የኢፌዲሪ የቴክኒክና ሙያ ስልጠና ኢንስቲትዩት
FDRE TECHNICAL & VOCATIONAL TRAINING INSTITUTE

TECHNICAL AND VOCATIONAL TRAINING INSTITUTE (TVTI)

SCHOOL OF GRADUATE STUDIES

FACULTY OF ELECTRICAL AND ELECTRONICS TECHNOLOGY AND INFORMATION AND COMMUNICATION TECHNOLOGY (DEPARTMENT OF INFORMATION TECHNOLOGY)

POSTGRADUATE PROGRAM

Predictive Analytics of Industry Skills Gaps: Using Machine Learning – In Case of Tilahun Yigzaw TVET College

A Thesis Submitted to the Department of ICT in Partial Fulfillment of the Requirements for the Masters of Science in Information Communication Technology

M.SC. THESIS

BY Azmach Berhe

ADVISOR: Dr.Yihenew Wondie

May 2026

ADDIS ABABA ETHIOPIA

Declaration

I, Azmach Berhe Kelali, the under signed hereby declare that this thesis entitled**:**

"Predictive Analytics of Industry Skills Gaps: Using Machine Learning – In Case of Tilahun Yigzaw TVET College"

Is my original work. I have carried out the research independently under the guidance and support of my research advisor. This study has not been submitted for any degree or diploma in this or any other institutions, and all sources of materials used in the thesis have been properly acknowledged.

Declared by

Name Azmach Berhe Kelali

Signature _____

Date _____

This thesis has been submitted for examination with my approval as university thesis advisor.

Advisor: Dr. Yihenew Wondie (PHD)

Signature _____

Date _____

Certification

This is to certify that the thesis prepared by Azmach Berhe, entitled "**Predictive Analytics of Industry Skills Gaps: Using Machine Learning – In Case of Tilahun Yigzaw TVET College"** and submitted in partial fulfillment of the requirements for the Master of Degree in the faculty of ICT complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Name of Advisor: Signature:_____Date: _____

Name of Internal Examiner: _____Signature: _____Date: _____

Name of External Examiner: _____Signature: _____Date: _____

Name of the Chairperson: _____ Signature: _____Date: _____

Acknowledgements: Help received from others in the course of choosing research topics, designing the research proposal, offering materials, designing data collection formats, & data collection & data analysis and interpretation are recognized and thanked.

ACKNOWLEDGMENTS

I would like to express my genuine thanks to the TVT Institute (TVTI) for giving me the opportunity to conduct this research.

I am also glad and interested to acknowledge my advisor, **Dr. Yihenew Wondie,** for his constructive guidance, valuable feedback, and continuous support throughout the development of this proposal and the completion of the thesis.

Furthermore, I would like extend my appreciation to my best friend **Amanuel T.haymanot**, for his advice and encouragement during the preparation of this thesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ABS	Accounting & Basic Service
AI	Artificial Intelligence
AUC	Area under the Curve
BAW	Basic Accounting Works
BBC	Bar bending & concreting
BEI	Basic Electrical Installation
CSV	Comma separated values
DBA	Data Base Administration
EDA	Exploratory Data Analysis
FM	Furniture Making
FN	False Negative
FP	False Positive
GMFA	General Metal Fabrication & Assembly
HKO	Hotel & Kitchen Operation
HNS	Hardware & Network Servicing
HO	Hotel Operation
HR	Human Resource
ICS	Industrial & Control Service
ICT	Information Communication Technology
IEMD	Industrial Electrical Machine and Drives Servicing
ITSS	Information Technology Support System
KO	Kitchen Operation
LR	Logistic Regression

ML	Machine Learning
MOE	Ministry of Education
MSME	Micro, Small and Medium Enterprises
NLP	Natural Language Processing
NumPy	Numerical Python
OCM_MGT	Onsite Construction Management
OECD	Organization for Economic Co-operation and Development
RF	Random Forest
ROC	Receiver Operating Characteristic
SSOM	Secretarial Science & Office Management
TN	True Negative
TP	True Positive
TVET	Technical and Vocational Education and Training
WBS	Work Break Down Structure
WTC	Work force Training Center
XGBoost	Extreme Gradient Boosting

ABSTRACT

Although Technical and Vocational Education and Training (TVET) have gained increasing significance, many graduates remain unable to secure employment due to a persistent mismatch between the skills they acquire and the requirements of the labor market. This issue continues to contribute to high levels of unemployment among trainees. Within this context, this study attempts to analyze Tilahun Yigzaw TVET College in Tigray, Ethiopia. The research concludes that predictive analytics have the potential to enhance data-driven decision-making in TVET institutions by identifying critical factors influencing the skills gap and facilitating the alignment of training programs with industry demands.

The difference between Skills of the Technical and Vocational Education and Training (TVET) graduates and the needs of the industry is also an important concern to the graduate employability. The purpose of the given study is to predict the industry skill gap through predictive analytics and machine learning at Tilahun Yigzaw TVET College and analyze it. The study used secondary data that was collected as graduate tracer studies, student academic records, training data and data related to the labor market. The dataset was prepared to analyze by the use of data preprocessing methods of cleaning, encoding, feature selection and dataset balancing.

Three machine learning classification models were produced to classify the degree of skill gap among graduates namely, Logistic Regression, Random Forest, and XGBoost. Accuracy, precision, recall, and F1-score are the performance measures that were used to assess the models.

The findings indicated that the predictive performance of Logistic Regression was 90.29 accurate, the same was 91.79 with the Random Forest, and the best predictive performance was that of XGBoost with an accuracy of 96.60. The results suggest that machine learning models have the potential to detect factors that cause skill gap and make valid predictions to improve the results of training. The research achieves that predictive analytics have the possible to assist the data-driven decision making in TVET institutions to give the means of identifying the significant factors that influence the gap of the skills and assist in the arrangement of the training programs with the requirements of the industry. The findings can benefit educators, policymakers, and stakeholders in the industry to improve the way of designing the curriculum, enhancing the practical training opportunities, and improving the graduate employability outcomes.

CHAPTER ONE
1. INTRODUCTION

1.1. Back ground of the study

Technical and Vocational Education and Training (TVET) is strategically important in Ethiopia socio economic transformation. Efforts to industrialization of the country are also growing, and this is the reason why the pressure on the technical force is growing. Nonetheless, with the fastest growth of institutions, most graduates are not qualified and do not have the needed skills as demanded by the employers. This displacement between the skills trained in the institution programs and the skills required by industry, often termed as the industry skill gap is a key obstacle to the optimal exploitation of youth employment and economic growth.

Technical and Vocational Education and Training (TVET) is a world of education and training that prepares citizens to flexibly respond to changing technology and the labor market by bringing knowledge, attitudes and skills across different occupations and technologies. International practice has shown that the mere development of TVET does not solve the problems of unemployment and low productivity of the economy. Represents a significant milestone must react to the competency needs of the labor market and create a competent, motivated, and flexible workforce capable of driving economic growth and development. Understanding of the aforementioned national development requires clear and strong policy and strategy. The 1994 Policy of Ethiopia is a big sight in distinguish the main plan to create a lower- and middle-level, competent, motivated, adaptable, and innovative workforce [1].

Knowing the availability of the technical graduates, and the prevailing skill gaps assist in the developing superior skill development strategies and employment policies. This report highlights key statistics depicting trends in the number of technician graduates entering in the labor market. The report also sheds light on some of the causes for the skills mismatch, and unpacks what employers mean when they say graduates are not “employable. We hope that this research will be valuable in the development of strategies to mitigate skill deficiency and assist in bridging the gaps in existing skills [2].

Skill mismatch is the lack of balance between the skills and qualifications of the labor market and the skills and qualifications of the workers. The concept is broad and includes a number of distinct types of imbalances, both qualitative and quantitative. Mismatch in skills may happen in

various forms and it may include the lack of the necessary skills, education that is not compatible with the job, obsolete skills or excessive skills, or, the skills may be excessive or even too few in a given job. Mismatches can be horizontal (field of study mismatch) and/or vertical (over/under qualification). These various forms of skill mismatch differ in their expression, measurement, causes and effects they cause. Moreover, several types of skill mismatch might coexist with each other, which complicates the general problem [3].

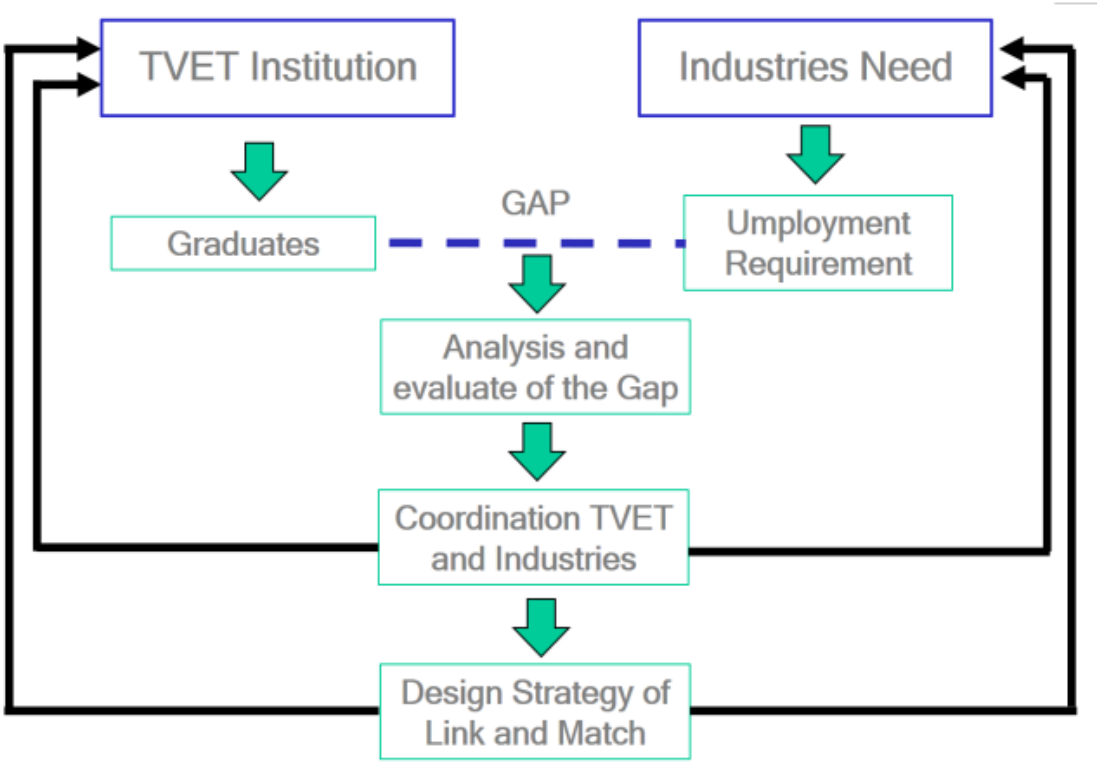


Figure 1 Gap Analysis between TVET and Industry Needs source [4]*)*

Graduates of TVET institutions provide labor in industries. Technical institutions therefore should have close relationships with the labor market to have the industry support them in enhancing the practical training such as offering work placements to trainees and exchange trainings to instructors [5].

1.2. Statement of the problem

Tilahun Yigzaw TVET College has some measures which can enhance graduate employability but at present there is an issue of the training modules in TVET programs and the industry employment expectation. The current structure of curriculum development and revision is not line with the drive of the needs of various industries. The institutions need a forecasted strategy that uses data to identify the gaps in skills and formulate appropriate intervention strategies. To improve the quality of TVET Institutions, it is necessary to analyze the key factors whether TVET has links and matches with industry needs. The results of the analysis of TVET and the industry will produce gaps which then need to be improved to minimize the gap between TVET and the industry. The gap between TVET and industrial needs can be narrowed through good communication and relationships between TVET Institutions and related industries and involving relevant parties such as the Ministry of Education, Ministry of Skill and Labor, Ministry of Trade, and other parties [4].

The study was conducted to address a specific local challenge at Tilahun Yigzaw TVET College, where tracer study data indicate a significant proportion of graduates remain unemployed or underemployed after completion of training. For example, recent tracer results show that a considerable percentage of graduates are not employed in their field of study, suggesting a mismatch between acquired skills and labor market requirements. Despite completing vocational training, many graduates lack industry-relevant competencies demanded by employers. This problem is not solely due to limited job availability but is strongly associated with skill gaps, as evidenced by discrepancies between training outcomes and employer expectations reported in institutional data. Existing approaches such as recruiting new graduates have not sufficiently addressed this mismatch, while structured identification of skill gaps remains limited. Therefore, this study aims to apply machine learning techniques to analyze tracer study data, identify patterns of skill deficiencies, and provide data-driven insights to support targeted interventions for improving graduate employability at Tilahun Yigzaw TVET College [6].

1.3. Objective

1.3.1. General objective

The main objective of this study is to develop and evaluate machine learning models for predicting industry skill gaps among TVET graduates.

1.3.2. Specific objective

- To assess the existing skill gap levels among graduates of Tilahun Yigzaw TVET College based on their acquired competencies and labor market requirements.
- To analyze the relationship between graduate characteristics (e.g., field of study, internship, sex) and skill gap levels.
- To develop and apply machine learning models to classify the skill gap levels of TVET graduates.
- To evaluate the performance and effectiveness of the developed machine learning models using appropriate metrics (e.g., accuracy, precision, recall, F1-score).

1.4. Research Questions

- What is the distribution and nature of skill gaps among graduates of Tilahun Yigzaw TVET College based on training and employment data?
- What key factors (e.g., internship, field of study, demographics) are associated with variations in skill gap levels?
- What patterns and relationships between training outcomes and labor market requirements can be identified using machine learning techniques?
- How effectively can machine learning models classify and predict the level of industry skill gaps among graduates?

1.5. Scope / Delimitations of the Study

This study looks at how predictive analytics and machine learning can be used to better understand and close the gap between the skills students are learning at Tilahun Yigzaw TVET College and the skills employers are actually looking for. It mainly focuses on the college's current training programs, data on where graduates end up working, feedback from employers, and ongoing trends in the job market. The research is limited to information gathered from

Tilahun Yigzaw TVET College and a few selected companies that often hire its graduates. It doesn't include data from other TVET colleges. More over as the machine learning algorithm is modeled after the available past data; they may not accommodate any sudden alterations in the industry needs or emerging trends that are yet to be reveled in the data. This study does not trying to build a one-size-fits-all solution. The approach seeks to develop a practical and data driven method that would assist the college to revise its training programs by aligning them with the requirements presented by the real world jobs and enhancing student job placement. While some of the insights may be useful for other institutions, the focus is specifically on what works best for Tilahun Yigzaw TVET College given its unique situation and resources.

1.6. Limitation of the study

Although this study wills aims to give meaningful insights into how predictive analytics and machine learning can help close the skills gap at Tilahun Yigzaw TVET College, there are a few important limitations to be aware of:

- Access to reliable data: the research relies on historical data of the College and employer feedback. Nevertheless, such data may not be complete, current, and consistent all the time, which may influence the degree to which the findings and predictions will be accurate.
- Limited coverage: the results may not be representative of the situation in other TVET institutions or industries in the country as the study will be based only on one college and a particular group of employers.
- Changes demands on the labor market: the labor market is dynamic and may be transformed at a very fast rate, particularly with the emergence of new technologies. Since we make predictions using previous and current data, there is a risk that new or rapidly emerging skills that have not yet been reflected in the data will not be represented in the predictions completely.
- Practical challenges: It takes the appropriate tools, technical expertise and computer power to develop and train machine learning models. These are aspects that may limit the extent or the quality of our analysis.

- Applicability Elsewhere: Also the result of the study can be useful guidance; they are specific to the environment of Tilahun Yigzaw TVET College. So, they might not apply directly to other institutions without adjustments.

1.7. Organization of the paper

The paper is organized into five chapters related to each other and systematically leading the reader to the research process.

- Chapter one presents the introduction of the study, where the background, problem statement, objectives, research questions, and scope are stated.
- Chapter Two examines the literature review and important concepts like predictive analytics, machine learning, skills gap, and employability.
- Chapter Three describes the research design including data sources, data collection procedure, and data analysis. The results are summarized and discussed in
- Chapter Four explores research results of the study by applying predictive analytics and machine learning models to reveal insights and developments and patterns of employment outcomes and skills gap.
- Chapter Five: Summary, Conclusion and Recommendations. The last chapter is a summation. It describes the main findings of the research, concludes using the data and gives the practical guidance on how the training programs in the college can be enhanced. It also points out areas where the future researchers can continue their research.

CHAPTER TWO
2. LITERATURE REVIEW

2.1. Theoretical Literature Review

2.1.1. Introduction

In today's rapidly evolving job market, predictive analytics and machine learning (ML) are playing a growing role in shaping education and workforce development strategies. For Technical and Vocational Education and Training (TVET) institutions, these technologies are not merely popular they're practical tools that can help better align educational outcomes with industry demands. By analyzing patterns and forecasting trends, predictive analytics and ML offer powerful ways to identify skills gaps and tailor training programs to meet real-world labor market needs. This theoretical literature review explores the core ideas, frameworks, and theories behind the use of predictive analytics and machine learning in the context of TVET. It focuses on how these technologies can support efforts to bridge the gap between the skills graduates possess and what employers are actually looking for. The goal is to highlight how data-driven approaches can lead to more effective training and improved employment opportunities for TVET graduates.

Skills gap is described as the difference between what students are taught in school and college and what employers are actually looking for when hiring recent graduates.

Skill gap has been acknowledged as a global issue and is experienced even in developed economies.

2.1.2. Defining Skills Gaps and Related Terms

There are numerous definitions that have been established for the term 'skills gap'. Based on the analysis of different studies, the conceptualization of the term is mostly guided by the direction in which an organization takes to address skills gaps.

Early conceptualizations of the skills gap generally describe it as a situation where the demand for specific skills exceeds their supply in the labor market. This definition highlights the imbalance between employer needs and the available workforce, forming the foundation for later interpretations that emphasize competency shortages and workforce misalignment [7].

The concept of skills gap has been defined in various ways in the literature, but most definitions share a common idea: a mismatch between the skills required by employers and those available in the labor market. Early definitions describe it as a situation where demand for certain skills exceeds supply; meaning employers are unable to find workers with the required competencies. Similarly, other studies define it as a mismatch between workforce capabilities and employer requirements, highlighting the imbalance between available talent and job demands [8].

More recent literature focuses on the employer view, where a skills gap refers to a scenario where the workers do not have the skills to effectively carry out the job tasks. In all these views, the theme is the same; there is a mismatch between the skills and job demands in the workforce, either because of the old skills, new technologies, or lack of proper training systems. Moreover, there is the phenomenon of occupational mismatch where people end up in occupations that are not suitable to their skills and qualification and in many cases; human capital is not fully utilized. The education and training gap accentuates the mismatch between competencies built in the course of formal education and those required by employers, frequently owing to the outdated curriculum. Later, the digital skills gap has been brought to the fore, but it entails the lack of digital skills needed in the contemporary workplaces as the world grows more technologically advanced. From the synthesis of these concepts, this study adopts the definition of skills gap as the difference between the skills employees possess and the skills required to perform job tasks effectively in a given labor market. This definition is suitable because it captures both employer expectations and workforce capability, which aligns with the focus of this research [9].

Skills gap refers to the mismatch of what employers require and what employees or job seekers are offering. These skills could be generic and therefore transferable across occupations (such as reading, writing, computer skills, or ‘soft’ skills), or they could be technical in that they are linked to a specific occupation or job. The gap occurs when the skills and knowledge required for a job or industry change faster than workers can learn them or when there is a shortage of workers with specific skills this may be because existing employees have outdated skills, which then creates a gap that can be difficult to fill, especially in a competitive labor market. In terms of the broader workforce, a skills gap is created when organizations struggle to find talent to meet their needs. Often, confusion arises between the terms ‘skills mismatch’, ‘skills gap’, and ‘skills shortage’. These terms are sometimes used interchangeably, yet they are quite distinct. The concept of Skills mismatch refers to an imbalance (an over-supply or under-supply) between

the types or level of skills available and what the labor market needs. Closely related to this Skills gap refers to a shortfall in the aggregate supply of a certain skill or set of skills broadly sought by employers (e.g., communication or computational skills). A more specific form of this challenge is known as Skill shortage refers to a shortfall in the supply of specific skills associated with particular occupations (e.g., a dearth of workers prepared to work as nurses or special education teachers) [10].

Another important concept discussed in the literature is Occupational mismatch refers to when an individual's skills, qualifications, or experience do not align with the requirements of their current job or the job market. This can guide to underutilization of skills and knowledge, as well as reduce job happiness and productivity. To address skills gaps, literature highlights two key strategies: up skilling and reskilling. Up skilling refers to improving existing skills to enhance job performance or career progression, while reskilling involves acquiring new skills to transition into different roles or occupations. Both approaches aim to align workforce competencies with changing labor market demands [11].

The Education and training gap refers to the gap between the skills and knowledge acquired through formal education and training programs and the skills required by employers. It can occur when the education system fails to keep pace with the changing demands of the labor market, resulting in graduates who lack the skills needed for available job opportunities [12].

With the rapid advancement of technology the digital skill gap has become increasingly prominent. According to the OECD, a digital skills gap refers to the disparity between the demand for digital skills in the workforce and the availability of individuals with these skills. As technology continues to advance rapidly, many industries require employees who possess digital literacy, proficiency in using digital tools, and the ability to adapt to new technologies [13].

To solve these problems experts stress that we need to improve our skills. Up skilling means getting skills or bettering the ones we already have to get better jobs or move up in our careers. Re skilling is different. It is about learning skills to switch to a completely different job or field. Both up skilling and re skilling are used to fix the gap, in skills people have and help them fit into the changing job market. The goal is to help individuals meet the demands of their work [14].

Over all the term skills gap is commonly defined as the difference “a gap between the skills an employee has and the skills they need to perform a job” [15].

These definitions indicate that the term is understood differently by different organizations, but for this research, the definition used by the Training Industry will be applied. People think that skills gaps are a problem that stops employees from doing their jobs. When companies have trouble finding people with the skills to do the tasks that the company needs to work well that is when a skills gap happens. The skills gap is what gets in the way of employees performing the tasks that they are supposed to do. The concept of skills gaps is a problem that the skills gaps cause, for employees [16].

Reported what could be described as empirically-supported situation, where skill culture was recommended and puts: “I am going to share statistics and observations to illuminate the 'Skills Gap' problem, then present why I think a Skills Culture is the mind set and Skills Based Approach them methodology/application to tackle this and other problems related to lifelong learning” [17].

Global economy, changing labor markets generate new demands, requiring the provision of new skills and knowledge from the Technical and Vocational Education and Training (TVET) sector. As the world becomes more connected the Ethiopian government has worked hard to improve the sector. The goal is to make it more accessible, better quality and more relevant. This will help modernize society and attract industries to set up production in the country. This is a challenging task since it not only aims at providing industries with adequately skilled manpower, but also embraces the rights of all young people to receive relevant and high-quality training, and thereby to reduce poverty. TVET is generally considered to be weak. This is similar to other countries in the Sub-Saharan region. Ethiopia does not have the resources to provide learners with a broad foundation of skills. These skills are required for getting a job in the sector. TVET, in Ethiopia needs to provide these skills to learners. In addition, there is no well-regulated labor market [18].

Many cities in the United States have experienced job loss due to a shift from industrial based to information and service economies, as well as the outsourcing of labor jobs to overseas locations. In the absence of industrial jobs once occupied by low-skilled workers, the problem that compelled this study was perceived gaps between the skills required by the employers who now hire such workers and the actual skills those workers have.. The research questions guiding this

study were designed to help explore and understand the needs of employers and whether the WTC was adequately training low-skilled workers to meet those needs. The WTC training programs were pretty good because they taught people the skills that employers said they wanted. The World Trade Center did not teach skills for some jobs though. Some employers did not know about the World Trade Center services. Did not use them when they were looking for new people to hire. The World Trade Center curricula were still good because they had the skills that employers said they needed for the World Trade Center jobs. Recommended strategies to increase collaboration between the WTC and local businesses are provided in a policy project paper. Positive change happens when Work Training Centers prepare people who do not have jobs and then place these people in jobs and careers [19].

TVET is about learning things that help people get jobs. It includes education and training that teaches people skills for different kinds of work. TVET is for people who want to learn how to do something to earn a living. People can learn these skills in school. After they finish school TVET can happen when people are in school or after they finish high school or even when they are, in college. TVET is something that people can do their lives. It includes work-based learning, continuing training and professional development that may lead to qualifications. TVET offers skill-development chances that fit national and local needs. TVET is now more focused on getting people ready for work in the Information Age. This is because we are moving away from the Industrial Age. The world of work is changing fast. We need different skills now. TVET helps prepare workers with the knowledge, for these changes. It also meets human resource needs that come with this shift [20].

2.1.3. Using Data-Driven Techniques to Close Skills Gaps

Predictive analytics is really helping companies manage their workers better. It lets them know what skills they will need in the future. They can also find out if their employees are missing some skills. Predictive analytics is a tool in workforce management. It enables organizations to anticipate skill requirements and identify gaps in employee competencies, with predictive analytics. By leveraging data from various sources, such as performance metrics, industry trends, and emerging technologies, predictive models can forecast the demand for specific skills and identify areas where current employees lack expertise. This approach helps companies plan

ahead for hiring, training and re-training so they stay competitive in a changing job market. It helps them prepare for needs. Companies can use analytics to support their HR plans by figuring out what kind of workers they will need based on company goals, market conditions and changes in the industry. This way business can make sure their workers have the skills for the future avoid having workers with the wrong skills and reduce the risk of not having enough workers. By using analytics in workforce planning organizations can make better decisions work more efficiently and adapt quickly to new business challenges. They can align their workforce capabilities, with demands. It also enables them to minimize the risk of talent shortages and reduce skills mismatch. This approach is particularly beneficial in sectors facing fast technological advancements or regulatory changes, where the skills of today may not meet the requirements of tomorrow. This means predictive analytics can help companies plan for talent needs. It gives businesses a way to look ahead and manage their workforce effectively [21].

Predictive analytics is a powerful approach used to forecast future events or trends based on historical data and statistical algorithms. At its core, predictive analytics involves the use of various techniques to analyze past data and identify patterns that can be leveraged to make informed predictions about future outcomes. This process often involves complex statistical models and algorithms that can process and interpret vast amounts of data to generate actionable insights [22].

Predictive analytics is getting really popular because we have much real world data available. So it is very important to choose the predictive algorithm. There are predictive algorithms that people use for predictive analytics. It is still hard to pick the right algorithm for the real world dataset and the problem we are trying to solve. Many people use algorithms for predictive analytics and the key is to find the best one for the specific real world dataset and problem like the ones we are studying. Predictive analytics and predictive algorithms are crucial, in this process [23].

Machine learning is the study of computer algorithms that provides systems the ability to automatically learn and improve from experience. It is generally seen as a sub-field of artificial intelligence. Machine learning algorithms allow the systems to make decisions autonomously without any external support. Such decisions are made by finding valuable underlying patterns within complex data [24].

Both machine learning and predictive analytics are used to help make sense of data. However, predictive analytics is category of data analytics. Machine learning is a tool that is used in analytics. Predictive analytics is like a box that has lots of things inside including machine learning. Machine learning uses data to make good guesses about what might happen. Predictive analytics helps companies by giving them ways, to group things predict what will happen in the future and make models of things [25].

As the workforce changes universities are really struggling to give students not just book knowledge but also practical skills and skills that can be used in many jobs so they can succeed in their careers [26].

People are starting to realize that it is really important to know how students are doing on. This is why there is a lot of research, on machine learning models. These models can help us figure out how students will do in the future accurately. The idea is to use these models to get a sense of student performance. This is where machine learning models come in and they can be used to assess student outcomes [27].

Embracing data-driven methods can ensure a more agile and adaptable workforce, ready to meet the challenges of tomorrow. As we move forward, continuous learning and data-informed decision-making can be key to closing skills gaps and driving sustainable growth [28].

Machine learning algorithms are organized into taxonomy, based on the desired outcome of the algorithm [29].

Machine learning is an evolving branch of computational algorithms designed to emulate human intelligence by learning from the surrounding environment [30].

Machine learning is the study of computer algorithms that provides systems the ability to automatically learn and improve from experience. It is generally seen as a sub-field of artificial intelligence. Machine learning algorithms allow the systems to make decisions autonomously without any external support. So when we make decisions like this we look for patterns that are really important in complicated information. We can group machine learning algorithms into main categories. These categories are based on how the algorithms learn, what kind of data they take in and give out and what kind of problems they solve. The main categories of machine learning algorithms are machine learning algorithms they're supervised, machine learning algorithms that are unsupervised and machine learning algorithms that use reinforcement

learning. We use machine learning algorithms to make decisions and machine learning algorithms are really good, at finding patterns in data. There are a few hybrid approaches and other common methods that offer natural extrapolation of machine learning problem forms [31].

Machine learning algorithms are organized into taxonomy, based on the desired outcome of the algorithm [32].

In today's world machine learning is really good at giving us estimates in many different applications. Machine learning learns a lot from the data it gets from projects that are already done. Researchers have shown that machine learning can do the job by itself once it figures out how to work with the data. This is something that researchers have found out. It is a big deal, for machine learning. Machine learning is getting better and better at doing things on its own [33].

This paper is about HR Analytics in the workforce. It looks at how HR Analytics helps with planning. Data-driven approaches are changing how human resources are managed. The focus is on HR Analytics and its role in the workforce.HR Analytics is important, for planning and improving human resource management practices. It uses data to make decisions. This helps HR to be more effective. The goal is to see how data can improve HR. By integrating advanced analytical devices and technologies such as artificial intelligence and machine learning, HR enables analytics organizations to predict the needs of the workforce, adapt talent management to adapt to talent management and increase employee engagement and retention. The study highlights the moral and organizational challenges associated with adopting HR analytics and outlines the importance of transparency, fairness and data privacy. In addition, it identifies emerging research opportunities in implementing HR analytics to develop workforce models including remote and gig employment. Ultimately, this letter argues that today's dynamic business environment required HR analytics to plan agile, responsible and strategically aligned workforces [34].

This study explores the role of HR analytics in identifying skill gaps and optimizing training needs within Micro, Small, and Medium Enterprises (MSMEs) in Andhra Pradesh. The research is focused on providing data-driven solutions to enhance workforce capabilities and improve employee retention and productivity [35].

The Skill Gap Analysis Using Machine Learning project is about figuring out what skills people already have and what skills they really need to do their jobs. This Skill Gap Analysis Using Machine Learning project checks out peoples resumes seeing what skills they actually have. It is like taking a look at what people can do and what they are missing to do their jobs better. The main goal of the Skill Gap Analysis Using Machine Learning project is to find the difference, between the skills people have and the skills they need for their jobs. It then finds out what skills they are missing. Give them a plan to learn those skills. The "Skill Gap Analysis Using Machine Learning" project helps people get the skills they need. It does this by checking their resumes and suggesting what they need to learn. The project uses Machine Learning give people a path to improve their skills. The goal is to help people have the skills for their jobs. It examines resumes to find skill gaps. Then it gives a plan, for learning skills using Machine Learning and NLP [36].

Skill shortages are a problem for our society. They hamper economic opportunities for individuals, slow growth for firms, and impede labor productivity in aggregate. So it is really important for people who make policies and educators to be able to know about skill shortages before they happen this will help them make things better and reduce the effects of skill shortages, on society and skill shortages will not be so harmful. This study uses a Machine Learning method to figure out when we will have a lack of people, with certain job skills. The Machine Learning approach is used to predict skill shortages [37].

The random forest algorithm, which was proposed by is really good, at classification and regression. It works by combining decision trees that are made in a random way. These trees make predictions. Then the algorithm averages them. This approach is great when there are variables but not many observations. The random forest algorithm has done well in such cases [38].

A logistic regression model is used to understand how one thing that can only be one of two choices is related to one or more things that can affect it. This thing that can only be one of two choices is called the dependent variable or the outcome variable. The other things that can affect it are called variables or covariates or explanatory variables. The logistic regression model looks at the relationship, between the dependent variable and the independent variables [39].

Logistic regression is a powerful statistical method widely used in health research to model and predict the probability of binary and categorical outcomes [40].

A random forest is a type of machine learning model that people use for classification and forecasting. This random forest model is really good at helping us make predictions about things. We use the forest model to look at a lot of information and make smart guesses about what might happen. The random forest model is very useful, for classification and forecasting because it can look at different things at the same time. To train machine learning algorithms and artificial intelligence models, it is crucial to have a substantial amount of high-quality data for effective data collecting. System performance data helps a lot, in making algorithms. It also makes software and hardware work efficiently. We use it to study how users behave. This helps us find patterns. With these patterns we can make decisions. We can even predict what might happen next. It also helps in solving problems. All this leads to results and more accurate answers. It is really useful as mentioned in some studies [41].

Tree boosting is a highly effective and widely used machine learning method. In this paper, we describe a scalable end to-end tree boosting system called XGBoost, which is used widely by data scientists to achieve state-of-the-art results on many machine learning challenges [42].

The main goal of machine learning is to get information from the huge amount of data that is made every day. This information is then used to learn and do something with it. Machine learning is used in different areas like understanding language finding patterns making search engines better helping doctors diagnose people and understanding biology and chemistry. There is a machine learning tool called Boost that is really good at understanding complicated systems. Boost is thought to be the machine learning tool because it is very good at predicting things it is easy to understand and it can be used in many different ways. Boost is a kind of tool that can be used on many different computers at the same time. It is very powerful. Boost uses a way of working with data to make machine learning algorithms better. It is also very good, at solving problems that involve a lot of data. It can do this quickly and accurately [43].

Gradient boosting algorithms have become a vital component in the realm of machine learning, thanks to their outstanding performance in a wide range of predictive modeling tasks [44].

2.2. Empirical Literature Review

Table 1 Empirical Literature Review

Author	Year	Study Title	Data Set Used	Key Findings	Gaps Identified
Sudha Rajeev Menon	2025	The Role of HR Analytics in Workforce Planning	Secondary data (academic journals, industry reports, interviews)	HR analytics supports workforce forecasting and talent management using AI/ML	Limited application in dynamic environments
Greta Braun	2024	Bridging Skill Gaps: A Systematic Literature Review	40 peer-reviewed articles (PRISMA method)	Stakeholder collaboration is essential for closing skill gaps	Lack of empirical validation and implementation roadmap
Muhammed Busari	2025	Predictive Analytics for Skill Gaps and Workforce Requirements	HR internal data + labor market data	ML models forecast skill gaps and workforce needs	Data quality issues and algorithmic bias
Dawson et al.	2020	Predicting Skill Shortages in Labor Markets	Job listings and government databases	ML achieved 83% F1-score in forecasting skill shortages	Not applicable to TVET systems
Singh & Verma	2023	Skill Gap Identification in Higher Education	Student academic and curriculum data	ML identifies mismatch between curriculum and industry needs	No real-time labor market integration
Ahmed et al.	2022	Data-Driven Workforce Planning	Enterprise HR datasets	Improved recruitment and training efficiency	Limited scalability to education sector
Li & Zhang	2023	AI-Based Skill Forecasting for Industry 4.0	Industrial workforce + automation data	Improved prediction of future skill demand	Limited focus on developing countries
Njoroge et al.	2021	ML in Education Performance Prediction	University student performance data	Predicts student success and dropout risk	Focuses on academic performance only
Kurniawan et al.	2022	Big Data Analytics for Workforce	LinkedIn job data	Identifies emerging skill trends	No integration with training systems

		Skills			
Tesfaye & Mekonnen	2024	TVET Skill Mismatch in Ethiopia	Survey of TVET graduates and employers	High mismatch between training and industry needs	No ML-based prediction approach
Brown & Wilson	2023	AI in Human Resource Management	HR case studies and corporate systems	AI improves workforce planning	Ethical and transparency concerns
Patel et al.	2024	Predictive Workforce Demand in Emerging Economies	National labor statistics	ML improves workforce demand forecasting	Weak policy implementation alignment

2.3. Summary of the gap analysis:

The root cause is a widened gap in the TVET training output at Tilahun Yigzaw TVET College and skill requirement in the industry due to the gap is primarily driven by outdated curricula that are unable to keep pace with rapid technological advancements, weak stakeholder collaboration, and the absence of data-driven forecasting mechanisms. Empirical results inform the important implementation gaps like invalidated strategies, small/non generalizable samples, and the possibility of algorithm bias, and a lack of an operational roadmap to implement the predictions in curriculum changes, aggravated by resource constraints unique to Ethiopia. This mismatch secures in graduate underemployment, discontent among the employers and slow economic progression, which requires a combined machine learning framework that can bridge the divide between the skills-gap forecasts on a localized scale, lawful curriculum reforms, and ethical engagement of the stakeholders to bridge the divide.

The key tool used to fill the skills gap is Machine Learning-based prediction analysis supported by data-driven skill significance analysis. It allows decision-makers with foresight in to labor market/industry/ requirements, allowing hands –on skills growth, program updates, and policy involvements.

Machine learning is implemented as a decision-support workflow that classifies each incoming cohort into High, Medium, and Low skill-gap groups at training intake based on their feature

profiles. This early classification enables targeted interventions, such as curriculum adjustments, internship prioritization, and focused training for High skill gap groups.

2.4. Conceptual Framework

This research paper presents a practical frame work that can be used to show how machine learning and predictive analytics can be incorporated in to the environment of Tilahun Yigzaw TVET College to enable the minimization of the gap between the skill level of graduates and demands of the industry. The frame work aims at evidencing based decision making in the areas of curriculum development, training design, and graduate employability improvement. It is a conceptual prism and practical tool to comprehend compared how institutional information can be process in to actionable information to stakeholders.

A. Data sources (Input Variables)

The framework makes use of the institutional and labor market data that are applicable to the training results in TVET. The main data sources will be graduates have acquired during training. Furthermore, curriculum specific information is also included in order to record program design, course content, work based training hours, and curriculum revision rate. In order to represent the demand of the demand of labor market, labor market data like job posts, skills required and industry trends are added. In addition, the employer's feedback, based on the organized interviews provides information on the level of employer satisfaction, the skills needs, and training gaps. Collectively these contributions can be viewed as the important determinants of graduate employability and skills match.

B. Predictive analytics and machine learning process

The framework is built around a machine learning-driven predictive analytics system that takes the data collected and patterns and relationships. The first step involves the data cleaning, transformation, and feature selection preprocessing. XGboost, Random Forest and Logistic Regression as the models of supervised learning are then trained to predict employment status, types if skill gap and skill gap level among graduates. The system forecasts the level of skill gaps (high, medium, or low), the graduate employability status, and provides the growing trends of skills demands, which are applicable to the TVET context through these modeling processes.

C. Output variable (Dependent variable)

The framework presents decision-support information to the administrators of TVET, instructors and policy makers through the outputs of the framework. These are skill gap category based on the occupation; sector, program, total training hour or skill based which assists in areas needing intervention. Curriculum alignment score is produced to show how well any training programs being offered are aligned with the labor market needs. Also, the framework generates a graduate employability forecast, skill gap level and types of skill gap which approximates the probability of employment after existing training routes. Depending on these outputs, practical recommendations are drawn, including curriculum revision, more training on soft skills, and better internship and industry attachment programs.

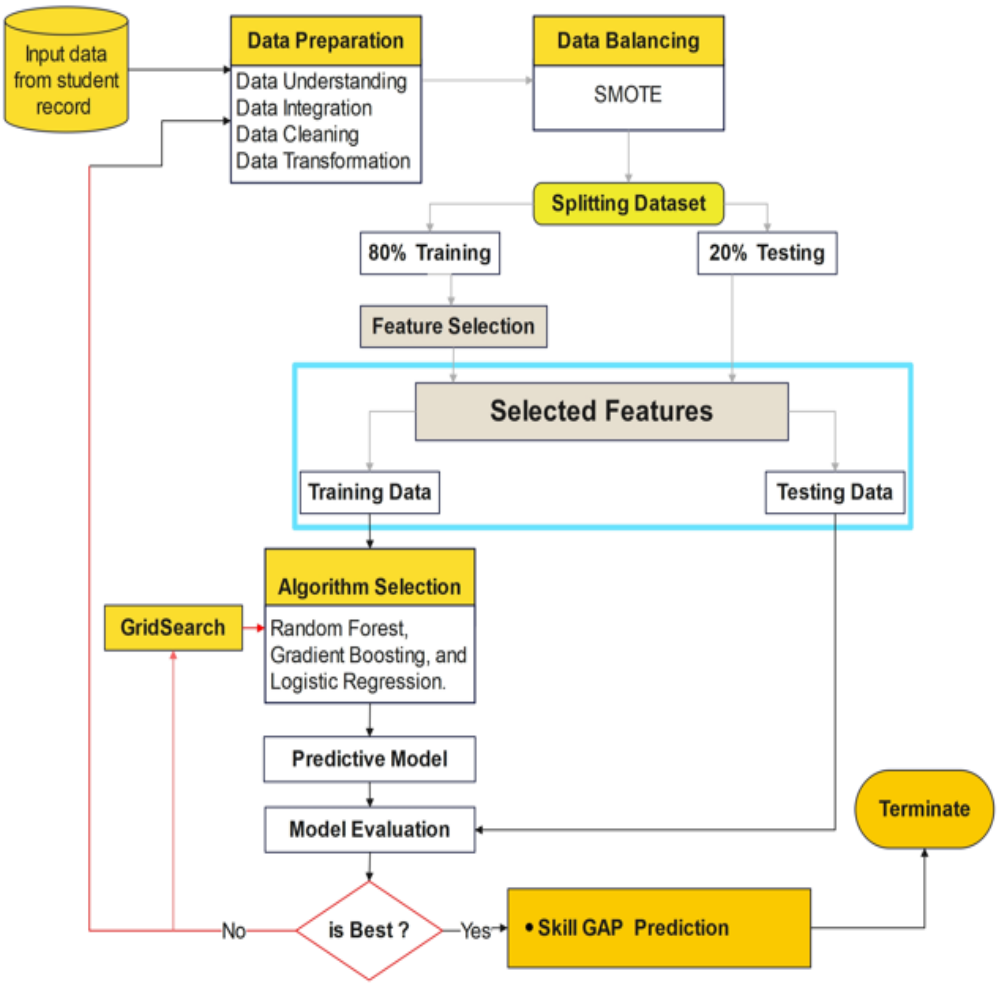


Figure 2 Conceptual design process

CHAPTER THREE
3. RESEARCH METHODOLOGY

3.1. Introduction

This chapter describes our approach to addressing research problems and achieving the study's objectives. This section provides a detailed description of the proposed prediction model. This includes data collection; description, pre-processing, hyper-parameter optimization based on nature, model development and selection, evaluation, and model explains ability.

3.2. The Proposed Prediction Model Architecture

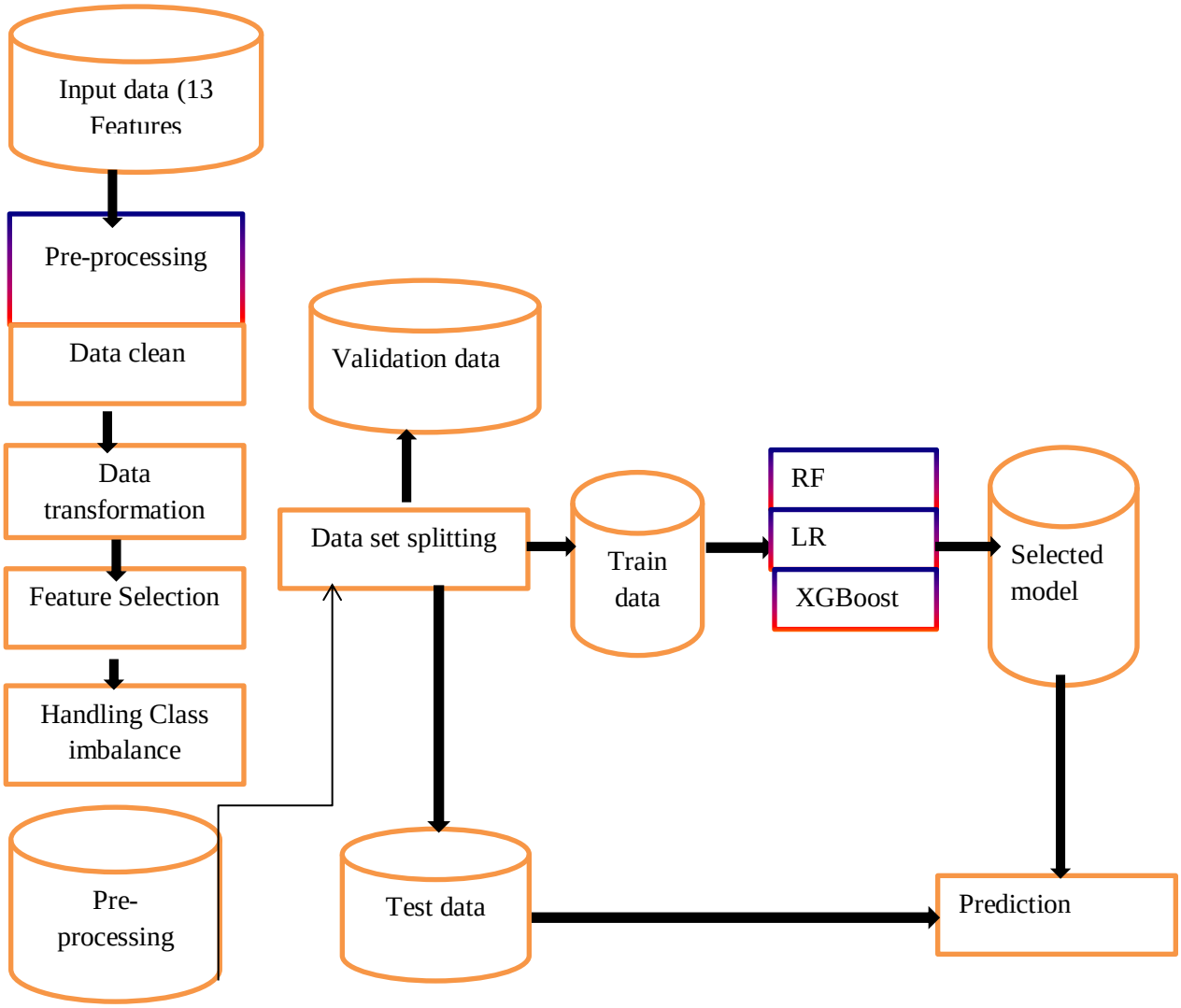


Figure 3 Proposed Architecture Model

The above Figure 3 illustrates the general approach of the Predicting analytics to bridge Industry skill gap Level prediction model. This model has four major components. The first part involves input the collected data and then preparing the data using different preprocessing techniques, which consists of feature selection, data transformation, and data balancing. The second step involves splitting our dataset into training-validating and testing sets. The third section involves model building by training and validating the machine learning models (LR, RF, XGBoost). The forth section involves models explain ability is applied to the selected model.

The final model used in this study is Extreme Gradient Boosting (XGBoost), selected based on superior performance over Random Forest and Logistic Regression. As shown in Figure 3, the model takes 13 preprocessed input features (after label encoding and normalization) and applies an XGBoost classifier configured with n_estimators = 100, max_depth = 6, learning_rate = 0.1, and objective = multi:softprob. It outputs probabilities for three classes (High, Medium, Low), with the highest probability determining the final prediction.

3.3. Research Design

A research design is the structured collection and analysis of data, along with a working scheme, in a manner that aims to address the research problem. This study accepts a quantitative predictive analytics research design connecting machine learning (ML) methods to integrate to find, model, and prediction industry skill gap in relation to graduate competencies at Tilahun Yigzaw TVET College. The design allow a systematic analysis of structured data sets, development of predictive models, and generation of data driven insights to support TVET curriculum development, industry arrangement, and training involvement. We have conducted results with a step-by-step procedure, such as data collection, data preprocessing, and splitting the processed data into training, validation, and test sets, and finally, model building. Therefore, we have examined the feature selection process, which is suitable for identifying the Skill gap level. The data preprocessing tasks, including handling missing values, outlier detection, and feature engineering, have been carefully undertaken. Additionally, the classification techniques have been intensively assessed to determine the most appropriate model for predicting the Skill gap level. Importantly, the study correctly defines the variable relationship: the Skill Gap Level is treated as the dependent variable (target variable), while factors such as Sex, Age, Level,

Occupation, Sector, Program, Total Training Hours, and Internship are considered independent variables (predictors). These independent variables are used to explain and predict variations in the dependent variable.

3.4. Software and Programming Tools

The data must first undergo pre-processing to create the correct format and data type required for accurately predicting the Skill gap level. Proper data preparation is essential for the classification process to work more accurately and efficiently. Consequently, data transformation, feature selection, and handling of imbalanced datasets have been conducted. The pre-processing task was carried out using a Python a comprehensive suite of packages like NumPy, Matplotlib and Pandas. A Python software library, offers advanced data manipulation and analysis capabilities, allowing you to work with numerical tables and perform statistical data analysis, import CSV files, align data, handle missing values, and preprocess the data.

3.5. Data Collection

One of the most important aspects of research is gathering data. The process of gathering, quantifying, and evaluating exact data from several relevant sources in order to address research questions, respond to inquiries and evaluate findings is known as data collection. Data collection is vital in research as it aids in undertaking research questions, challenges, and conclusions. This involves obtaining, quantifying, and evaluating precise data from appropriate sources to fulfill particular research goals. Data collection is crucial for enhancing products and services, facilitating informed business choices, and addressing research issues. It is vital to identify the required data and collection methods before starting the data collection process. The secondary datasets include tracer study reports, curriculum records, job market information, and employer feedback datasets obtained from Tilahun Yigzaw TVET College and associated industry stakeholders. These datasets were previously gathered through structured methods such as surveys, institutional reporting systems, and labor market assessments conducted by the respective organizations

A precondition to any machine learning model project is the data itself. For the purpose of this research the data is collected from Tilahun Yigzaw TVET College.

Labeling Procedure

(a) The Skill_GAP_Level labels (High, Medium, and Low) were assigned by domain experts, specifically TVET instructors and industry supervisors affiliated with Tilahun Yigzaw TVET College, based on their professional evaluation of students' competencies.

(b) A structured assessment rubric was used, which evaluates the gap between students' demonstrated skills and the expected occupational standards this rubric incorporates indicators such as practical performance, assessment scores, and supervisor evolutions.

(c) The rubric was adapted from institutionally recognized competency frameworks and further refined by the researcher to fit the study context, making it a hybrid of institution-provided and self-designed criteria.

(d) The labels were generated independently of the input features used in the machine learning

3.5.1. Data description

A data set description is a detailed overview of the content and characteristics of a specific data set. It typically includes information about the source of the data, the variables or columns included in the data, and any relevant details needed to understand and analyze the data effectively. This description helps users understand the data and its potential uses before working with it. The goal of data description is to provide upcoming researchers with the necessary information to interpret and make use of preserved or gathered datasets, as well as to comprehend the characteristics and patterns of the data. For this research, data was gathered from Tilahun Yigzaw TVET. The dataset comprises 15 attributes.

The data in this research has various categorical variables explaining the demographic, educational and working features of TVET graduates, and details on the skills gaps. The variables will be Sex, Occupation, Sector, Level, Program, Internship, Employment Status, Skill Gap Type and Skill Gap Level. These variables are significant in the analysis of how training background is related to any skill gaps in the labor market.

The Sex variable will determine the gender of the respondents' distribution; this will assist in checking whether there are diverse skill gaps between the male and female graduates. Occupation and Sector indicate the nature of work and industry sector where graduates are engaged so that the study can examine variation in the level of skills demanded among different sectors. The Level and Program variable is used to indicate the level of education and field of

training in the TVET system and this is a significant variable that determines the competency and employability of graduates.

Moreover, the Internship variable shows the level of whether the graduates have undergone hands-on industry training in the course of their education, which is likely to affect their preparation to work in the market. The Employment Status variable indicates the presence of graduates who are and those who are not employed, which is used to determine the relationship between skill gaps and the results on the labor market. Lastly, Skill Gap Type and Skill Gap Level are the type and the level of the identified skills mismatch, which are the key variables to analyze and forecast skill gaps with the help of machine learning models.

All these variables combined will give a thorough foundation of investigating the elements that cause the skills gaps among the TVET graduates and will be able to assist in creating predictive models that will be used to determine in which areas training programs need to be refined.
In this paper, the set of categorical variables was transformed into numerical values in a label encoding procedure to allow application in machine learning models. The Sex variable was coded into 0 Female (F) and 1 Male (M). Such transformation enables the model to take gender information and retain the original categorical meaning. The need to encode categorical variables into numerical form is attributed to the fact that most machine learning algorithms have to be trained and predicted using numeric data.

Likewise, the Sector variable that denotes field of training/employment area of TVET graduates was coded into numeric variables. These are 0 = Accounting & Finance, 1 = Business, 2 = Construction, 3 = Electrical, 4 = Furniture, 5 = Hotel and Tourism, 6 = ICT, and 7 = Manufacturing. Such classification assists the model to examine the difference in skill gaps in various industrial sectors and training areas.

Besides, the target variable Skill Gap Level was coded 0 = High, 1 = Low and 2 = Medium. This variable will be used to evaluate how severe the skill mismatch between graduate competencies and industry requirements will be. Using the numerical labels of the level of skill gaps by converting them into numerical values, the machine learning algorithms will be able to efficiently classify and predict the level of the skill gaps among graduates.

In general, this process of encoding enhances the compatibility of the dataset with machine learning algorithms, and the meaning of the categorical variables does not get lost. It also enables efficient processing of the data and aids proper prediction of the level of the skill gap in the developed predictive models.

Table 2 Dataset Description

No	Variable Name	Description	Data Type
1	Stu_ID	Unique student identifier	Integer
2	Sex	Gender of graduate(Male/Female)	Categorical
3	Age	Age of graduate students	Numerical
4	Sector	Employment sector	Categorical
5	Occupation	Current job role	Categorical
6	Level	TVET Qualification level(II-V)	Ordinary
7	Program	Regular/Extension	Categorical
8	Graduate_Year	Year of graduation	Numerical
9	In_TVET_Traning_hours	Training hours received in TVET	Numerical
10	In_Company_Traning_hours	Training hors received in industry	Numerical
11	Total_Traning_hours	Sum of TVET & industry training hours	Numerical
12	Internship	Participating in internship YES/NO	Binary
13	Employment_Status	Employed/Unemployed	Categorical
14	Skill_GAP_Type	Type of skill gap(technical, soft, digital)	Categorical
15	Skil_GAP_Level	Degree of skill gap (High, Low, Medium)	Categorical

3.5.2. Study Area

The study conducted at Tilahun Yigzaw TVET College, which is a typical TVET College providing vocational education and skills training to a variety of technical fields.

3.5.3. Population and Sampling

The study population consists of graduates from Tilahun Yigzaw TVET College utilized the entire available dataset of graduate records; it follows a census approach rather than a sampling-based design**.** Therefore, no sampling technique was applied for selection purposes. The study

population consist 5,000 graduate records available in the college's alumni database was utilized for the analysis of graduates from Tilahun Yigzaw TVET College. The method is a census approach**,** to guarantee the representation of the sectors, programs, Occupation level and graduate year. The final data set used for model training comprised records containing demographic attributes, Training hours, occupational data, and skill gap assessment results.

$$n = N \tag{3.1}$$

Where:

n= sample size

N= Total population size

3.6. Data Sources

3.6.1. Secondary data:

The secondary data used in this study were Tilahun Yigzaw TVET College academic records, college tracer studies, and market assessment report which gave information about the academic performance of the students, employment outcome after graduation and also the industry skills requirements which were used to determine the gaps in skills and predictive models to increase the training and employment adaptability.

3.7. Description of variables

3.7.1. Independent variables

- Sex
- Age
- Sector
- Occupation
- Level
- Program
- Graduate_Year
- Internship
- In_TVET_training_hrs
- In_Companytraning_hrs
- Total_Traning_hrs
- Skill_GAP_Types
- Employment status

3.7.2. Dependent variable

- The dependent variable in this study is the Skill_GAP_Level, which represents a single multiclass categorical target variable used for classification. This variable is defined at three ordinal levels: High, Medium, and Low, representing the degree of skill gap among TVET graduates.

3.7.3. Exploratory data analysis (EDA)

Before development of the model, data analysis using exploratory data analysis (EDA) was used as an initial analytical step to have knowledge of the structure, distribution and the nature of data. The EDA procedures were employed to observe the missing values, outliers, the distribution of variables and also the relationship between variables. Numerical variables were calculated using descriptive statistics measures like mean, median, standard deviation, minimum, and maximum. In case of categorical variables frequency distributions and percentages were obtained. Visual exploratory methods were used such as histograms, box plots, bar charts, and correlation heat maps to investigate patterns and trends and possible relationship between features. The insights gained in the course of EDA were used in the data preprocessing decision, feature selecting, and model creation.

3.7.4. Data Analysis Techniques

The cross-tabulation analysis was used to analyze the relationship between significant categorical variables in the dataset. In particular, cross-tabulation was employed in order to justify the distribution of Skill Gap Level in such variables as sector, sex, and participation in the internship at Tilahun Yigzaw TVET College. This method helped the research to detect patterns, associations, and proportional differences among categories and then implement machine learning models. Findings of the presented cross-tabulation analysis formed the groundwork of the selection of features and the construction of a model.

3.8. Data Preprocessing

Data preprocessing include converting raw data into suitable for machine learning models, it is essential stage in model development. Data preparation is processing applied raw data to ready it for the following basic steps. Data collection is gathering of raw information, translating it into a

form interpretable and analyzable by computers and machine learning algorithms. Following data collection, the information undergoes preprocessing, encompassing tasks such as data cleaning, transformation, feature selection, and addressing imbalanced datasets. Machine learning algorithms are then applied to this preprocessed data. Data cleaning entails correcting or removing incorrect, corrupt, improperly formatted, duplicate, or incomplete data. Data transformation involves technically transferring data from one format or structure to another without altering its original content. Feature selection refers to selecting the most vital variables (features) crucial for predicting the outcome, essentially identifying characteristics pivotal for the model's accuracy.

In machine learning it is common to encounter data that is not clean or properly formatted. Therefore, it is important to clean and format the data before performing any data operations. This is why the task of data pre-processing is utilized. Preprocessing can be an iterative process, where a series of steps may need to be repeated until the data is correctly formatted for analysis. It is essential to exercise attention during preprocessing to avoid introducing bias that could potentially impact the outcomes of high risk. In the area of machine learning, thorough data investigation and preparation are essential tasks that significantly influence the accuracy of classification. The integrity of any decision-making process is compromised in the presence of incomplete or inaccurate datasets [45].

Data preprocessing in machine learning involves cleaning and structuring raw data to prepare it for training and developing machine learning models with various algorithms. Real-world data often includes noise, missing values, and may be in an unusable format, making it unsuitable for direct use in machine learning models. Data preprocessing is essential for making the data compatible with a machine learning model, thereby enhancing the accuracy and efficiency of the model. Common anomalies such as missing values, noise, inconsistencies, and redundancy are often found in raw data, affecting the performance of subsequent learning processes. Therefore, a preprocessing step is usually undertaken to minimize the effects of these data abnormalities [46].

After data collection data preprocessing is always the first step in any machine learning. The collected data that we have collected are raw data some of them are incomplete which needed further preprocessing for preparing of the data for model development. The aim of involving the data preprocessing techniques is to prepare it in a compatible format for machine learning and to improve the quality of the data so that the models can perform in more accurate way. In this

study we do an important thing to get our data ready. We handle the null values. We also encode the values. Then we select the features and balance the dataset. This is all part of the preprocessing activities, for the dataset. The main things we do are handling missing values encoding values selecting the best features and balancing the dataset. It incorporates the process of cleaning; processing and sorting raw data in order make it accurate, consistent and model ready. A crucial step in the data analysis process is preprocessing, which involves converting raw data into a format that computers and machine learning algorithms can understand [47].

Before model development the data set, passed through the following steps.

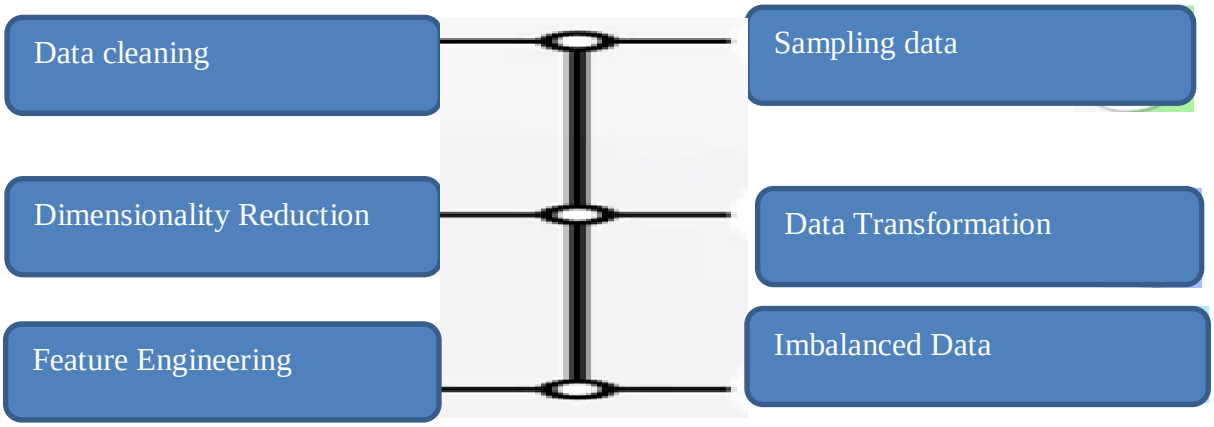


Figure 4 Data Preprocessing

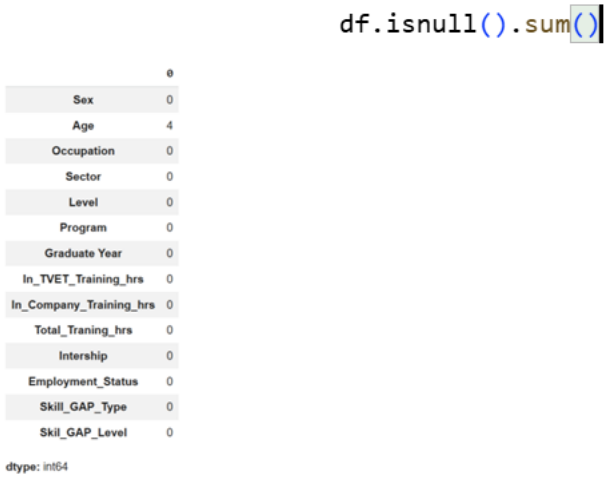
3.8.1. Data normalization

Normalization is the process of transforming data to a standardized format or scale. This technique is applied to ensure that the features or attributes have consistent ranges and distributions, which make the model interpretable. Normalization is a pre-processing stage of any type problem statement. Especially normalization takes important role in the field of soft computing, cloud computing etc. for manipulation of data like scale down or scale up the range of data before it becomes used for further stage. In study performing label encoder on data features is a crucial preprocessing step that serves numerous purposes machine learning. Using label encoder improves the performance of sensitive algorithms, addresses the impact of varying feature ranges and outliers, enables more meaningful feature comparisons, accelerates the convergence of optimization methods, ensures consistent feature importance evaluation, enhances data visualization, and promotes numerical stability [48].

3.8.2. Handling Missing values

Handling missing values in machine learning refers to the process of identifying and managing empty, NaN, or null entries in a dataset to prevent reduced model performance, bias, or algorithm failure. Common techniques include deletion of incomplete records or imputation using statistical measures such as mean, median, or mode.

Missing value analysis revealed that the variable Age contained 4 missing observations, while all other variables had complete data. Since the proportion of missing values was very small (0.08%), the missing Age values were handled using median imputation, which is robust to outliers and preserves the central tendency of the variable. No records were deleted in order to maintain dataset integrity and avoid information loss.



```
df.isnull().sum()
```

	0
Sex	0
Age	4
Occupation	0
Sector	0
Level	0
Program	0
Graduate Year	0
In_TVET_Training_hrs	0
In_Company_Training_hrs	0
Total_Traning_hrs	0
Intership	0
Employment_Status	0
Skill_GAP_Type	0
Skil_GAP_Level	0

dtype: int64

Figure 5 Handling Missing Values

3.8.3. Delete duplicate column

A duplicate data check was conducted to identify repeated rows or columns that could affect model performance or bias results. Duplicate values can increase dataset size unnecessarily and distort statistical analysis. The dataset was examined using appropriate data cleaning techniques; however, no duplicate rows or duplicate columns were detected. Therefore, no data removal or transformation was required for this step.

3.8.4. Data Transformation

Data transformation is the process of converting raw data into a structured format suitable for machine learning algorithms. This step improves data consistency, model performance, and

interpretability. In this study, data transformation involved converting categorical variables into numerical representations and standardizing relevant features to ensure compatibility with machine learning models. This step is essential because most machine learning algorithms require numerical input data to perform accurate predictions. The main objective of data transformation is to enhance model efficiency and improve predictive accuracy by ensuring that all variables are in an appropriate and usable format.

3.8.5.Data cleaning

Involves correcting and determining the errors in the data set, such as missing or inconsistent data, eliminates duplicates, handles outliers. It is important to note that you must ensure that you train the machine learning mode on valid and correct data.

- Removal of duplicate records
- The absence of values filed with median(numeric)and mode (categorical) imputation

3.8.6. Feature Engineering /Selection

The development of transforming raw data in to information that is useful to the machine learning models. In other words, feature engineering is the process of creating predictive model features. A feature-also called dimension – is an input variable used to generate model prediction.

Feature selection plays a significant role in improving the performance of the machine learning algorithms in terms of reducing the time to build the learning model and increasing the accuracy in the learning process. Therefore, the researchers pay more attention on the feature selection to enhance the performance of the machine learning algorithms. Identifying the suitable feature selection method is very essential for a given machine learning task with high-dimensional data. Hence we need to study feature selection methods for the research community. This is especially important for developing a method to improve machine learning tasks on high-dimensional data. To achieve this paper reviews feature selection methods for high-dimensional data. The goal is to enhance performance of machine learning tasks on data. Feature selection methods are crucial for this. This literature review covers methods. Dimensional data poses challenges, for machine learning tasks. Therefore a suitable feature selection method is required. The research

community needs to focus on developing this method. Machine learning tasks need to be improved on dimensional data. Feature selection methods can help achieve this goal [49].

The data was scaled on features so as to standardize the range of the input variables and subsequently trained the machine learning models. The message “Features scaled successful states that the numerical features were scaled using a scaling method to ensure that the scale of all the variables is closer to each other. This is a significant step since machine learning algorithms especially models like Logistic Regression work better when the input features are normalized or standardized.

The scaling also assists in avoiding the dominance of variables that have higher numerical values in the learning process thus enhancing the stability and efficiency of the model training process. Here, the feature scaling provides that those aspects like training hours, education level and other numerical variables will be scaled in a way that they play equal roles in predicting the levels of skill gaps.

The standardization of the features allows the model to learn patterns better and also converge more quickly as the optimization process. Consequently, feature scaling leads to the enhanced operation of models as well as more valid prediction outcomes in the skill gap examination.

3.8.7. Train test split

Train–test split is a fundamental technique in machine learning used to evaluate model performance by dividing the dataset into two subsets: a training set and a testing set. In this study, 80% of the dataset is used for training the model, while the remaining 20% is reserved for testing its performance on unseen data. The training set is used to learn model parameters, while the test set is used to evaluate the model’s ability to generalize to new data [50].

The data was separated into the training and testing sets to assess the performance of the machine learning models. The outcome ((3993, 13), (999, 13)) shows that the training set has 3,993 observations (records) that have 13 features, and the testing set has 999 observations (records) that have 13 features. This implies that 80 percent of the data was trained and 20 percent was to be used in testing.

The training data was working to enable the machine learning algorithms to acquire patterns and relationship between the independent variables and the target variable (Skill Gap Level). The more of the data is used in training, the better the model is able to be generalized and to capture significant patterns in the data.

The testing dataset on the other hand was required to assess the predictive performance of the trained model on the unseen data. This assists in establishing the extent to which the model can accurately predict the level of skill gap on new cases and to prevent the problem of over fitting of the model.

On the whole, by dividing the data into training and testing data sets, the reliability and validity of the machine learning findings can be enhanced and a more realistic assessment of the predictive ability of the model is offered.

3.9. Developing Machine Learning Models

This research, used machine learning methods that demonstrate diverse levels of performance across various algorithms, resulting in the development of multiple models. The model applied a variety of machine learning algorithms including XGBoost, Random forest (RF) as well as Logistic Regression (LR).

Three predictive models were developed and evaluated.

3.9.1. XGBoost classifier

I choose as the major model because of

- Process structured learning datasets
- Works with mixed variables (categorical and numerical)
- High predictive accuracy
- Suitable for both regression and classification algorithms
- Ability to handle none leaner data interaction
- Built in regularization to reduce over fitting

3.9.2. Random Forest

A random forest in machine learning is an ensemble method that builds many decision trees during training and outputs the mode (classification) or mean (regression) of the individual trees predictions.

- For interpretability and strength on varied data types
- Making it highly accurate and robust by reducing over fitting

3.9.3. Logistic Regression

Logistic regression is widely used supervised machine learning algorithm for classification problems, where the goal is to predict a categorical outcome rather than a continuous value.

Therefore, XGBoost, Random Forest, and Logistic Regression were selected because they provide a balance between accuracy, interpretability, robustness, and suitability for structured educational data, which aligns with the objectives of this study.

3.10. Data representation and understanding

This defines methods and analytical actions used to understand, summarize, and visualize the data set.

Understanding research results and sharing conclusions demands the presence of data visualization:

- Common visualization tools include Matplotlib, Seaborn, and Streamlit
- Important Visualizations:
- The first visualization displays skill match scores according to different Occupations
- The second visualization represents employment status based on curriculum alignment
- Heat maps demonstrate which skills appear most frequently in job postings

3.11. Hyper parameter Tuning

Hyper parameter tuning is essential for optimizing the performance and generalization of machine learning (ML) models. The hyper parameter tuning in Machine Learning is really important. It plays a role in Machine Learning. We need to understand how important hyper parameter tuning is in Machine Learning. We also need to know about its applications and the different ways we can optimize it. There are some things that affect how well Machine Learning works. These include how good the data is, which algorithm we choose and how complicated the model is. We also need to think about how things like the learning rate and batch size affect the training of the model. When it comes to tuning there are methods we can use. We can try grid search. We can use random search. Some people also like to use Bayesian optimization or meta-learning for hyper parameter tuning, in Machine Learning [51].

3.12. Model Evaluation and Validation

This activity is about checking how well a specific model works. To do this we use a confusion matrix and some key measures. These measures include accuracy, precision, recall, F1-score and ROC-AUC. They help us understand how good the model is, at making predictions. Checking the models performance involves looking at things: How accurate it is, how precise it is, how well it recalls important information, Its F1-Score.We also use cross-validation to see how strong the system is. This helps ensure the model works well in situations.

3.11.1. Confusion Matrix

The Confusion matrix is considered a basic and easy-to-understand metric for assessing the accuracy and correctness of a model. It consists of values that represent both correct and incorrect classifications and is commonly employed in the evaluation of different classification models. This matrix enables a comparison of the number of data points that are accurately versus inaccurately classified, making it particularly useful for classification scenarios involving multiple classes [52].

True Positive (TP):- indicates the number of instances that are correctly predicted as well as the actual outcomes of "low risk".

✓True Negative (TN):- refers to the number of instances where both the predicted outcome and the actual outcome are categorized as “high risk ".

✓False Positive (FP):- indicates the instances where the actual outcome is classified as "high risk ", but the predicted outcome is inaccurately labeled as "low risk".

✓False Negative (FN):- refers to the instances where the actual outcome is categorized as "low risk", but the predicted outcome is mistakenly classified as "high risk ".

3.11.2. Performance Measure

✓Accuracy is the effectiveness of a machine learning classification algorithm can be assessed by its accuracy, which reflects the frequency of correct classifications of data points. Accuracy quantifies the ratio of correctly predicted data points to the total number of data points. This metric is determined by dividing the sum of true positives and true negatives by the sum of true positives, true negatives, false positives, and false negatives.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \qquad (\mathbf{3.2})$$

Precision representing the quality of accurate predictions produced by the model, serves as a metric to gauge its performance. It's computed by dividing the sum of all positive predictions by the ratio of actual positives, encompassing both true positives and false positives.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} \qquad (\mathbf{3.3})$$

A recall referred to as the true positive rate, measures the machine learning model's capability to correctly identify a high proportion of actual positive instances. It highlights the model's effectiveness in capturing positive cases within the dataset. In essence, recall evaluates the model's ability to avoid missing actual positives, minimizing false negatives and ensuring a comprehensive identification of positive instances within the dataset.

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}} \qquad (3.4)$$

The F1-score acts as a comprehensive performance metric that unites recall and precision, computed as the harmonic mean of these two measures. Obtained by averaging recall and precision, the F1 score equally emphasizes both recall and precision. Precision, a constituent of the F1 score, calculates the ratio of correctly [53].

$$\boldsymbol{F1 - Score} = \frac{\boldsymbol{2 * Precision * Recall}}{\boldsymbol{Precision + Recall}} \qquad (\mathbf{3.5})$$

3.13. Regularization in machine learning

Regularization is a technique used in machine learning to prevent over fitting, which otherwise causes models to perform poorly on unseen data by adding a penalty, for complexity regularization helps create models. These simpler models are also more generalizable [54].

3.14. Ethical Considerations

Agreement requested from all the participants. Data Stored secretly and securely ethical clearance request from the institution's research ethics board.

CHAPTER FOUR

4. RESULT AND DISCUSSION

4.1. Introduction

This chapter explains the procedures of Predictive Analytics of Industry Skills Gaps Using Machine Learning – In Case of Tilahun Yigzaw TVET Colleg**e** and discusses their outcomes. This study dataset has been pre-processed and is ready for testing, as declared in the previous chapter. It also describes the events and methods used through the research.

4.2. Exploratory Data analysis (EDA)

The data collected from Tilahun Yigzaw TVET College contains 5000 records with 15 attributes to predict the Skill_GAP_Level. But after applying pre-processing method there are 4992 instances with 13 attributes including one target class.

4.2.1. **Distribution of Skill Gap Levels**

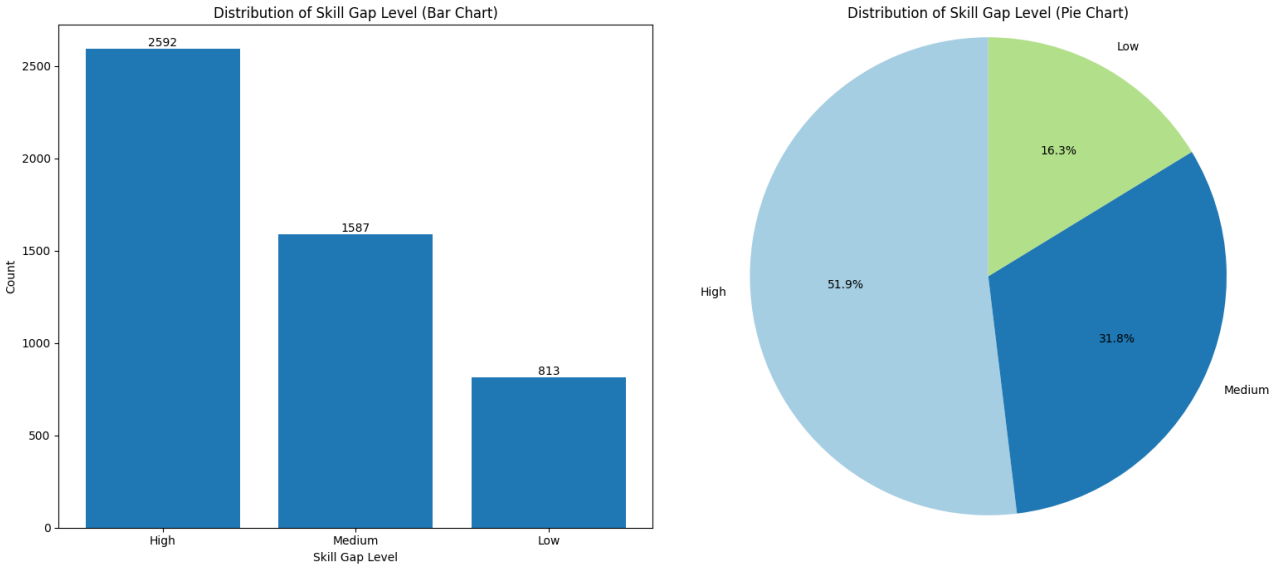


Figure 6 Distribution of Skill Gap Level

As shown in the figure 6 the level of Skill Gap is imbalanced among graduates. High skill gap type overcome in the dataset (2,592/51.9%), then there is Medium (1,587/31.8%) and Low skill gap (813/16.3%). It means that a high percentage of graduates have high gaps existing between

the skills obtained and the industry demands. The amount of high and medium levels of skill gaps assist to classify efficient gaps in training significance and industry arrangement that need predictive analytics and specific involvement to address the gap of the skills and improve employment outcomes.

4.2.2. Distribution of Sex

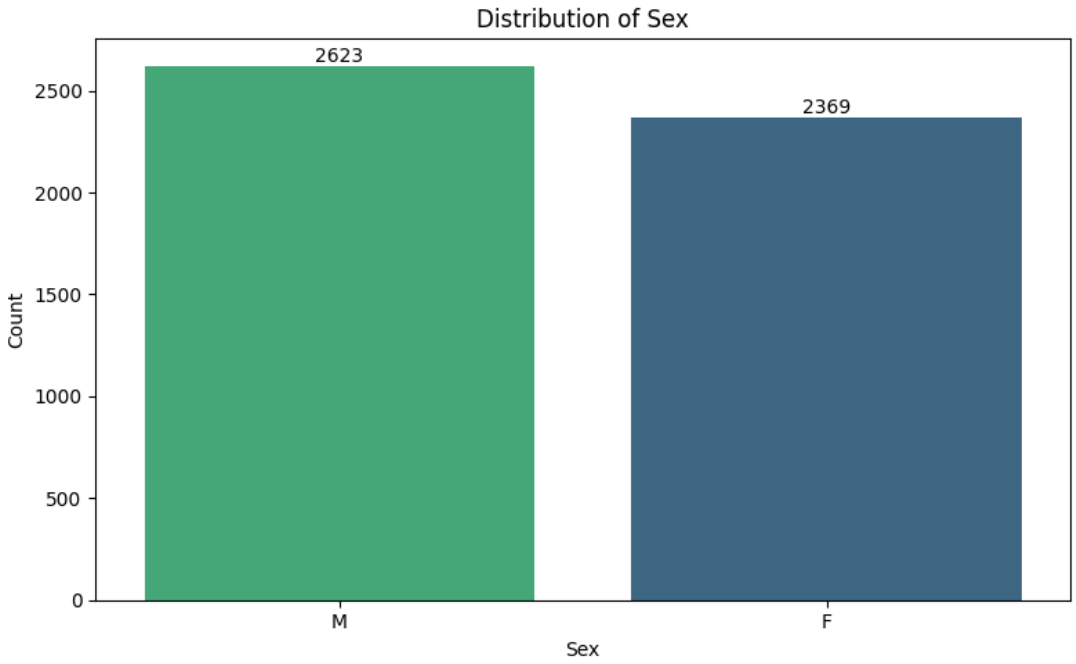


Figure 7 Distribution of Sex

Figure 7 demonstrate the sex of the study population. The total number of the samples is 4,992 which are split into the following parts: Male (M): 2,623 participants. Female (F): 2,369 members.

Demographic study shows that there is a rather equal sex ratio with a slight male majority. The amount of males is about 52.5 percent of the total population and that of females is 47.5 percent. The information that the two numbers are close to each other (254) show that the procedures of employment or join have almost an equal gender ratio. Such similarity is also important to the strength of the study, since there is no serious grade in the outcome which would need a big amount of increase or subgroup analysis.

4.2.3. Distribution of Occupation

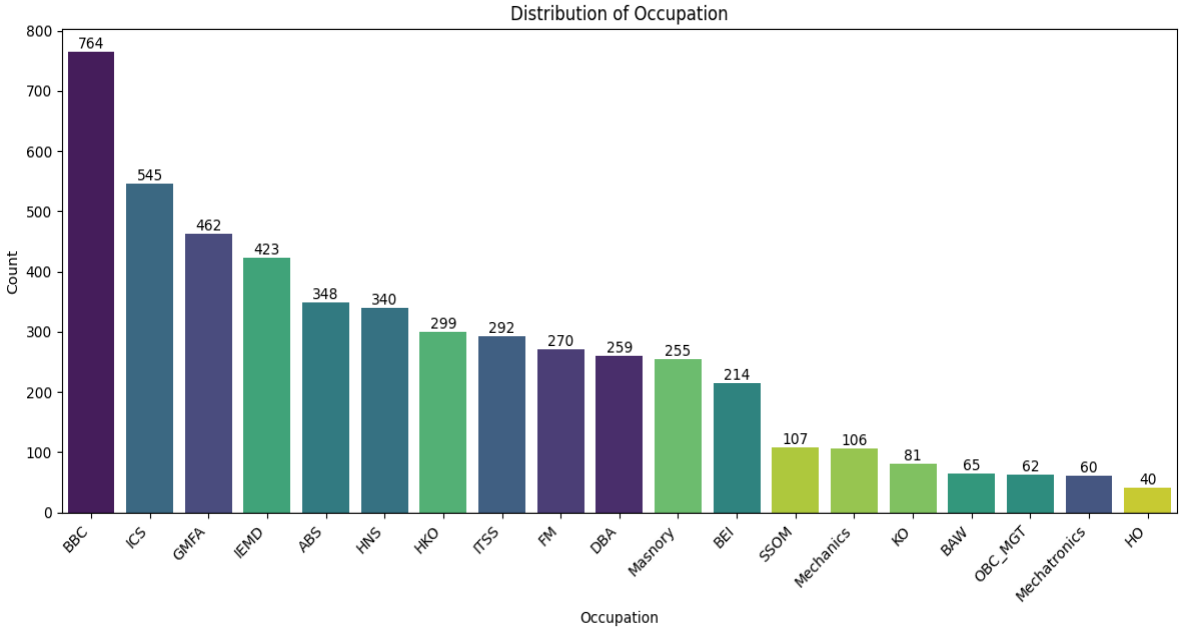


Figure 8 Distribution of Occupation

There is important variation in the occupation distribution as per field of training. The maximum number of graduates is in Building Construction (764) and ICS (545) and GMFA (462), which means that there is a good number of enrolled and created workers in these areas. They are characterize mid-level in jobs like IEMD(423), ABS(348, HNS 340, HKO 299 and TTSS 292, and the number of graduates in technical and specialized areas such as Mechatronics 60, OBCMGT 62, BAW 65 and HO 40, are relatively small.

Such imbalanced distribution implies that there may be a gap in the supply of training and demand of labor in the labor market, and this might be one of the reasons why the level of skill gap may vary among occupations. These outcomes validate occupation-specific prediction of skill gaps and occupation-specific curriculum- Industry alignment.

4.2.4. Distribution of Sector

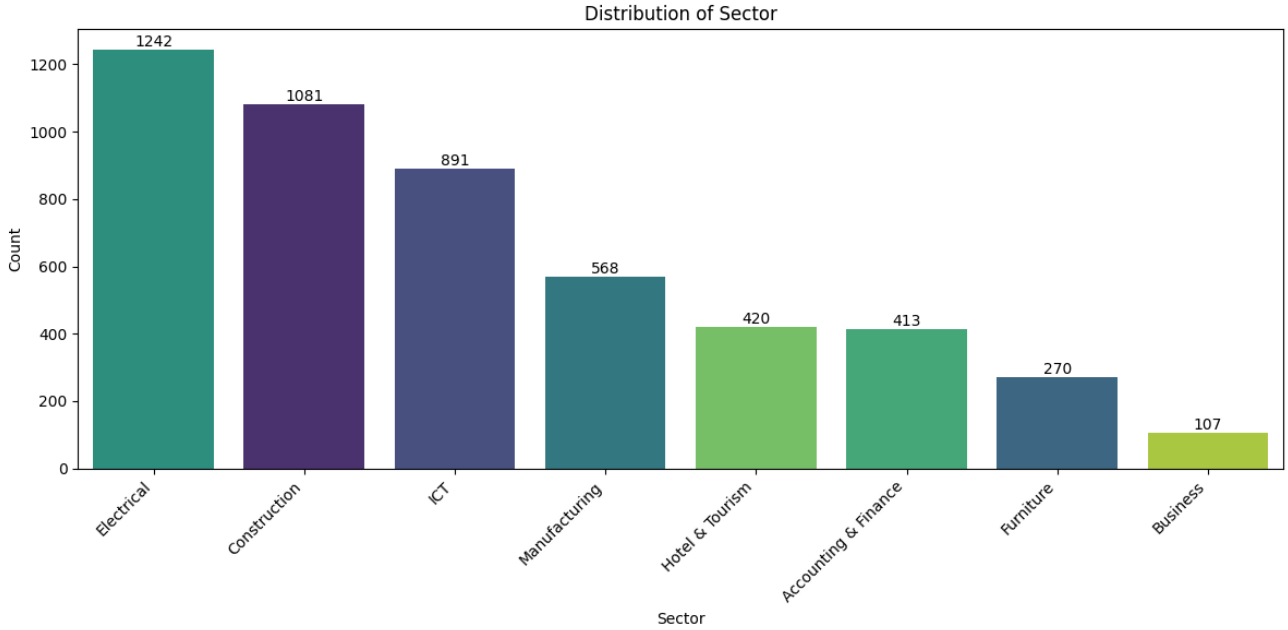


Figure 9 Distribution of Sector

The sector-wise distribution indicates that Electrical sector is the highest with the largest number of graduates (1,242), followed by Construction (1,081) and ICT (891), which all explain the high training force and enrolment in the technical sector. An average share (568) is found in the manufacturing and a more fair but less participation in Hotel and Tourism (420) and Accounting and Finance (413). Furniture (270) and Business (107), on the other hand, are the most underrepresented.

Such unequal sector distribution involve that the area of TVET training is extremely focused on technical, infrastructure-based areas, possibly tracking the priorities of national development. But the underrepresentation in the service-based sectors could be a cause of sector skills shortages and labor market mismatches. These results highlight the importance of industry-specific analysis of the skills gap and the predictive model to facilitate the growth of a balanced workforce and more efficient corresponding training-based supply with the demand in the labor market.

4.2.5. Distribution of Level

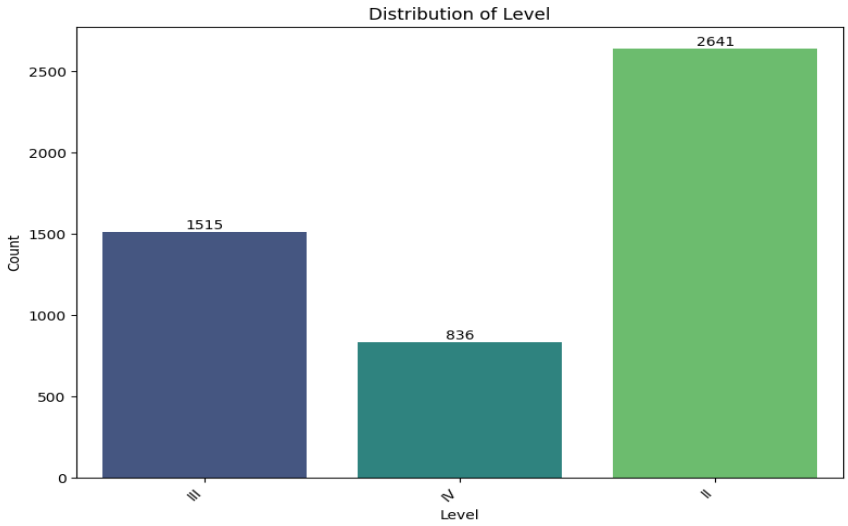


Figure 10 Distribution of Level

The distribution of the member in three academic or professional levels is shown in Figure 10. The analysis shows that the total sample size is 4,992 students and is divided into the following levels: Level II: The largest group consisting of 2,641 students. Level III: The number of students is 1,515. Level IV: This group had the least number of people with 836 students.

The analysis shows the clear negative correlation between the level development and the student's number. Level II forms the largest percentage of the group with some 2,641(52.9%) percent of the total population.

This distribution practice, such that the population decreases as the level grows is that of the hierarchical or pyramid arrangement. The high turnover between Level II and Level IV (a decline of about 68.3%) show that the slow destruction rates are high, there are more supplies to the development, or the area focus becomes focused at the higher levels. In particular, the relatively low percentage in Level IV (16.7%) could indicate a move to the higher level of knowledge or management positions, in which there are lower numbers of applicant that fit the required requirements of progressing to the next level.

4.2.6. Distribution of Program

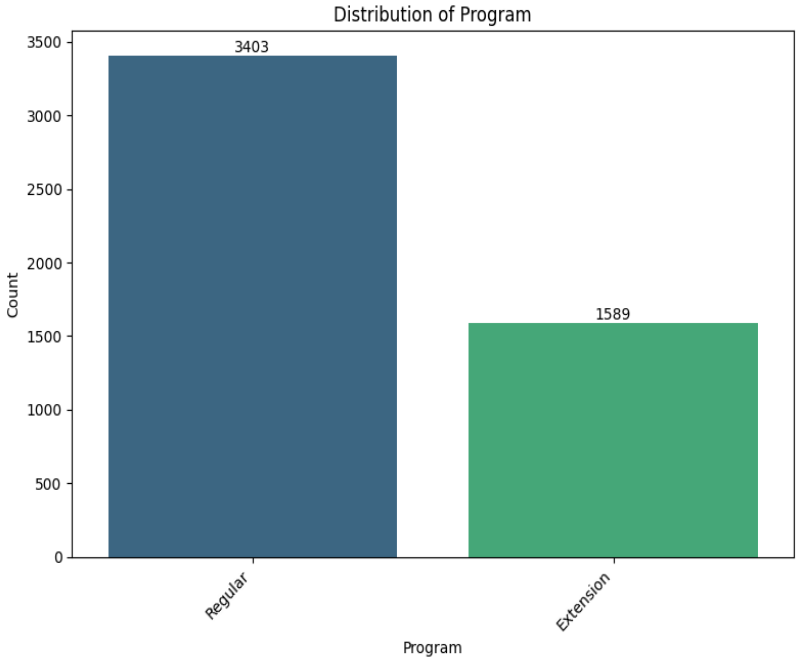


Figure 11 Distribution of Program

Figure 11 indicates the distribution of the students in the two types of programs. The better part of the illustration is the Regular program which has 3,403 participants whereas the Extension program has 1,589 students. The dataset consists of 4,992 students in total.

The figures show that there is a very big difference in both enrollments in the two tracks with the Regular program having about 68.1(3403) percent of the total number of trainers in the group. The overpowering presence of the Regular program involve that it is the main alternative or the straight delivery model with the population. The higher number of Regular enrollment as compared to Extension program could be an attribute of accessibility, institutional capacity, or favorite.

These results can be used to notify the distribution of resources so that it is equal to the increased number of the Regular track when examining possible obstacle to entry to the Extension program.

4.2.7. Distribution of Graduate Year

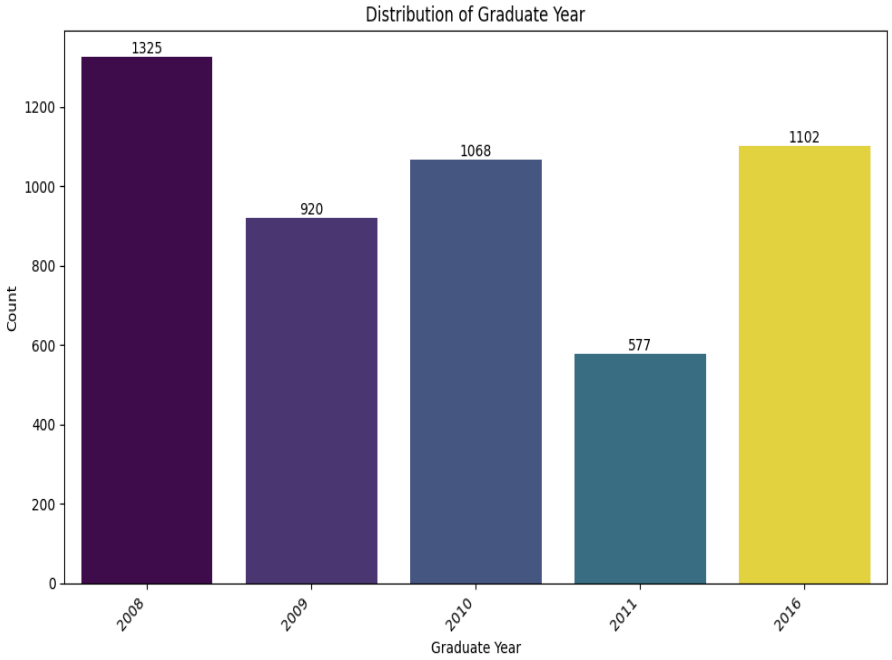


Figure 12 Distribution Graduate Year

The number displays how graduates per year are distributed in the data. The majority amount of graduates evidence was in 2008 (1,325) then, 2016 (1,102) and 2010 (1,068). The moderate number of graduates was in 2009 (920) and the lowest number was in 2011 (577). This dispersion suggests that certain cohorts had a larger sample to the data, and thus, it is better represented in the dataset of graduates and skills gap patterns.

4.2.8. Distribution of Internship

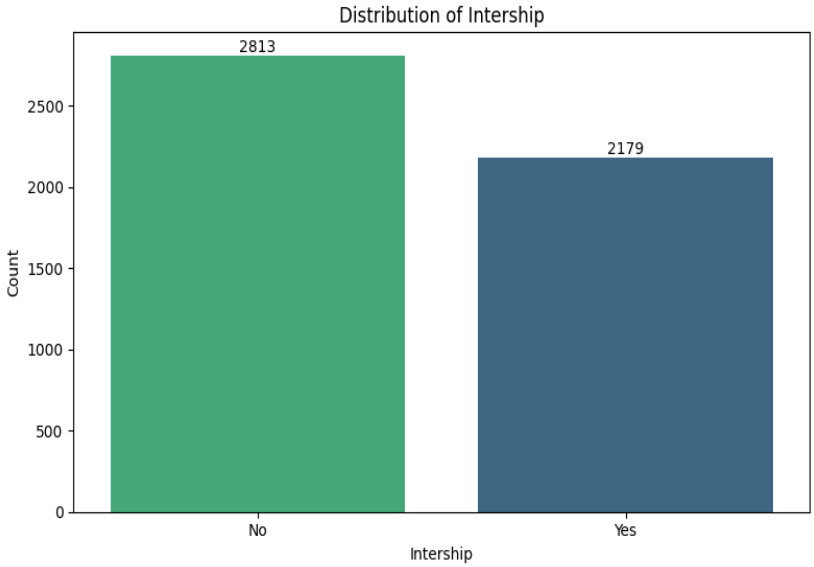


Figure 13 Distribution of Internship

The figure 13 provides the distribution of the number of students who contribute in the internship. Among the total 2,813 students (56.3%) assumed that they had not taken an internship whereas 2,179 students (43.7) said that they had taken an internship. This indicates that a higher percentage of students failed to do internship programs in the course of their study.

The findings specify that there is an important variation in the internship attendance with more than half of the student having no practical industry exposure. As internships are a solution factor in the gap between theory and practice, the low participation rate could be the cause of higher skill gaps and a lack of employment training in graduates.

This difference can be a significant issue in the model results in the case of your skill gap and predictive analytics study, in that students who have not had the internship experience have a higher probability of indicating a higher skill gap. Therefore, internship participation can be measured a significant predictor variable in explaining variations in industry skill gap levels.

The findings show the need to ensure TVET organization have more links to industries and give more internship programs and provide structured workplace learning programs to advance the employability of students and moderate the obvious skills gap.

4.2.9. Distribution of Employment Status

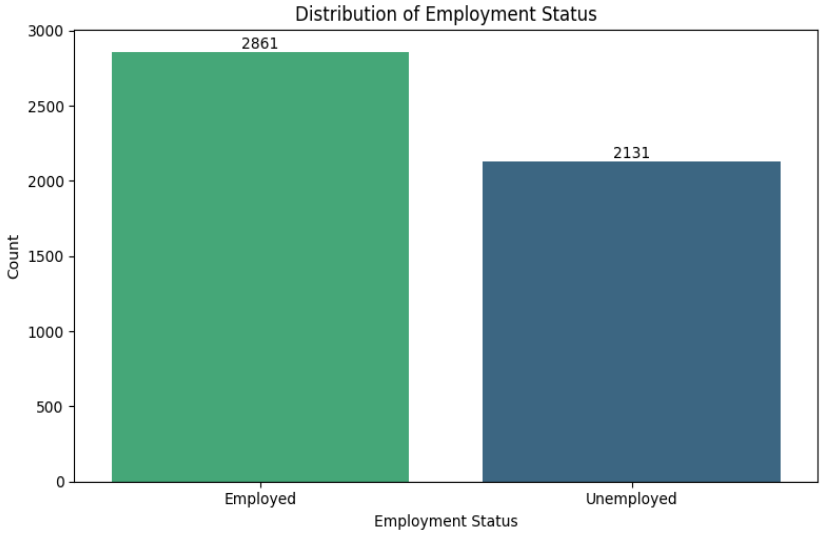


Figure 14 Distribution of Employment Status

The Figure 14 shows the percentage application of the employment status of students. Among the entire number of 2,861 students (57.3%) are employed and 2,131 students (42.7%) are not. This shows that there is a slight increase in the percentage of graduates who are in employment as opposed to the percentage of unemployed.

The result informs that there are moderate graduate employment rates that advise that more than 50 percent of the students get work. However, the rate of employment of 42.7 % (2,131) is comparatively important and involve that there are still some alarm concerning aligning the outcomes of TVET training with the requirements of the working market.

With regard to the skill gap analysis, it is achievable that unemployment is strongly connected with an enhancement in the level of the skill gap. Graduates who do not possess competencies in the industry, have no practical exposure, or have never worked as an intern can be challenged in getting employment. Thus, the employment status is an important outcome variable of training program and predictive models evaluation.

The idea of the results are that although fairly favorable employment outcomes, a high percentage of graduate's still needs specific interference, including updating the curriculum to be industry-driven, reinforcing internship programs, and continuing skill improvement programs to improve employability and moderate the industry skills gap.

4.2.10. Distribution of Skill Gap Type

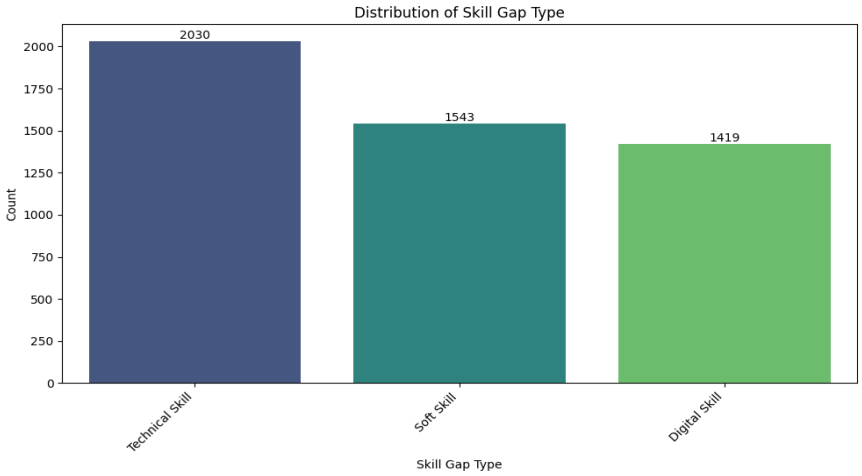


Figure 15 Distribution of Skill Gap Type

The figure 15 illustrates the types of skill gaps as distributed amongst respondents. Technical Skills have the highest level, 2,030 cases (40.7%). The next in line are Soft Skills, 1,543 cases (30.9%), Digital Skills, and 1,419 cases (28.4%).

Such findings show that the most predominant of the three categories is the skill gap in terms of technical skills.

The results indicate that the lack of technical skills is the key problem of graduates. Considering that the main purpose of the TVET institutions is to provide occupation-specific technical training, the fact that the real skills shortage is dominated by technical skills points to the possibility of the mismatch of the content of the curriculum with the current industry demands.

The soft and digital skills gaps are not as high as they are in comparison, but their numbers are high. Soft skills like communication, teamwork, and problem-solving play a vital role in work place performance, whereas digital are becoming fundamental to all fields of work because of the development of technology. These gaps reveal the necessity of combined training methods that incorporate technical skills and transferable and digital ones.

In general, the findings indicate the necessity of the balanced intervention plan, which can strengthen practical technical training; incorporate into the curricula soft skills development and increase digital literacy programs. These three dimensions need to be handled simultaneously in a bid to close the general industry skills gap and enhance the employment rate of graduates.

4.3. Predictive Statistical Analytics of Skill Gap Level Using Cross Tabulation

Cross-tabulation analysis was conducted to examine the distribution of skill gap levels (High, Medium, and Low) across key categorical variables. The results reveal statistically meaningful variations, indicating that skill gaps are not randomly distributed but are influenced by structural, educational, and experiential factors. Descriptive statistics, including frequency distributions and cross-tabulations, were first applied to establish baseline patterns before developing predictive machine learning models. This progression from descriptive to predictive analysis enabled both understanding of current disparities and forecasting of future skill gap risks.

Overall, 51.92**%** of students fall under high skill gap, 31.79% under medium, and only 16.29% under low, indicating a widespread mismatch between training outcomes and labor market requirements.

4.3.1. Skill Gap Level by Occupation

Skill gap levels vary considerably across occupations. Technical fields such as GMFA, ITSS, IEMD, and BAW show extremely high concentrations of skill gaps (approaching or at 100%), indicating substantial training–industry mismatch. In contrast, business-oriented fields such as OBC_MGT and SSOM are concentrated in the low skill gap category, reflecting better alignment with labor market needs.

Some occupations (e.g., ICS, Mechanics, and Mechatronics) are entirely concentrated in the medium category, suggesting moderate competency development that requires targeted improvement rather than complete restructuring.

Overall, occupation is a critical predictor variable and should be prioritized in predictive modeling and curriculum design.

4.3.2. Skill Gap Level by Sector

The analysis shows strong sectorial disparities in skill gap distribution. Technical and production-oriented sectors such as Manufacturing (81.34%), Furniture (84.07%), Construction (65.59%), and ICT (60.38%) exhibit high proportions of students in the high skill gap category.

This suggests a significant misalignment between training content and industry requirements in these sectors.

In contrast, Business (100% low skill gap) and Accounting & Finance (69.25% low) demonstrate better alignment with labor market expectations. Sectors such as Electrical and Hotel & Tourism show moderate performance, with most students falling under the medium skill gap category.

These findings confirm that sector is a strong predictor of skill gap variation, highlighting the need for sector-specific curriculum reform, enhanced practical training, and stronger industry partnerships.

Table 3 Skill Gap Distribution According Sector

Skill_GAP_Level	High	Low	Medium	Total
Sector				
Accounting & Finance	94 (22.76%)	286 (69.25%)	33 (7.99%)	413 (100.00%)
Business	0 (0.00%)	107 (100.00%)	0 (0.00%)	107 (100.00%)
Construction	709 (65.59%)	62 (5.74%)	310 (28.68%)	1081 (100.00%)
Electrical	562 (45.25%)	0 (0.00%)	680 (54.75%)	1242 (100.00%)
Furniture	227 (84.07%)	0 (0.00%)	43 (15.93%)	270 (100.00%)
Hotel & Tourism	0 (0.00%)	184 (43.81%)	236 (56.19%)	420 (100.00%)
ICT	538 (60.38%)	174 (19.53%)	179 (20.09%)	891 (100.00%)
Manufacturing	462 (81.34%)	0 (0.00%)	106 (18.66%)	568 (100.00%)
Total	2592 (51.92%)	813 (16.29%)	1587 (31.79%)	4992 (100.00%)

4.3.3. Skill Gap Level by Level

A clear progression is observed across qualification levels. Level II students are predominantly in the high skill gap category (82.13%), indicating insufficient foundational skill development. Level III students are largely in the medium category (73.99%), reflecting partial competency

acquisition. In contrast, Level IV students are mainly in the low skill gap category (75.24%), demonstrating strong alignment with industry requirements.

This trend confirms that higher qualification levels significantly reduce skill gaps, emphasizing the need to strengthen competency-based training at lower levels, particularly Level II.

4.3.4. Skill Gap Level by Program

Program type shows the most pronounced variation in skill gaps. Students enrolled in the Extension program exhibit an extremely high proportion of skill gaps (87.79% high), with no students in the low category. In contrast, Regular program students display a more balanced distribution, with significantly lower high skill gaps (35.17%) and a meaningful proportion achieving low skill gaps (23.89%).

This indicates that program structure and delivery significantly influence skill acquisition, with Extension programs likely lacking sufficient practical exposure and industry linkage. Strengthening these components is critical to reducing skill gaps.

Table 4 Skill Gap Distribution According Program

Skill_GAP_Level	High	Low	Medium	Total
Program				
Extension	1395 (87.79%)	0 (0.00%)	194 (12.21%)	1589 (100.00%)
Regular	1197 (35.17%)	813 (23.89%)	1393 (40.93%)	3403 (100.00%)
Total	2592 (51.92%)	813 (16.29%)	1587 (31.79%)	4992 (100.00%)

4.3.5. Skill Gap by Internship

Internship participation emerges as the most influential factor affecting skill gap levels. Students without internship experience show extremely high skill gaps (92.14%), whereas those with internship experience have no representation in the high category and are distributed between medium and low levels. This highlights the critical role of practical exposure in reducing competency mismatch.

4.3.6. Skill Gap by Employment Status

Employment status further supports this finding, as unemployed graduates exhibit higher skill gaps compared to employed graduates, indicating a direct relationship between skill alignment and employability. In contrast, sex shows minimal variation, suggesting that skill gaps are systemic rather than gender-based.

4.3.7. Overall Skill Gap Distribution

The overall distribution indicates that more than half of the students fall under the high skill gap category, with a relatively small proportion achieving low skill gaps. This highlight a substantial disconnect between training outcomes and labor market expectations. The findings underscore the urgent need for targeted interventions, including curriculum revision, expanded internship opportunities, and strengthened industry collaboration to improve graduate employability.

4.4. Machine Learning Model Development

The type of information that was used in this study is the evaluation of models that were created in order to predict degree of industry skills gaps among the trainees and graduates of the TilahunYigzaw TVET College. Development of these models was aimed at finding patterns in the dataset and coming up with insights that can be used to make predictions and aid in decision-making to enhance training programs and employment outcomes.

The development of the model has started once data preprocessing and exploratory data analysis were completed. The ready dataset included various input variables that concerned student characteristics, the training background, and the participation in internship, sector, and skill Gap type. The variable being predicted was the Skill Gap Level as a categorical variable with three classes that included High, Medium, and Low

The various machine learning classification algorithms were used to estimate the levels of skill gaps. These were Logistic Regression, Random Forest and Extreme Gradient Boosting (XGBoost). The use of Logistic Regression has been chosen as a baseline model because it is simple and easy to understand. The application of Random Forest was based on its capabilities to deal with nonlinear relationships and over fitting reduction due to the use of an ensemble learner. XGBoost was also used as it is a strong gradient boosting algorithm that does not fail to perform highly on structured data in most cases.

In the model training process, cross-validation methods were used in evaluating the stability and reliability of the models. Cross-validation splits the training data into several folds and performs multiple trials on the model to provide an assurance that the outcomes are not particular to one data split. This will enhance the strength of the predictive models.

Several metrics of classification that were used to evaluate the performance of the developed models include accuracy, precision, recall, and F1-score. These assessment measures give a clear explanation of the ability of the models to forecast the various levels of skill gaps. Accuracy determines the general accuracy of the predictions whereas precision and recall analyze the capability of the model to identify each of the classes correctly. F 1 -score is a balanced score of precision and recall.

The outputs of such machine learning models give valuable information about the causes of skill differences and allow identifying what aspects training programs should be improved. The study has the potential to aid in devising specific measures to reduce the skills gap and improve employment outcomes among TVET graduates by means of predictive analytics.

4.4.1. The Logistic Regression Model

Logistic Regression Model Training

In the process of training the Logistic Regression model, data preprocessing and model convergence warnings were generated. A Future Warning of pandas suggested that the usage of inplace=True of a chained assignment in the imputation of missing values of the Age variable may not work in future versions. Further a convergence warning of scikit-learn indicated that the Logistic Regression model with lbfgs solver had reached maximum repetitions of the process

before it converged. Nevertheless, this model was fitted successfully and was able to be used to make predictions and evaluations.

Initially, the model was trained using a maximum iteration value (max_iter) of 100, which is the default setting in most machine learning libraries. However, the solver failed to converge within this limit, suggesting that the optimization process required more iteration to stabilize the coefficient estimates.

To address this issue, the max_iter parameter was increased to 500, and subsequently to 1000, to allow the model sufficient iterations for convergence. After increasing the iteration limit, the model showed improved convergence behavior, and the warning was no longer observed.

Feature Scaling

The data characteristics were scaled successfully then the machine learning model was trained. To standardize the range of the variables, feature scaling was used and it facilitates the improvement of the convergence and performance of the Logistic Regression model when training.

Logistic Regression Model Performance (Scaled Data)

A retraining of the Logistic Regression model was performed after the feature scaling so that the prediction performance would improve. Model gave an overall accuracy of 90.29 which means that it is capable of classifying the level of skill gaps well.

Logistic Regression Accuracy

The Logistic regression model resulted in overall accuracy being 90.3 which means that the model predicted most of the levels of the skill gap in the data set. The outcome of this study shows that the model is effective in predicting the categories of skill gaps and can be deemed to be effective in the classification of data in this study.

Confusion Matrix Interpretation

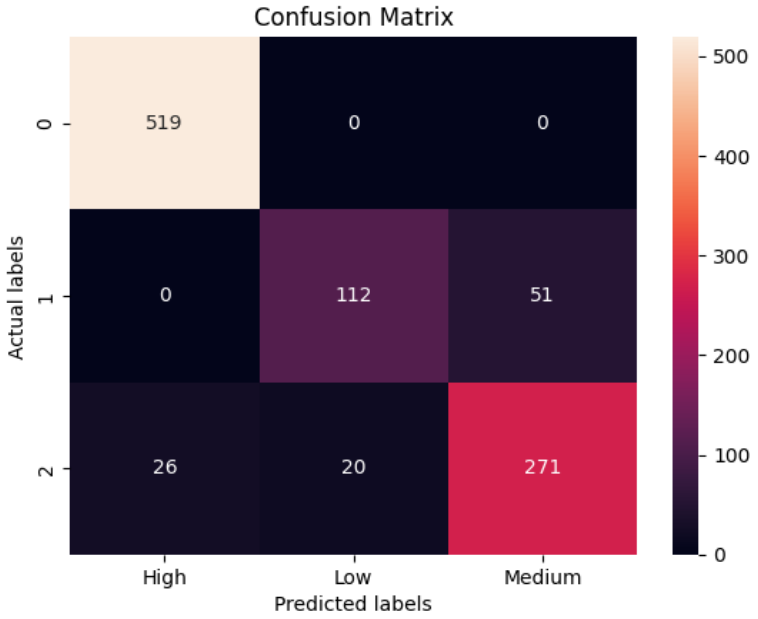


Figure 16 Confusion matrix Logistic Regression

The confusion matrix for the Logistic Regression model demonstrates that the High skill gap group was best predicted with 519 hits and no misclassifications. On the Low skill gap group, 112 were correctly identified, and 51 were incorrectly identified as Medium. The Medium skill gap class had 271 accurate and 26 inaccurate instances of being classified as High and Low respectively. All in all, the findings show that the model is highly effective as regards to identifying the High category and fairly effective as regards to differentiating between Low and Medium levels of skill gaps.

Classification Report Analysis

The overall accuracy of the L1 Regularization applied to the Logistic Regression model was 91 which show a high predictive performance. High skill gap category registered the highest with precision of 0.96, recall of 1.00 and F1-score of 0.98, which implied that almost all cases of High were detected. The Medium category was also showing good performance with F1-score of 0.86 and the Low category was showing a little lower recall (0.70) which implies that some

misclassification occurred. On the whole, the model shows good results in the classification of levels of skill gaps throughout the dataset.

```
••• Classification Report for L1 Regularized Logistic Regression (with text labels):

              precision    recall  f1-score   support

        High       0.96      1.00      0.98       519
         Low       0.85      0.70      0.77       163
      Medium       0.85      0.86      0.86       317

    accuracy                           0.91       999
   macro avg       0.89      0.85      0.87       999
weighted avg       0.91      0.91      0.91       999
```

Figure 17 Classification Report Logistic Regression

Confusion Matrix Interpretation (L1-Logistic Regression)

The L1 Regularization based Logistic Regression model has a good classification as evidenced by the confusion matrix. High skill gap group was well classified with 519 correct predictions and none of the misclassifications. In the case of the Low skill gap category, there were 114 correct predictions and 49 false predictions as Medium. The Medium skill gap category registered 274 hits and 23 false alarms as High and Low respectively. All in all, the model is very accurate, and the number of errors is found between the Low and Medium categories.

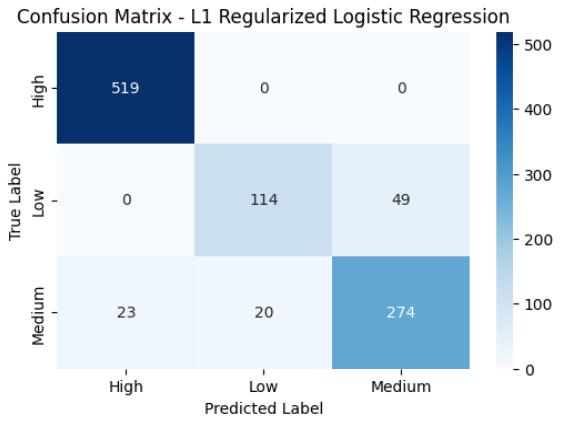


Figure 18 Confusion Matrix L1 Regularized Logistic Regression

Logistic Regression Model Performance

The Logistic Regression model was able to attain a training accuracy of 91.76% and a testing accuracy of 90.29%. The difference in the accuracy of training and testing is small meaning that

the model is applicable to the unknown data and that it does not have an over-fitting effect which illustrates high accuracy when predicting the level of skill gaps.

4.4.2. The Random Forest Model

Random Forest classifier was first trained to come up with a baseline performance of the predictive model. The model had the baseline accuracy of 91.79, which implies that it correctly classified all the cases in the evaluation dataset. Although this finding confirms that the Random Forest algorithm has a high predictive power.

Hyper parameter Tuning Using Grid Search

Hyper parameter tuning of the Random Forest model was performed through GridSearchCV with 5-fold Stratified Cross-Validation to enhance the performance of the model. There were 72 parameter combinations, and 360 models fits (5 folds × 72 combinations). The parameters that have been taken into account in tuning process included criterion (Gini, entropy), max depth (None, 10, 20, 30), min samples split (2, 5, 10), min samples leaf (1, 2, 4) and n estimators (1000). Upon assessing all options, the best model that was chosen by GridSearchCV was the RandomForestClassifier of 1000 estimators. This tuned model offered the most accurate results of the tested setups suggesting that more trees added to the model resulted in the improved predictive accuracy and stability of the model on the skill gap prediction task.

Hyper parameter Optimization of Random Forest Model

In order to maximize the performance of the Random Forest model, the GridSearchCV was used with 5-fold stratified cross-validation. There were 72 different combinations of hyper parameter that were tested and that gave rise to 360 iterations of training the model (5 folds x 72 combinations). These parameters were adjusted, criterion (Gini, entropy), max depth (None, 10, 20, 30), min samples split (2, 5, 10), min samples leaf (1, 2, 4) and n-estimators (1000). Following the evaluation of all the configurations, the selected best estimator was RandomForestClassifier with 1000 trees (n_estimators = 1000) that had the highest level of accuracy and also had the high predictive power concerning the skill gap classification task.

```
•••  Classification Report for Random Forest (after even stronger regularization):
                precision    recall  f1-score   support

          High       0.96      0.99      0.97       519
           Low       0.94      0.71      0.81       163
        Medium       0.85      0.91      0.88       317

      accuracy                           0.92       999
     macro avg       0.91      0.87      0.89       999
  weighted avg       0.92      0.92      0.92       999
```

Figure 19 Classification Report Random Forest

Random Forest Model Performance

The random forest model had an accuracy of 91.79 in the prediction of Skill Gap Level of TVET graduates.

This finding suggests that the model was able to categorize most of the cases in the testing dataset, which is a strong predictor. According to the classification report, the model was the most effective in detecting the High skill gap category with a precision of 0.96, a recall of 0.99, and an F1-score of 0.97 meaning that there was very high precision and recall of high skill gap cases. In Medium skill gap category, the model had precision of 0.85, recall of 0.91 and F1-score of 0.88 which is good predictive ability. The Low skill gap category, however, had a lower recall of 0.71 indicating that there were some incidences of the low skill gaps which were not classified under the other categories.

The macro average F1-score of 0.89 and weighted average F1-score of 0.92 in general shows balanced and trustworthy performance of the models in the various classes of the skill gaps. The findings indicate that the Random Forest algorithm can be applied successfully in forecasting the level of skill gaps and it can be employed to give useful information to enhance the training programs and match the skills of graduates with industry requirements.

Confusion Matrix – Random Forest

The confusion matrix shows how the Random Forest classifier performs in predicting the 3 levels of the skill gap; High, Low and Medium.

The findings demonstrate that the model appropriately categorized 514 of 519 High skill gap cases, and only 5 cases were categorized as Medium meaning that it has very high performance

in identifying high skill gap cases. In the Low skill gap category, the model was accurate with 116 cases and the 47 cases were mistaken as Medium.

This implies that sometimes the model blurs the line between the low and medium levels of skill gap as these two categories are similar. In the Medium skill gap category, 287 were properly classified and 22 were wrongly classified as Higher and 8 cases Low.

Although these slight misclassifications did exist, most of the cases were correctly predicted. On the whole, the confusion matrix shows that the Random Forest model is highly efficient in detecting the High and Medium levels of skill gaps, whereas slightly lower accuracy can be observed in the Low category of skill gaps. This affirms that the model has a high predictive capacity of the classification of skill gaps in the research.

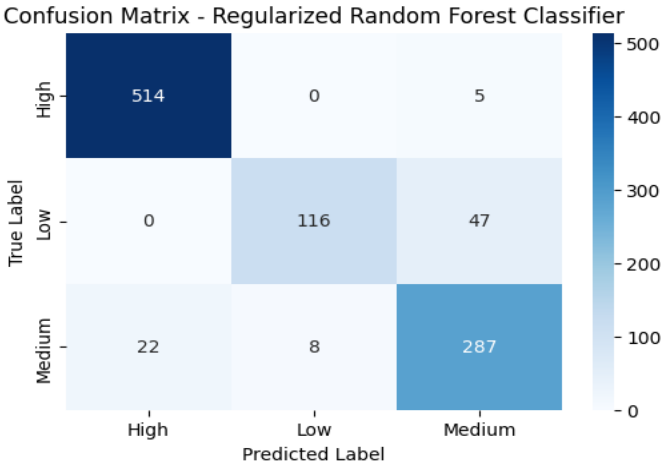


Figure 20 Confusion Matrix Random forest

Random Forest Accuracy

Random Forest model obtained training and the testing accuracy of 91.81 and 91.79 respectively in forecasting the Skill Gap Level. The similarity between the training and testing accuracy shows that the model is good in generalizing to unknown data and it does not experience serious over fitting.

This finding indicates that the Random Forest algorithm can be useful in fitting the associations between the input variables and the target variable. The good and steady accuracy of the model makes it reliable in predicting the level of skill gaps among the TVET graduates.

In general, the Random Forest model has a high predictive success and can be regarded as a solid model to examine and predict skill gaps in the research

4.4.3 The XGBoost

XGBoost Model Configuration

XGBoost classifier was applied to forecast the level of Skill Gap of TVET graduates. This model was fitted with certain hyper parameters, such as learning rate was 0.05, maximum tree depth was 2, number of estimators (trees) was 75, gamma was 0.7 and colsample by tree was 0.6. The choice of these parameters was done to ensure that the complexity of the model is controlled and to ensure that over fitting is minimized as well as good predictive performance is achieved.

The comparatively low learning rate aids the model with learning patterns in a gradual manner, enhancing generalization. The low tree depth makes sure that the model is not too complicated and will not have overly complicated decision boundaries. Moreover, with various trees (n estimation = 75) the model will have a chance to combine numerous weak learners aiming to increase the accuracy of the prediction with boosting.

On the whole, the developed XGBoost model is oriented to effectively modeling dependencies between training characteristics and the level of skill gaps and stabilizing models and avoiding over fitting. This model is incorporated as among the advanced machine learning methods of predicting skill gaps in the study.

XGBoost Model Performance

The XGBoost model had an accuracy of 96.60 which implies that it has a very high predictive performance in the classification of the Skill Gap Levels of the TVET graduates. This fact indicates that high accuracy shows the model is good at capturing the relationships between the input features and the target variable.

Based on the classification report, the model had the highest performance in the High skill gap category, having a preciseness of 0.98, recall 1.00 and a F1-score of 0.99, and it identified 518 out of the 519 cases correctly. In the Low skill gap category the model performed with precision of 1.00, recall of 0.87 and F1-score of 0.93 meaning that majority of the low skill gap cases were also predicted correctly although some of them were confused with Medium. In the Medium skill

gap category, the model had high scores with a precision of 0.93, recall of 0.96 and F1-score of 0.95.

```
...  Accuracy: 0.965965965965966

     Classification Report:
                   precision    recall  f1-score   support

             High       0.98      1.00      0.99       519
              Low       1.00      0.87      0.93       163
           Medium       0.93      0.96      0.95       317

         accuracy                           0.97       999
        macro avg       0.97      0.94      0.96       999
     weighted avg       0.97      0.97      0.97       999
```

Figure 21 Classification Report XGBoost

The confusion matrix also proves the great predictive ability of the model. A majority of the cases were rightly categorized, and few incorrectly categorized among Bed medium and Low categories. In general, the macro average F1-score of 0.96 and weighted average F1-score of 0.97 represent equal and trustworthy model performance of all the classes. These findings imply that XGBoost model is the most predictive of all the models used in the experiment and thus can be used well in predicting the level of skill gaps in this experiment.

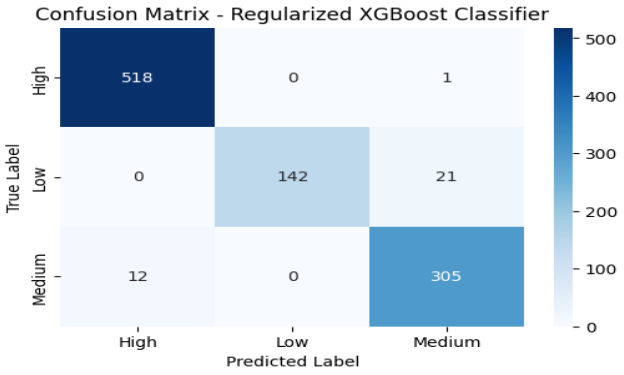


Figure 22 Confusion Matrix XGBoost Classifier

XGBoost Model Accuracy

XGBoost model had accuracy during training of 96.92 and testing accuracy of 96.60. The slight variation in the accuracy of the training and the testing states that the model is generalizable and

exhibits a low-over fit. This finding shows that the model is very useful in forecasting the skills gap levels of the TVET graduates.

4.5. Comparison of Machine Learning Model Performance

The performance of three machine learning models, which include Logistic Regression, Random Forest (Regularized), and XGBoost (Regularized), to predict the level of skill gap (High, Medium, and Low) is provided in Table 8. The findings indicate that the XGBoost model was the most successful and had an accuracy of 96.60% an overall performance more than the other models. It had also the highest macro average precision (0.9700), macro average recall (0.9438), and macro average F1-score (0.9553). XGBoost was also effective in class-level evaluation, where it performed excellently in predicting the High skill gap category with the precision of 0.9774, recall of 0.9981, and F1-score of 0.9876. Likewise, it recorded good scores on the Medium skill gap category (F1-score of 0.9472) and the Low skill gap category (F1-score of 0.9311). The findings suggest that the XGBoost model is the most balanced and trustworthy prediction in all classes.

Random Forest model has moderate results with an average score of 91.79%. Although it was good in forecasting the High skill gap level (F1-score = 0.9744), its forecast qualities were comparatively low in the Low skill gap sub-set (F1-score = 0.8084). The overall macro average of F1-score of 0.8859 shows that, in spite of the fact that the Random Forest works better than the Logistic Regression, it is not as effective as XGBoost in the relationship complexities of the data.

Logistic Regression model performed the worst among the three models having accuracy of 90.29% and a macro average F1-score of 0.8610. It had a good prediction of the High skill gap level (F1-score = 0.9756) but performed worse (F1-score = 0.7593) on the Low skill gap category which has shown that it is not as effective in dealing with more complex/non-linear patterns in the data.

In general, the results show that XGBoost can be the most suitable model to predict the level of skill gap in the study. Its high performance implies that the ensemble boosting techniques will be more appropriate in measuring different relationships between training, internship and

educational variables that we employed in this study. Thus, XGBoost model was chosen as the most successful model of predictive analysis in this paper

Table 5 Model Performance

Model	Accuracy macro	Precision macro	Recall macro	F1 Score macro	Precision weighted	Recall weighted	F1 Score weighted
Logistic Regression	90.29	0.8808	84.73	0.8610	0.8952	90.29	0.8980
Random Forest	91.79	0.9137	0.8691	0.8859	0.9250	0.9179	0.9195
XGBoost	96.60	0.9700	0.9438	0.9553	0.9650	0.9660	0.9655

CHAPTER FIVE

5. CONCLUSION AND RECOMMENDATION

5.1 Conclusion

The study was aimed at studying and forecasting the skills gap in the industry amongst the graduates of Tilahun Yigzaw TVET College through machine learning. This paper has shown that an apparent difference is there between the competencies acquired during training and the skills required in the labor market. Using exploratory analysis of data, cross-tabulation and predictive modeling, a number of the factors that significantly contributed to the occurrence of skill gaps were revealed, such as training hours, participation in internship programs, and sector of work, type of program and level of education. The machine learning models developed, especially Random Forest and XGBoost had good predictive power in determining the extent of skill shortages among graduates. These models had the capability to learn patterns based on historical training and employment data and make dependable predictions that can be used to facilitate institutional planning. The findings also suggest that data-driven solutions can play an important role in helping TVET institutions to know the requirements of the labor market and enhance the alignment of the training programs. On the whole, the research confirms the fact that predictive analytics may be an effective decision-support instrument of educational institutions, policymakers, and industry stakeholders. With the help of machine learning-based models of skills gap prediction, TVET colleges are able to enhance their curriculum design, better their collaboration with industries, and increase graduate employability outcomes.

5.2. Recommendations

1. Meeting the industry needs through curriculum alignment. In the reduction of the gaps between training and employment requirements, the TVET institutions must regularly revise the curriculum in accordance with analysis of the demand of skills in the industry.

2. Enhancing Industry-TVET Partnership. Partnerships, dual training and advisory boards should also be enhanced between TVET colleges and industries so as to guarantee that training is based on actual workplace demands.

3. Growth of Internship and Practical Training. The number of internship opportunities and hours in industry based training can be increased and this may reduce the skill gaps significantly as it enhances practical abilities of students.

4. Implementation of Data-Driven Decision Systems. Predictive analytics and machine learning systems should be integrated into the institutional planning process at the TVET institutions to track the outcomes of the graduates and predict the future needs of skills.

5. On-Going Skills Training Programs. Graduates need to be equipped by institutions with up skilling and re skilling in order to suit the demands of the labor market which is changing very fast.

5.3. Future Work

1. Diverse TVET Institutions Inclusion.

The research can also be extended in the future with more TVET colleges in order to enhance the generalizability of the predictive models.

2. Include additional variables

Such as socioeconomic background, learning environment, trainer competence, and technological infrastructure could be included to enhance prediction accuracy and provide deeper insights in to skill development factors.

3. Real-Time Data on the Labor Market.

The data on the real time situation in the job market including online job adverts and skill requirements could contribute to better prediction.

4. Advanced Machine Learning techniques have been used.

Future research should explore advanced methods such as deep learning, neural networks, and hybrid ensemble models to improve predictive performance and usability. While logistic regression and tree-based models work well for structured data, they struggle with complex nonlinear relationships. Deep learning and ensemble approaches can better capture hidden patterns in high-dimensional data and improve accuracy, robustness, and scalability in skill gap prediction.

5. Creation of Decision Support Dashboard.

A skill gap monitoring dashboard should be created in real time to assist policymakers and educational administrators in making their decisions at the appropriate time.

6. Longitudinal Skill Gap Analysis.

Research may be further done to carry out longitudinal studies to examine how the skill gaps change with time as there is a change in technology and industry demands.

Reference

[1] MoSHE, “Ethiopian TVET Policy and Strategy,” 2020.

[2] C. Bayudan-dacuycuy and L. B. Dacuycuy, “Labor M arket S tructures , P ay G ap , and S kills in the Philippines Labor M arket S tructures , P ay G ap , and S kills in the Philippines,” 2021.

[3] ILO & OECD, “Global skills gaps measurement and monitoring: Towards a collaborative framework,” *Ilo Oecd*, no. January, pp. 1–39, 2023, [Online]. Available: https://www.ilo.org/wcmsp5/groups/public/---dgreports/---ddg_p/documents/publication/wcms_867533.pdf

[4] M. Ali, D. Mardapi, and T. Koehler, “Identification Key Factor in Link and Match Between Technical and Vocational Education and Training with Industry Needs in Indonesia,” no. July, 2020, doi: 10.2991/assehr.k.200521.053.

[5] B. Singh and B. Tolessa, “TVET- Industry Linkage and Collaboration in Ethiopia : A Necessity for Improving Employability Skill,” pp. 3526–3532, 2019.

[6] G. Braun, P. Rikala, M. Järvinen, R. Hämäläinen, and J. Stahre, “Bridging Skill Gaps - A Systematic Literature Review of Strategies for Industry,” *Adv. Transdiscipl. Eng.*, vol. 52, pp. 687–696, 2024, doi: 10.3233/ATDE240209.

[7] R. Daniels, *Skills Shortages in South Africa: A Literature Review*, no. June 2007. 2012. doi: 10.2139/ssrn.992111.

[8] J. G. Whittaker, “Skills Gap – A Strategy for Increasing Knowledge Worker Supply & Demand,” *J. Bus.*, vol. 1, no. 4, p. 13, 2016, doi: 10.18533/job.v1i4.42.

[9] R. S. Kaki, R. C. Gbedomon, F. S. Thoto, D. M. Houessou, K. Gandji, and A. K. N. Aoudji, “Skills mismatch in the agricultural labour market in Benin: vertical and horizontal mismatch,” *Int. J. Lifelong Educ.*, vol. 41, no. 3, pp. 343–365, 2022, doi: 10.1080/02601370.2022.2075480.

[10] S. A. Donovan, A. Stoll, D. H. Bradley, and B. Collins, “Skills Gaps : A Review of Underlying Concepts and Evidence Skills Gaps : A Review of Underlying Concepts and

Evidence," 2022.

[11] P. C. Patel, "Occupational skill mismatch in self-employment: prevalence and income implications," *Int. J. Manpow.*, vol. 46, no. 10, pp. 148–168, 2025, doi: 10.1108/IJM-07-2024-0493.

[12] S. Khattri, P. Partner, and H. S. Khanal, "Bridging Education-Employment Gap Through Curriculum Innovation in Higher Education," no. September 2023, 2024, doi: 10.13140/RG.2.2.25479.00162.

[13] P. K. Tee, L. C. Wong, M. Dada, B. L. Song, and C. P. Ng, "Demand for digital skills, skill gaps and graduate employability: Evidence from employers in Malaysia," *F1000Research*, vol. 13, no. June, 2024, doi: 10.12688/f1000research.148514.1.

[14] P. H. Cappelli, "Skill gaps, skill shortages, and skill mismatches: Evidence and arguments for the United States," *Ind. Labor Relations Rev.*, vol. 68, no. 2, pp. 251–290, 2015, doi: 10.1177/0019793914564961.

[15] M. Wiseman, "The skills gap is real. Here's what to do about it," pp. 1–8, 2022.

[16] A. Binaben and SSACI, *Identification of Skills Gaps in South Africa N S F National Skills Fund FUNDING TO SKILL OUR NATION*. 2024. [Online]. Available: www.dhet.gov.za

[17] N. Azih, "Nigerian Journal of Business Education (NIGJBED) Volume 6 No.1 October 2019," vol. 6, no. 1, pp. 107–116, 2019.

[18] S. Melesse, A. Haley, and G. B. Wärvik, "Bridging the skills gap in TVET: a study on private-public development partnership in Ethiopia," *Int. J. Train. Res.*, vol. 21, no. 3, pp. 171–186, 2023, doi: 10.1080/14480220.2022.2159854.

[19] A. A. Wilson and A. Wilson, "Businesses Walden University This is to certify that the doctoral study by," 2022.

[20] A. Grech and A. F. Camilleri, *The digitization of TVET and skills systems*. 2020. [Online]. Available: https://www.um.edu.mt/library/oar/handle/123456789/108201

[21] M. Busari, "Predictive Analytics for Identifying Skill Gaps and Future Workforce

Requirements," no. March, 2025.

[22] Patrick Azuka Okeleke, Daniel Ajiga, Samuel Olaoluwa Folorunsho, and Chinedu Ezeigweneme, "Predictive analytics for market trends using AI: A study in consumer behavior," *Int. J. Eng. Res. Updat.*, vol. 7, no. 1, pp. 036–049, 2024, doi: 10.53430/ijeru.2024.7.1.0032.

[23] D. Theng and M. Theng, "Machine Learning Algorithms for Predictive Analytics : A Review and New I . INTRODUCTION II . RELATED WORK , METHODS AND DATASET," vol. 26, no. 6, pp. 537–545, 2020.

[24] S. Sah, "Machine Learning : A Review of Learning Types," no. July, 2020, doi: 10.20944/preprints202007.0230.v1.

[25] T. M. Mitchell, *Machine Learning*.

[26] C. Preparedness, "education sciences Learning Analytics for Bridging the Skills Gap : A Data-Driven Study of Undergraduate Aspirations and Skills Awareness for Career Preparedness," no. 2018, 2025.

[27] B. B. Alkan, S. Kuzucuk, N. Alkan, and A. Sinan, "Using machine learning to predict student outcomes for early intervention and formative assessment," pp. 1–18, 2025.

[28] M. Chukwube, "Using Data-Driven Techniques to Close Skills Gaps," 2024.

[29] H. Hormozi, E. Hormozi, and H. R. Nohooji, "The Classification of the Applicable Machine Learning Methods in Robot Manipulators," *Int. J. Mach. Learn. Comput.*, no. July, pp. 560–563, 2012, doi: 10.7763/ijmlc.2012.v2.189.

[30] D. O. Otewa and J. Samaniego, "THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE Machine-Learning Model for Program Selection in Technical and Vocational Educational Training (TVET) in Kenya," vol. 12, no. 8, pp. 8–14, 2024.

[31] Y. Matsuo *et al.*, "Machine Learning: A Review of Learning Types," *Neural Networks 152*, vol. 7, no. 1, pp. 267–275, 2020, doi: 10.20944/preprints202007.0230.v1.

[32] T. Oladipupo, "Types of Machine Learning Algorithms," *New Adv. Mach. Learn.*, no.

February 2010, 2010, doi: 10.5772/9385.

[33] S. S. Dash, S. K. Nayak, and D. Mishra, “A review on machine learning algorithms,” *Smart Innov. Syst. Technol.*, vol. 153, no. January 2021, pp. 495–507, 2021, doi: 10.1007/978-981-15-6202-0_51.

[34] M. Sudha, R. Menon, and D. Sethi, “The Role of HR in Workforce Planning,” vol. 28, no. 5, pp. 395–404, 2025, doi: 10.69980/ajpr.v28i5.395.

[35] S. Jahnavi, P. Naga, D. Prasad, and S. Tripathy, “International Journal of Research Publication and Reviews HR Analytics for Skill Gap Analysis and Training Needs in MSMEs of Andhra Pradesh,” vol. 6, no. 2, pp. 2258–2267, 2025.

[36] J. Market, “CCT College Dublin ARC (Academic Research Collection) Identifying Skills Gaps and Labour Market Trends Using Machine Learning : A Data Analysis of Tendencies Within the Irish Online Identifying Skills Gaps and Labour Market Trends Using Machine Learning : A Data Analysis of Tendencies Within the Irish Online Job Market A Thesis Submitted in Partial Fulfilment Degree of Master of Science in Data Analytics February 2025 Supervisor : Taufique Ahmed,” 2025.

[37] N. Dawson, M. A. Rizoiu, B. Johnston, and M. A. Williams, “Predicting Skill Shortages in Labor Markets: A Machine Learning Approach,” *Proc. - 2020 IEEE Int. Conf. Big Data, Big Data 2020*, no. September, pp. 3052–3061, 2020, doi: 10.1109/BigData50022.2020.9377773.

[38] E. Scornet, “A Random Forest Guided Tour,” pp. 1–42.

[39] E. C. Zabor, C. A. Reddy, R. D. Tendulkar, and S. Patil, “Logistic Regression in Clinical Studies,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 112, no. 2, pp. 271–277, 2022, doi: 10.1016/j.ijrobp.2021.08.007.

[40] K. J. Olowe, N. L. Edoh, S. Jean, C. Zouo, J. Olamijuwon, and I. Researcher, “Comprehensive review of logistic regression techniques in predicting health outcomes and trends,” 2024.

[41] H. A. Salman, A. Kalakech, and A. Steiti, “Random Forest Algorithm Overview,” vol.

2024, pp. 69–79, 2024.

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

[43] Z. A. Ali, Z. H. Abduljabbar, H. A. Tahir, A. B. Sallow, and S. M. Almufti, “Exploring the Power of eXtreme Gradient Boosting Algorithm in Machine Learning : a Review,” vol. 12, no. 2, 2023.

[44] B. Kumar and T. Kumar, “Available online www.jsaer.com Comparative Analysis of ML based Gradient Algorithms : XGBoost , CatBoost , and LightGBM,” vol. 7, no. 8, pp. 235–239, 2020.

[45] E. Journal, S. Kalyana, and P. Buddiga, “The Importance of Data Cleaning in Machine Learning : Best Practices and Techniques,” vol. 7, no. 10, pp. 70–73, 2020.

[46] O. Alotaibi and E. Pardede, “Cleaning Big Data Streams : A Systematic Literature Review,” pp. 1–24, 2023.

[47] A. Brijith, “Data Preprocessing for Machine Learning,” vol. 03, 2023.

[48] S. G. K. Patro and K. Kumar, “Normalization : A Preprocessing Stage”.

[49] D. A. A. Gnana, “Literature Review on Feature Selection Methods for High-Dimensional Data,” vol. 136, no. 1, pp. 9–17, 2016.

[50] D. E. Birba, “study of data splitting algorithms for machine learning,” 2020.

[51] J. A. Ilemobayo, O. Durodola, A. Ogungbire, and A. Osinuga, “Hyperparameter Tuning in Machine Learning : A Comprehensive Review,” vol. 26, no. 6, pp. 388–395, 2024.

[52] S. Sathyanarayanan and B. R. Tantri, “Confusion Matrix-Based Performance Evaluation Metrics,” vol. 27, no. 4, 2024.

[53] P. Machine and L. Algorithms, “Assessing the Effectiveness of Predictive Machine Learning Algorithms Based on Classification,” vol. 12, no. 4, pp. 13882–13895.

[54] A. Y. Ng, “Feature selection , L 1 vs . L 2 regularization , and rotational invariance,” 2004.

Appendix

Appendix 1 approval letter for data Access for Postgraduate Research

	Institution Name: የኢፌዴሪ ቴ/ሙ/ስ ኢንስቲትዩት FDRE TVT INSTITUTE	Document No. OF/FTI/ALL/01	
Effective date 01/09/2016	Title: ወጭ ደብዳቤ Outgoing Letter	Issue No. 2	Page No. Page 1 of 1

Ref. RD/08-100/18
Date: 20/09/2018 EC

To:
Tilahun Yigzaw TVET College
Maichew

Subject: Request for Collaboration and Data Access for Postgraduate Research

We are pleased to request your kind collaboration in supporting the postgraduate research of **Mr. Azmach Berhe** (ID No. TTMR/050/2016), a postgraduate student at the FDRE Technical and Vocational Training Institute.

Mr. Azmach Berhe is currently conducting his postgraduate thesis entitled **"Predictive Analytics and Bridging Industry Skill Gaps to Enhance Training and Employment Outcomes Leveraging Machine Learning: The Case of Tilahun Yigzaw TVET College, Tigray."** The study aims to apply machine learning techniques to analyze graduate outcomes and identify skill gaps, with the objective of improving training relevance and employability. For the successful implementation of this research, access to **datasets related to graduate trainees** (such as training background, employment status, and competency records) is essential. We therefore respectfully request your support in granting him access to relevant datasets strictly for **academic and research purposes**.

We assure you that all data will be used ethically, confidentiality will be maintained, and any conditions set by your institution will be fully respected. Proper acknowledgment of your support will be made in all research outputs.

We kindly request your positive consideration and look forward to your cooperation.

Yours sincerely,

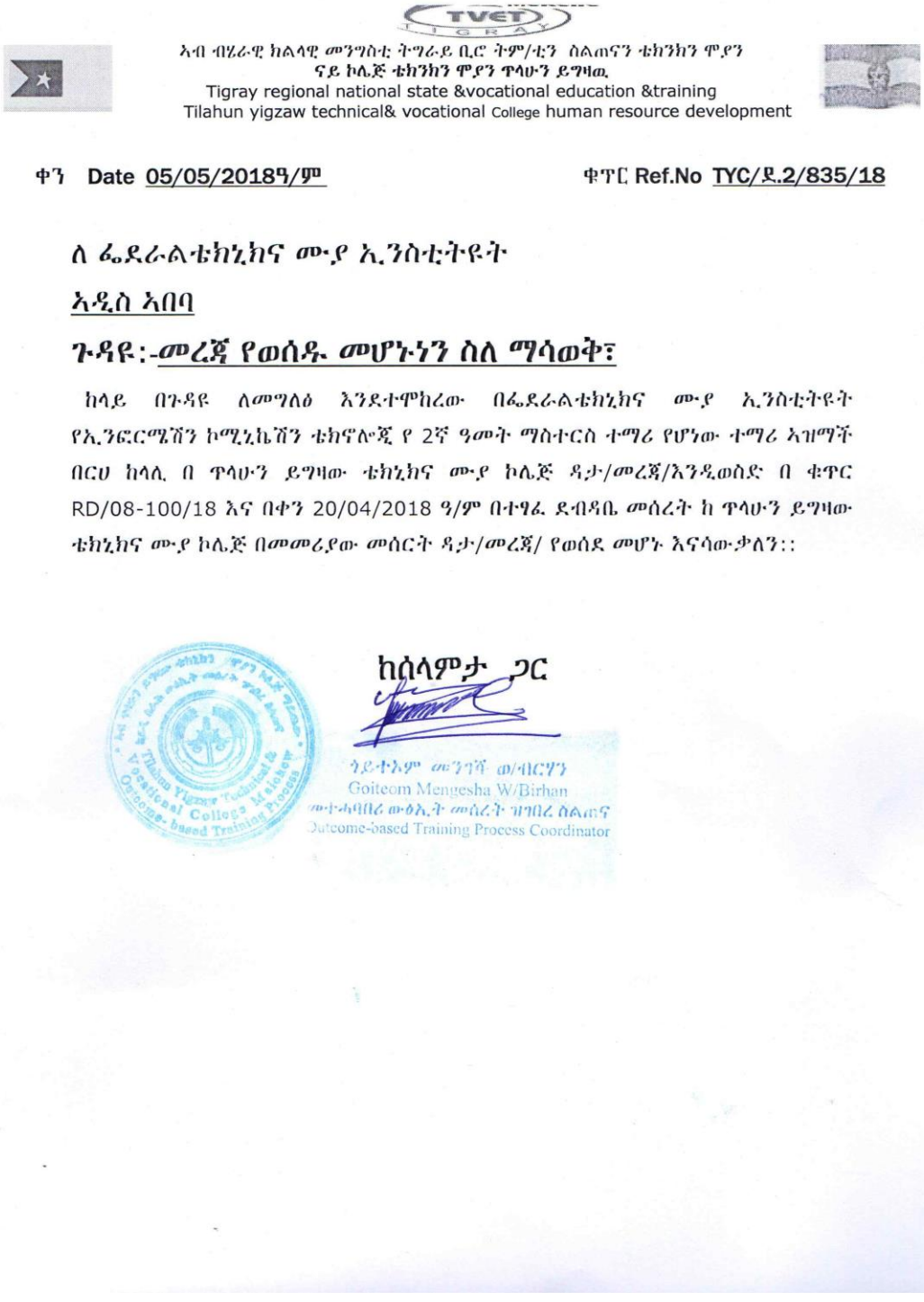
Animaw Tadesse (PhD)

Director for Research and Dissemination

Tel: 0116464455 Email: info@ftveti.edu.et PO. Box: 190310 Website: www.ftveti.edu.et

Year of Excellence

Appendix 2 approval letter for data collection fromTilahunYigzaw TVET College



ኣብ ብሄራዊ ክልላዊ መንግስቲ ትግራይ ቢሮ ትም/ቲን ስልጠናን ቴክንክን ሞያን
ናይ ኮሌጅ ቴክንክን ሞያን ጥላሁን ይግዛዉ
Tigray regional national state &vocational education &training
Tilahun yigzaw technical& vocational College human resource development

ቀን Date 05/05/2018ዓ/ም ቁጥር Ref.No TYC/ደ.2/835/18

ለ ፌደራል ቴክኒክና ሙያ ኢንስቲትዩት

ኣዲስ ኣበባ

ጉዳዩ:-መረጃ የወሰዱ መሆኑንን ስለ ማሳወቅ፣

ከላይ በጉዳዩ ለመግለፅ እንደተሞከረው በፌደራል ቴክኒክና ሙያ ኢንስቲትዩት የኢንፎርሜሽን ኮሚኒኬሽን ቴክኖሎጂ የ 2ኛ ዓመት ማስተርስ ተማሪ የሆነው ተማሪ ኣዝማች በርሀ ከላሊ በ ጥላሁን ይግዛው ቴክኒክና ሙያ ኮሌጅ ዳታ/መረጃ/እንዲወስድ በ ቁጥር RD/08-100/18 እና በቀን 20/04/2018 ዓ/ም በተፃፈ ደብዳቤ መሰረት ከ ጥላሁን ይግዛው ቴክኒክና ሙያ ኮሌጅ በመመሪያው መሰርት ዳታ/መረጃ/ የወሰደ መሆኑ እናሳውቃለን::

ከሰላምታ ጋር

ጎይቶኣም መንገሻ ወ/ብርሃን
Goiteom Mengesha W/Birhan
ውጥሓባዘ ውዕኢት መሰረት ዝገበረ ስልጠና
Outcome-based Training Process Coordinator

Appendix 3 Overview of the TVET Skills Gap Dataset

```
import pandas as pd
df=pd.read_csv ('/content/drive/My Drive/PredictiveAnalytics/Industry_skill_gap_5000_records.csv')
print(df.shape)
df.head()
```

```
(5000, 16)
```

	Stu_ID	Name	Sex	Age	Occupation	Sector	Level	Program	Graduate Year	In_TVET_Training_hrs	In_Company_Training_hrs	Total_Traning_hrs	Intership	Employment_Status	Skill_GAP_Type	Skill_GAP_Level
0	STU_00001	Meselu Teame Menasbo	F	17.0	BBC	Construction	III	Regular	2011	670	52	710	Yes	Unemployed	Technical Skill	Medium
1	STU_00002	Lamrot g/medhin w/kidan	F	22.0	DBA	ICT	IV	Regular	2016	772	85	861	Yes	Employed	Soft Skill	Low
2	STU_00003	Abadi Gebremeskel Reda	M	20.0	FM	Furniture	II	Extension	2009	515	157	675	No	Unemployed	Soft Skill	High
3	STU_00004	Tadese Hailu	M	17.0	GMFA	Manufacturing	II	Regular	2008	504	272	773	No	Employed	Technical Skill	High
4	STU_00005	Genet Haftu Gebrehiwet	F	21.0	Mechatronics	Electrical	IV	Extension	2009	557	237	793	No	Employed	Digital Skill	Medium

Appendix 4 Managing Data Duplication

```
df[df.duplicated()]
```

	Sex	Age	Occupation	Sector	Level	Program	Graduate Year	In_TVET_Training_hrs	In_Company_Training_hrs	Total_Traning_hrs	Intership	Employment_Status	Skill_GAP_Type	Skill_GAP_Level
792	F	19.839672	FM	Furniture	II	Extension	2009	518	155	674	No	Unemployed	Soft Skill	High
1957	F	19.839672	HNS	ICT	III	Extension	2008	677	165	837	No	Employed	Digital Skill	High
2111	F	19.839672	BBC	Construction	II	Regular	2009	542	151	705	No	Unemployed	Digital Skill	High
2929	F	19.839672	HKO	Hotel & Tourism	II	Regular	2008	595	164	759	Yes	Employed	Digital Skill	Low
3030	M	19.839672	IEMD	Electrical	II	Extension	2010	296	150	440	No	Unemployed	Technical Skill	High
3218	M	19.839672	BEI	Electrical	III	Extension	2008	510	223	742	No	Unemployed	Technical Skill	High
4096	M	19.839672	Mechanics	Manufacturing	III	Regular	2010	422	179	614	Yes	Employed	Technical Skill	Medium
4491	F	19.839672	BBC	Construction	II	Regular	2010	534	150	704	No	Unemployed	Technical Skill	High

Appendix 5 Numerical Representation of Features

```
df.head()
```

	Sex	Age	Occupation	Sector	Level	Program	Graduate Year	In_TVET_Training_hrs	In_Company_Training_hrs	Total_Traning_hrs	Intership	Employment_Status	Skill_GAP_Type	Skill_GAP_Level
0	0	19.839672	2	2	1	1	2011	670	52	710	1	1	2	2
1	0	19.839672	4	6	2	1	2016	772	85	861	1	0	1	1
2	1	19.839672	5	4	0	0	2009	515	157	675	0	1	1	0
3	1	19.839672	6	7	0	1	2008	504	272	773	0	0	2	0
4	0	19.839672	16	3	2	0	2009	557	237	793	0	0	0	2