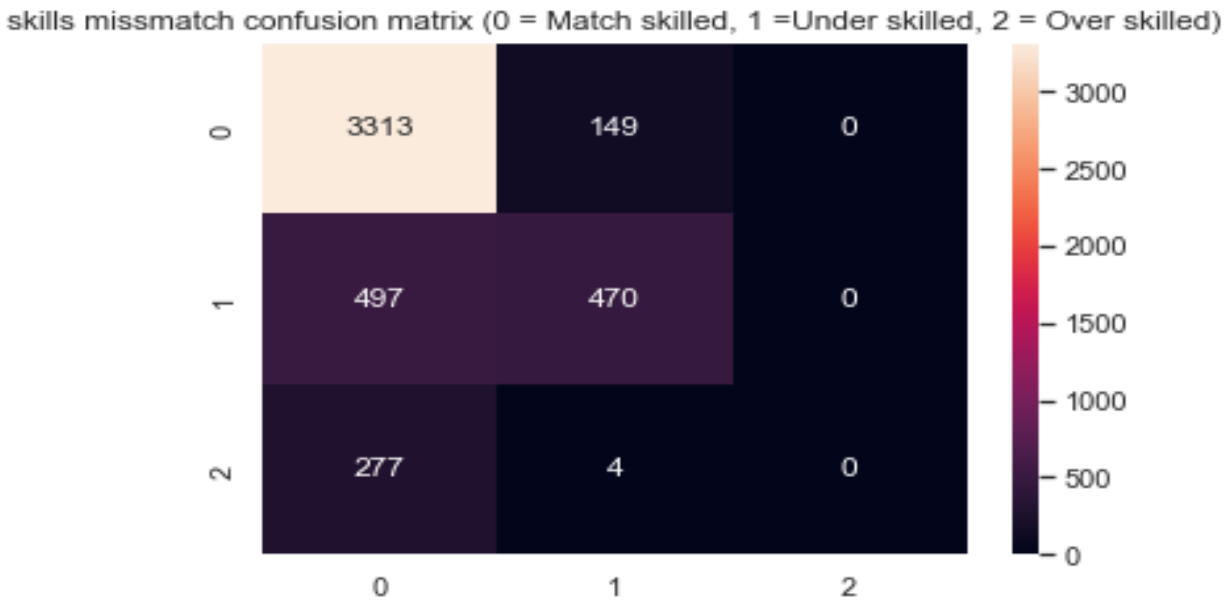


Figure 14: Confusion Matrix for Naïve Bayes

In conclusion, based on the accuracy and the degree of precision, we may conclude that the best model to estimate skilled mismatch is both Decision Tree and Random Forest as presented in "Table 4" below

Table 12: Summary of the predicted model

Mode	**Accuracy**	**Precision**	**Recall**	**F1_score**
Decision tree	0.79	0.76	0.79	0.77
Random forest	0.79	0.76	0.79	0.77
KNN	0.77	0.74	0.77	0.75
Naïve Bayes	0.63	0.71	0.63	0.62

According to accuracy, the following models do the best: Decision tree (0.79), Random Forest (0.79), KNN (0.77), and Naive Bayes (0.63). Decision tree, Random Forest, KNN, and Naive Bayes all received precision scores of 0.76, 0.76, 0.76, and 0.71 respectively for the classifiers that were assessed. The recall score for each model was 0.79 for the decision tree, 0.79 for the random forest,

0.75 for KNN, and 0.63 for Naive Bayes. The F1 score for each model was 0.77 for the decision tree, 0.79 for the random forest, 0.89 for KNN, and 0.62 for Naive Bayes.

4.4 Discussion

Skills mismatch refers to a situation where the skills possessed by the workforce do not match the skills demanded by employers or the labour market. It has significant economic and social implications. Job vacancies are open positions within organizations that employers actively seek to fill (Dolton P. and Vignoles, A. 179–198). They are advertised through various channels, providing details about the job title, responsibilities, required qualifications, and application instructions. Skills mismatch arises from a gap between the skills individuals acquire through education and training and the skills needed in the job market. Factors contributing to this gap include outdated education systems, lack of alignment between educational programs and industry needs, and technological advancements rendering certain skills obsolete (Freeman 1976).

Skills mismatch has negative consequences for individuals and the economy. It can result in unemployment, underemployment, and limited career growth, leading to reduced job satisfaction and lower wages. From an economic standpoint, it hinders productivity and innovation, with businesses struggling to find qualified workers or facing unemployment due to an oversupply of certain skills. Skills mismatch also contributes to income inequality and social exclusion, creating a divide between those with in-demand skills and those without, limiting economic participation and social mobility (Frenette 2004).

Addressing skills mismatch in Rwanda requires a multi-faceted approach involving collaboration between educational institutions, policymakers, employers, and individuals themselves (S. McGuinness 2003). Some strategies to mitigate skills mismatch include:

Enhancing educational and training programs: educational institutions should ensure their curricula align with industry needs, providing students with relevant and up-to-date skills.

Strengthening career guidance and counselling: Providing individuals with accurate and timely information about labour market trends and skill demands can help them make informed career decisions and choose educational paths that align with market needs.

Promoting partnerships between academia and industry: Internships, apprenticeships, and work-integrated learning programs can also bridge the gap between theory and practice.

Supporting retraining and upskilling programs: Government initiatives and employer-supported programs can help individuals acquire new skills or upgrade existing ones to adapt to changing job requirements.

In conclusion, skills mismatch is a complicated issue that calls for coordinated action from numerous parties. Rwanda can reduce skills mismatch and build a more skilled and competitive workforce by coordinating education and training with labour market needs, encouraging continuous learning, and assisting people in gaining relevant skills (RDB 2016). This, in turn, can contribute to economic growth, social inclusion, and individual well-being. It is crucial to keep in mind that due to the dynamic nature of labour markets, the effect of many external factors, and the accessibility of high-quality data, precisely forecasting job openings can be difficult.

The quality and applicability of the data collected, as well as the modelling approaches applied, have a significant impact on the forecasts' accuracy and efficacy. The study was therefore to analyze the advertised job and identify potential employment needed in the market. Further research can work on the prediction of skills that would be needed in the future for good preparation of the country’s performance. Different authors have described the various approach to stud job mismatch as the Normative approach, statistical approach or self-assessment approach where skills can be studied either by analyzing mismatch by skills or by qualification. We also recommend other researchers study the effect of skills mismatches and economic performance or use any other approach in addressing skills mismatch in Rwanda.

Chapter 5: Conclusion and Recommendations

5.1 Conclusion

The analysis in this study brought attention to the skills mismatch in Rwanda and emphasized the importance of the demand side of the labour market in determining training and skill gaps. Most importantly, it was found that skill development involves both work tasks and work structure on both the demand side (Institutions) and the supply side, not simply one or the other. Both the supply and demand sides of skills have been considered when developing policies. Otherwise, even highly qualified individuals run the risk of not using their abilities, which depreciates the value of educational investments.

The skills mismatch resulted in an overestimation of 42 percent of Rwanda's employed population, leaving 57.6 percent of the workforce unemployed. When compared to other major occupational groups, service and sales workers are relatively under-skilled (60 percent), followed by skilled agricultural workers (37 percent), artisans and workers in related trades (29 percent), and plant and machine operators (19 percent).

According to the normative approach, 73.3 percent of Rwanda's workforce has the skills required for their jobs, while 21.8 percent were classified as under-skilled, meaning their education levels did not correspond to their occupations, and only 4.8 percent were classified as over-skilled, meaning their education levels exceeded the minimum requirements. Skills are more well-matched among women (75%) than men (72%). Under-skilled were less prevalent among women and slightly more prevalent among men. With the exception of Kigali, we can claim that the rates in the other provinces are close to the national average.

In the years 2021–2022, there have been about 18 000 job advertisements. The majority of the jobs that were advertised required at least a Bachelor's degree for professionals and technicians, as well as technical and vocational training for plant and machine operators. According to this study's findings, 7.9% of the unemployed population lacks the required job-related skills given the degree level that has been acquired.

The best models to estimate skilled mismatch were both Decision Tree and Random Forest with an overall precision of 79 percent, according to the accuracy and degree of precision of the machine learning technique that was used to make this choice.

5.2 Recommendations

Finding potential means of assisting in the adoption of the use of current skills in economic performance has received special attention. Otherwise, even highly qualified individuals run the danger of not using their abilities, which depreciates the value of educational investments.

Education should be more responsive to labour market demand as a result of the policies that target the labour supply, and Create structures for adult education, training, and work-based learning. The adoption of technologies and procedures that maximize complementarities to the existing skill sets should be encouraged by the policies that target the labour demand.

Policies that coordinate the labour supply and demand should provide information and guidance by facilitating and promoting coordinated stakeholders in the ensemble. The role of the government (REB) and policy should be in elaborating competency-based curriculum but also considering conducting research to find out what is the market, The role of workers and organized labour should be well enhanced o find out how demand and supply for skills can be matched and The role of employers in regularly training workers on needed skills as a dynamic event.

REFERENCES

Adalet McGowan M. and Andrews. 2015. "Labour Market Mismatch and Labour Productivity." *Evidence from PIAAC Data", OECD Economics Department Working Papers, No. 1209, OECD.*

AfDB. 2012. *Labour force data analysis: Guidelines with African specificities. (African Development Bank, Abidjan).* AfDB.

Allen J., van der Velden R. 2021. "Educational mismatches versus skill mismatches:Effect on wages, job satisfaction, and on-the-job search." *Oxford Economic Papers 53(3)* (effects on wages, job satisfaction, and on-the-job search) 434–52.

Amador, Lourdes Badillo. 2006. "Education and competence mismatches: job satisfaction consequences for workers ." *Journal Economic Literature* 2-5.

Arsenau, D. and Epstein, B. 2014. *The Welfare Costs of Skill Mismatch Unemployment.*

Autor, D., Levy, F., & Murnane, R.,. 2003,. "The skill content of recent technological change: An empirical exploration, ." *Quarterly Journal of Economics, vol. 118, no. 4* pp. 1279-1333.

Becker, G. 1964. *Human capital.* New York: National Bureau of Economic Research. .

Belle Fille M., J. Felicien I , Idrissa K. 2023. "Comparison of Tree-Based Machine Learning Algorithms to Predict Reporting Behavior of Electronic Billing Machines." *Information 2023* 9-13.

Brownlee, Jason. 2019. *Machine Learning Mastery.* October. Accessed April 2023. https://machinelearningmastery.com/naive-bayes-classifier-scratch-python/.

Brun-Schammé, A. and Rey, M. 2021. "A new approach to skills mismatch." (OECD Productivity Working Papers).

Brun-Schammé, A. and Rey, M. 2021. "A new approach to skills mismatch." *OECD Productivity Working Papers.*

Brun-Schammé, A. and Rey, M. 2021. "A new approach to skills mismatch."

Cappelli, P. 2014. *Skill Gaps, Skill Shortages and Skill Mismatches: Evidence from the US.* Working Paper 20382.

CEDEFOP. 2010. *Skills Supply and Demand in Europe: Medium-term Forecast up to 2020.*

Daniel, B.K. 2017. "Contestable professional academic identity of those who teach research methodology." *Int. J.*

Dolton P. and Vignoles, A. . 179–198. "The Incidence and Effects of Overeducation in the U.K. Graduate Labour Market." *Economics of Education Review* 2000.

European Commission. 2018. *Labour Market and Wage Developments in Europe 2018.* Bruxels: European Commission.

Freeman, R. B. . 1976. "The Overeducated American." *Academic Press.*

Freeman, R. B. 1976. *The Overeducated American.* New York: Academic Press.

Frenette, M. 2004. "The Overqualified Canadian Graduate: The Role of the Academic Programme in the Incidence, Persistence and Economic Returns to Overeducation." *Economics of Education Review,23* 23, 28-16.

Green, A. and Owen, D., 1992,. n.d. "Skill Shortages: Causes and Consequences,." In *Skill Shortages: The Local Dimension*. Avebury: Aldershot.

Hossin, M.; Sulaiman, M.N. . 2005. "A Review on Evaluation Metrics for Data Classification Evaluations. Int. J. Data Min. Knowl. Managament process." *Int. J. Data Min. Knowl. Manag. process* 1-11.

ILO. 2013. *19th ICLS Guideline.* ILO.

ILO. 2008. *International Standards of Classification of Occupation.* ILO.

—. 1982. *Resolution concerning statistics of the economically active population, employment, unemployment and* . ILO. October.

Lehti, S., and E. Lehtinen. 2005. "Computer-supported Problem-based Learning in the Research Methodology Domain." *Scand. J. Educ. Res. 2005, 49, 297–324.* 49,296-323.

McGuinness, S. 2003. "Graduate Overeducation as a Sheepskin Effect: Evidencefrom Northern Ireland". ." *Applied Economics* 597-608.

McGuinness, S., and Sloane P. 2011. " Labour market mismatch among UK graduates: an analysis using REFLEX data. ." *Economics of Education Review, 30* 130-145.

RDB. 2016. *Tracer Survey for graduates.* Kigali: GoR.

Rokach, L., and O. Maimon. 2005. "Decision Trees. In Data Mining and Knowledge Discovery Handbook." *Springer* (Boston, MA, USA, 2005; pp. 165–192.) 165–192.

Rwanda, Governmentof. 2017. *National Strategic for Tranformation 1.* Kigali: Gor.

Sala, Gullem. 2011. *In his book on approaches to skills mismatch in the labor market.*

Stephan Kampelmann , François Rycx. 2012. *The Impact of Educational Mismatch on Firm Productivity: Evidence from Linked Panel Data.* Brussels: IZA DP No 7093.

Thurow, L.C. 1975. *Generating Inequality.* New York: : Basic Books.

Tsang, Mun C. 1987. "The impact of underutilization of education on productivity: A case study of the U.S. Bell companies." *Economics of Education Review* 239–254.

Vujovic, Ž.D. Classification. 2021. "Model Evaluation Metrics." *Appl.* , Appl. 12.

ILO, *Resolution concerning the measurement of working time*, 18th International Conference of Labour Statisticians, Geneva, 24 November – 5 December 2008. http://www.ilo.org/global/statistics-and-databases/standards-and-guidelines/resolutions-adopted-by-international-conferences-of-labour-statisticians/WCMS_112455/lang--en/index.htm

ILO, *International Standard Classification of Occupations, ISCO-08,* http://www.ilo.org/public/english/bureau/stat/download/res/futisco.pdf

United Nations, *International Standard Industrial Classification of All Economic Activities, ISIC* Rev. 4 http://unstats.un.org/unsd/cr/registry/isic-4.asp

Appendices: A Turn tin report

Thesis

ORIGINALITY REPORT

4%	4%	0%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dr.ur.ac.rw Internet Source	1%
2	statistics.gov.rw Internet Source	<1%
3	libstore.ugent.be Internet Source	<1%
4	www.ilo.org Internet Source	<1%
5	repository.smuc.edu.et Internet Source	<1%
6	observatoriocensal.files.wordpress.com Internet Source	<1%
7	scholarworks.rit.edu Internet Source	<1%
8	researchonline.ljmu.ac.uk Internet Source	<1%
9	etd.aau.edu.et Internet Source	<1%

10	palevel.unza.zm Internet Source	<1 %
11	scholarworks.utep.edu Internet Source	<1 %
12	repository.unej.ac.id Internet Source	<1 %
13	www.tandfonline.com Internet Source	<1 %
14	microdata.statistics.gov.rw Internet Source	<1 %
15	unstats.un.org Internet Source	<1 %
16	www.mdpi.com Internet Source	<1 %
17	doaj.org Internet Source	<1 %
18	doi.org Internet Source	<1 %
19	editorialexpress.com Internet Source	<1 %
20	lup.lub.lu.se Internet Source	<1 %
21	repository.kemu.ac.ke:8080 Internet Source	<1 %

22	www.edocs.maseno.ac.ke Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		