A data driven Machine Learning model for industry skill demand prediction and TVET admission alignment in Kenya

Jumbe Divison Caleb

MSC/COM/001/22

A research proposal submitted to the graduate school in partial fulfilment of for the requirement of degree of masters of computer science of Kibabi University

January 2026

Declaration

"This research proposal is my original work, prepared with no other than the indicated sources and support, and has not been presented elsewhere for a degree or any other award"

Signature ……………………………………………Date………………………………..

Supervisor s Certification

"This research proposal has been submitted for examination with our approval as university supervisors".

Signature ……………………………………………Date………………………………..

Name:

Department

University/institution

Signature ……………………………………………Date………………………………..

Name:

Department

University/institution

Abstract

Aligning education systems with evolving labor market demands has become a critical priority, unemployment remains a socio-economic challenge, especially among young people in developing economies. While most existing studies apply machine learning to predict individual student or graduate employability outcomes, less attention has been given to forecasting industry-level employment opportunities and integrating such forecasts into education admission planning. Admissions into academic and Technical and Vocational Education and Training (TVET) programs are often supply-driven rather than demand-informed, This contribute to graduate unemployment, and many end up hating Education; Young People will rather stay at home instead enrolling for a course in Tvet or even Universities. In this study I shift the analytical focus from employability prediction to opportunity forecasting by developing a machine learning model for forecasting future industry employment opportunities in Kenya. Using secondary data on historical and current labor market data trends including job vacancies, industry growth indicators, and skill requirements. This study models the labor market as a network of interconnected entities such as industries, skills, education programs, and economic indicators. Graph Neural Networks (GNNs), together with clustering and supervised learning techniques, are employed to learn structural patterns and forecast future industry demand. The predicted industry opportunity trends are then mapped to relevant TVET and academic programs to generate data-driven admission guidance that aligns enrollment decisions with projected labor market needs. Empirical validation using secondary labor market datasets demonstrates that the proposed framework improves forecasting accuracy and reveals hidden relationships between training pathways and employment opportunities. The findings indicate that opportunity-driven, network-centric labor market modeling provides a more proactive basis for admissions planning, workforce development, and education policy formulation. This research contributes a scalable and adaptive artificial intelligence models that supports evidence-based alignment between education systems and evolving labor market demands in Kenya.

ABBREVIATIONS AND ACRONYMS

GNN – Graph Neural Network

TVET – Technical and Vocational Education and Training

Contents

Table of figures

List of tables

List of Equations

CHAPTER ONE: INTRODUCTION

This chapter address the back ground of the study, statement of the problem, purpose, main and specific objective, justification, importance and scope of the study.

1.1 Background of the Study

The transition from education to employment remains a major challenge for many graduates, particularly in developing economies where labor markets are rapidly evolving.

Traditionally, education systems—especially Technical and Vocational Education and Training (TVET) institutions—have been positioned as key drivers of employability and workforce development.

Though there have been increased access to education and training in recent years, there is still, persistent youth unemployment, we need an effective alignment between admissions, curriculum offerings and labor market opportunities.

In Kenya, youth unemployment remains disproportionately high among individuals aged 15–34, even as the country continues to expand TVET infrastructure and enrollment. Although the economy generates new jobs annually, most opportunities are concentrated in the informal sector, offering limited job security and career progression.

This imbalance highlights a structural disconnect between the fields of training students are admitted into and the actual opportunities available within the labor market.

Recent advances in machine learning and artificial intelligence have transformed labor market analysis by enabling predictive modeling of employment trends and skill demand.

Existing studies largely focus on predicting graduate employability outcomes using academic and demographic variables ((al, 2025); (Alalwan, 2024); (Wang, 2025) While these approaches improve post-training assessment, they remain reactive—evaluating employability after students have already completed their programs.

Global policy institutions such as UNESCO (2025), the World Economic Forum (2024), and the OECD (2024) emphasize the need for education systems to adopt predictive, demand-driven planning models that anticipate labor market needs before training decisions are made. These reports advocate for the integration of artificial intelligence into education planning to reduce skills mismatch and improve workforce relevance.

Despite these recommendations, there is limited empirical research—particularly within the Kenyan TVET context—on using machine learning to forecast future industry employment opportunities and applying such forecasts to guide admission decisions.

Moreover, most existing models treat labor market variables independently, failing to capture the complex relationships between industries, skills, education programs, and economic indicators.

This study addresses this gap by proposing a machine learning model that models the labor market as a networked system and predicts future industry opportunities. By linking these predictions to TVET and academic admissions, the study advances a proactive, data-driven approach to aligning education pathways with labor market demand.

1.2. Problem statement

Unemployment remains a persistent socio-economic challenge, particularly among young people in Kenya, despite increased investment in education and skills development. Technical and Vocational Education and Training (TVET) institutions continue to enroll large numbers of students; however, many graduates remain unemployed or underemployed, indicating a misalignment between training pathways and labor market opportunities.

Most existing analytical approaches to addressing this challenge like time series rely on traditional econometric models or descriptive labor market statistics. These methods are limited by linear assumptions and aggregated data, which fail to capture the complex, interconnected relationships between industries, skills, education programs, and economic conditions. As a result, they provide limited predictive capability and insufficient guidance for admission planning and workforce development.

Recent machine learning studies improve employability prediction accuracy but primarily focus on individual graduate attributes, such as academic performance and demographic factors (Mahdin et al., 2025; Alalwan et al., 2024). While valuable, these models remain reactive and do not forecast

future industry demand or inform admission decisions prior to training. Consequently, TVET institutions often base admissions on institutional capacity rather than projected labor market opportunities, exacerbating skills mismatch and graduate unemployment.

In the Kenyan context, existing studies acknowledge the influence of digital skills and program selection on employability ((Muthoni, 2024); (Ochieng, 2024)) but stop short of developing predictive models that integrate labor market demand forecasting with admissions planning. Furthermore, limited use has been made of machine learning techniques capable of modeling relational labor market structures.

There is therefore a critical need for a data-driven, predictive, and context-aware machine learning model that can forecast future industry employment opportunities and translate these forecasts into actionable admission guidance for TVET and academic institutions. Addressing this gap is essential for reducing skills mismatch, improving graduate employment outcomes, and strengthening the alignment between education systems and labor market realities in Kenya.

Purpose of the Study

The purpose of this study is to develop a machine learning model that forecasts future industry employment opportunities and applies the forecasts to guide admission in TVET institutions.

General Objective

To develop and evaluate a machine learning model for forecasting industry skill demand and aligning them to TVET admissions in Kenya.

Objectives of the study

1. To preprocess and integrate industry Skills demand and TVET admission data for machine learning analysis
2. To model industry skill demand through feature engineering techniques.
3. To develop machine learning model for predicting industry skill demand.
4. To Evaluate the model

Research Questions

1. How can industry and TVET admission data be effectively preprocessed and integrated to support machine learning analysis of skill demand?
2. Which feature engineering techniques best capture and represent industry skill demand from integrated labour market and TVET datasets?
3. Which machine learning models are most effective for predicting industry skill demand based on the engineered features?
4. How accurately do the develeped machine learning models predict industry skill demand, based on appropriate evaluation metrics and validation techniques?

Justification of the study

This study is grounded in the methodological and conceptual limitations of existing unemployment and workforce planning approaches.

Advanced Technology in machine learning has demonstrated improved accuracy in predicting graduate employability outcomes. However, most existing models focus primarily on individual-level characteristics such as academic performance, demographic attributes, and personal competencies. These models assess employment outcomes after training has occurred.

They do not forecast future industry demand or provide guidance for admission decisions prior to enrollment. TVET institutions often base admissions on institutional capacity and historical trends rather than projected labor market opportunities, this exacerbating skills mismatch and graduate unemployment.

In Kenyan, existing studies acknowledge the role of program selection and digital skills in influencing employability outcomes. These studies stop does not develop integrated predictive frameworks that link labor market demand forecasting with admissions planning.

No applicable machine learning techniques capable of modeling relational and network-based labor market structures, example interdependencies between industries, skills, educational programs, and employment outcomes. This represents a significant gap in both theory and practice.

Therefore there is a need to develop a data-driven, predictive machine learning framework that can model the complex structure of Kenya’s labor market, and also translate this into actionable admission guidance.

By adopting machine learning algorithms, the study will conceptualizes the labor market as an interconnected network, enabling the capture of non-linear relationships and dependencies that are overlooked by traditional and attribute-based models. This innovation will enhance the ability to forecast future industry employment opportunities.

The use of real-world Kenyan labor market and education data strengthens the external validity and practical relevance of the proposed model.

The visualization and evaluation of the model’s outputs provide interpretable insights for policymakers, TVET administrators, and workforce development agencies,. Hence this will support evidence-based admission planning, and skills development strategies.

These insights contribute directly to reducing skills mismatch and improving graduate employment outcomes.

This study also contributes to the literature by extending the application of machine learning to unemployment modeling and education–labor market alignment in a developing economy context.

It also offers a scalable and adaptable framework that can inform policy reforms, enhance workforce responsiveness, and support sustainable employment strategies.

Significance of the Study

This study holds both academic and practical significance in addressing the persistent challenge of youth unemployment through a data-driven and intelligent modeling approach. By integrating machine learning techniques, the research provides an innovative framework for analyzing the complex relationships among education systems, labor market dynamics, and economic trends—factors that traditional statistical models often fail to capture.

Academically, the study contributes to existing literature by introducing a novel analytical model that bridges the gap between technical and vocational education (TVET) and labor market forecasting. It synthesizes strategies for aligning educational outcomes with real-world skill demands, offering a comprehensive understanding of how predictive modeling can inform curriculum development, policy planning, and institutional decision-making. The proposed framework expands the body of knowledge in machine learning applications, education–employment linkage studies, and labor market analytics.

Practically, the study's implementation of a forecasting model using real-world data enhances the reliability and relevance of its findings. By evaluating the model's performance and predictive accuracy, the study demonstrates how artificial intelligence and graph theory can support strategic decision-making in employment policy and workforce planning. The model's data visualization outputs provide an intuitive way for stakeholders—such as policymakers, educators, and employers—to interpret and act on emerging unemployment and skill-demand patterns.

The outcomes of this research will have direct implications for TVET institutions, enabling them to tailor their curricula and admissions based on predicted market needs. Governments and policymakers will gain actionable insights for designing targeted employment strategies, while employers can use the findings to anticipate and address future workforce gaps.

Ultimately, this study is significant because it advances both theoretical understanding and practical interventions in tackling unemployment. By capturing the interdependencies between education systems, labor markets, and economic dynamics through a graph-based machine learning model, it provides a robust foundation for evidence-based planning, improved graduate employability, and sustainable workforce development.

Scope

The study begins with a comprehensive review of existing research on unemployment modeling and workforce analysis, with particular emphasis on strategies that align educational outcomes with labor market demands. This review is limited to peer-reviewed journals, policy reports, and academic sources published within the past 1–5 years.

The implementation is based on publicly available or institutionally sourced datasets, which may include national labor statistics, education enrollment and graduation data, and employment trends. The study is limited to data that is structured, reliable, and relevant to the research objectives.

The performance of the model will be evaluated based on its predictive accuracy and its ability to offer meaningful insights. Visualization tools will be applied to present results in a form that supports data-driven decision-making for policymakers and educators.

The evaluation does not extend to longitudinal policy impact or implementation at a national level.

This study is limited to the scope of model development and performance assessment and does not cover direct intervention or policy implementation outcomes. Geographically, the focus may be constrained to one or a few representative regions or countries, depending on data availability

The study primarily focuses on tvets institutions in Kenya, considering local labor market conditions, educational systems, and economic trends.

However, the model's framework is scalable and can be adapted for other regions in Africa and Asia where youth unemployment from tvets has always been a major issue for years

Limitations of the Study

The study relies on available data sources, and any gaps or inconsistencies in labor market data may affect accuracy.

The model's generalizability may be limited to regions with comparable economic structures.

The research does not address socio-political factors (e.g., labor laws, government policies) that may also influence unemployment.

Assumptions of the Study

Availability of Reliable Data: It is assumed that the datasets used for model training and evaluation (e.g., unemployment statistics, educational outcomes, and labor market data) are accurate, up-to-date, and representative of real-world conditions.

Existence of Measurable Relationships: The study assumes that there are measurable and meaningful relationships between educational qualifications, labor market demands, and unemployment trends that can be captured using graph-based machine learning techniques.

Consistency in Classification and Labeling: It is assumed that variables such as job categories, education levels, and unemployment indicators are consistently defined and labeled across different data sources used in the study.

Users' Interpretation of Visualizations: The study assumes that the visual outputs generated from the model will be interpretable and meaningful to stakeholders, such as policymakers and educators, facilitating informed decision-making.

Conceptual Framework

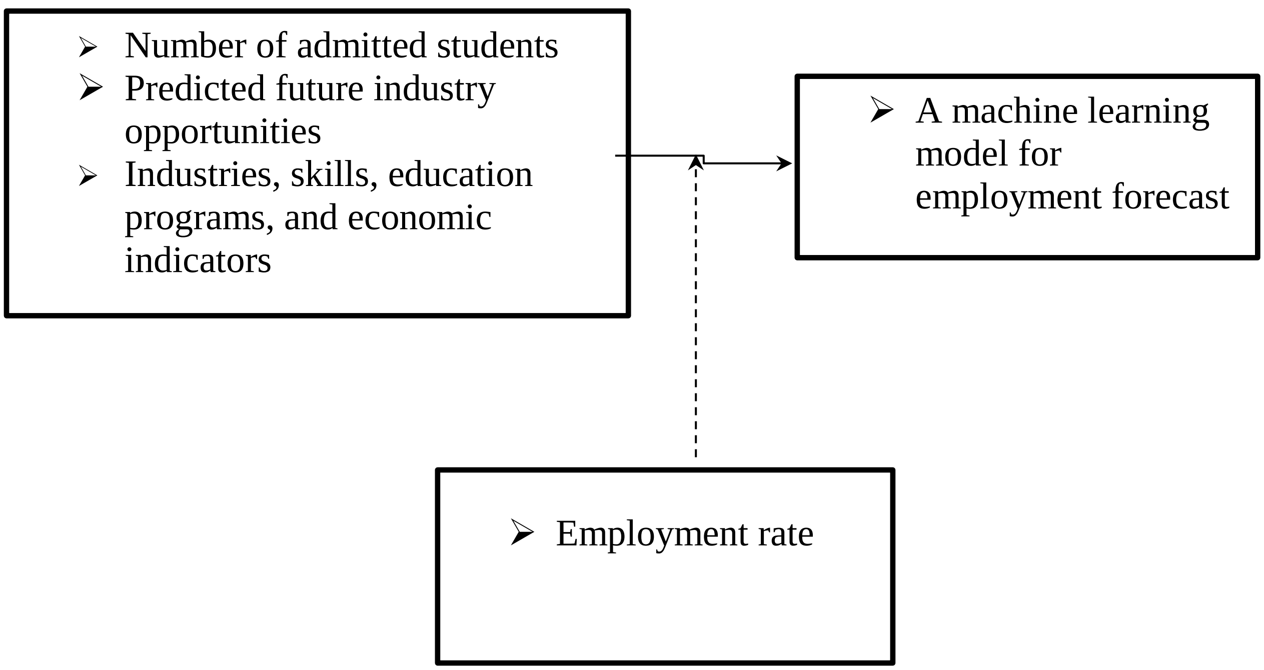


Figure 1 1.1 conceptual framework

CHAPTER TWO: LITERATURE REVIEW

2.1INTRODUCTION

This chapter reviews existing literature related to the use of data science and machine learning techniques in analyzing industry demand, forecasting employment opportunities, and informing higher education and TVET admission decisions. The review is guided by the study objectives: To learn patterns in industry demand and skill requirements across economic sectors using machine learning, (ii) to develop a machine learning model for forecasting future industry employment opportunities using historical labor market data, and (iii) to design a data-driven admission guidance model that maps predicted industry opportunities to academic and TVET program intake decisions. In addition, the chapter discusses the conceptual framework underpinning the study and the operationalization of key variables.

2.2 To preprocess and integrate industry Skills demand and TVET admission data for machine learning analysis

The first objective of this study is to learn patterns in industry demand and skill requirements across economic sectors using machine learning. Recent studies demonstrate increasing adoption of machine learning techniques to analyze employability and skill-demand patterns, driven by persistent skills mismatches and graduate unemployment.

Alalwan et al. (2024) compared Logistic Regression, Random Forest, and Support Vector Machine models to predict graduate employability using academic and experiential attributes. Their findings show that non-linear and ensemble models outperform traditional linear approaches, indicating the suitability of machine learning for capturing complex skill–employment relationships. However, the study relies on static, individual-level attributes and does not explicitly analyze sector-wide skill demand patterns across industries.

Similarly, Wang (2025) proposed a machine learning–based curriculum optimization framework to align vocational education programs with labor market requirements. The study demonstrates that data-informed curriculum design improves graduate employability by reflecting evolving industry needs. Despite this contribution, the framework focuses on curriculum structure and does not model cross-sectoral skill demand patterns or interdependencies between industries, skills, and training pathways.

In the African context, Ochieng (2024) examined the relationship between digital skills, technology acceptance, and employability among TVET graduates in Kenya. The study confirms that digital competencies significantly influence employability outcomes. However, the analysis adopts a descriptive and correlational approach and does not apply machine learning techniques to uncover latent or structural patterns in industry-wide skill demand.

Global policy reports by UNESCO (2025) and the OECD (2024) emphasize the growing demand for digital, transversal, and adaptive skills across economic sectors and advocate for the use of artificial intelligence to support skills anticipation. While these reports highlight macro-level skill trends, they do not provide computational models capable of systematically identifying and learning industry-specific skill demand patterns from labor market data.

Overall, existing studies acknowledge the importance of machine learning and data analytics in skills analysis but provide limited empirical modeling of sector-wide industry demand patterns. This study addresses this gap by applying machine learning particularly graph-based techniques to identify structural and cross-sectoral patterns in industry skill requirements.

Table 1.1 Summary of variables used with their strengths and limitations.

N	Title	Authors & Year	Variables Used	Strengths	Limitations
1	Predicting Graduate Employability Using Machine Learning Models	Alalwan et al. (2024)	Academic performance, internship experience, technical skills, soft skills	Demonstrates superior performance of non-linear and ensemble models; Captures complex relationships between skills and employability	Focuses on individual-level attributes; Does not model sector-wide industry demand patterns
2	Machine Learning–Based Curriculum Optimization for Vocational Education	Wang (2025)	Curriculum modules, labor market demand indicators, graduate employment outcomes	Supports data-driven curriculum design; Improves alignment between training and labor market needs	Does not capture cross-sectoral skill interdependencies; Limited focus on industry-wide demand patterns
3	Digital Skills and Employability Outcomes among TVET Graduates in Kenya	Ochieng (2024)	Digital competencies, technology acceptance, employment status	Highlights importance of digital skills in employability; Context-specific to Kenyan TVET institutions	Uses descriptive and correlational methods; Lacks machine learning-based pattern discovery
4	Skill Demand Prediction Using Job Advertisement Analytics	Liu et al. (2023)	Job postings, skill keywords, industry categories, employment trends	Effectively captures real-time industry skill demand; Uses large-scale labor market data	Heavily dependent on online job platforms; Informal sector skills may be underrepresented

5	Graph-Based Modeling of Labor Market Skill Networks	Khaouja et al. (2022)	Industry sectors, skill nodes, job transitions, skill co-occurrence	Identifies structural relationships and skill clusters across industries; Captures cross-sectoral patterns	Requires complex graph construction; Computationally intensive for large datasets
6	Artificial Intelligence for Skills Anticipation and Workforce Planning	OECD (2024)	Occupational data, skill taxonomies, economic sector indicators	Provides macro-level insights into emerging skill demands; Supports policy-level decision-making	Lacks operational machine learning models; Not tailored to institutional-level admissions decisions
7	Using Big Data Analytics to Forecast Workforce Skill Requirements	UNESCO (2025)	Labor market statistics, education data, digital skill indicators	Emphasizes AI-driven skills anticipation; Addresses global workforce trends	Focuses on policy guidance rather than empirical modeling; Limited sector-specific analysis

2.3 modeling industry skill demand through feature engineering techniques.

The second objective of this study is to apply supervised machine learning techniques to model the labor market as a network of interconnected industries, skills, education programs, and economic indicators. The study recognizes the growing need for predictive, data-driven approaches that move beyond descriptive admission planning methods toward demand-informed decision support systems.

The development process begins with a comprehensive understanding of labor market dynamics and education–employment linkages, including industry growth trends, skill requirements, and training pathways. This is followed by the identification of relevant secondary data sources such as job vacancy records, industry employment statistics, skill taxonomies, and TVET and academic program data. These datasets are integrated to form a unified labor market dataset capable of supporting supervised learning tasks.

Subsequently, data preprocessing and feature engineering are conducted to transform raw labor market data into meaningful predictive features. These include industry-level employment indicators, skill demand frequencies, program–skill mappings, and economic signals. Normalization, encoding, and dimensionality reduction techniques are applied to ensure data consistency and model efficiency.

Supervised machine learning algorithms are then employed to learn predictive relationships between historical labor market patterns and future employment demand. Unlike traditional admission planning approaches that rely on expert judgment or static statistics, the proposed approach explicitly models relationships between industries, skills, and education programs. To capture these interdependencies, the labor market is represented as a network structure in which nodes correspond to industries, skills, and programs, while edges represent demand, skill relevance, or transition relationships.

Model training is conducted iteratively, with performance evaluated using appropriate classification and regression metrics. The trained models are subsequently used to generate forecasts of industry employment demand, which are mapped to corresponding academic and TVET programs to support data-driven admission guidance. Through this approach, the study provides an integrated framework that translates labor market intelligence into actionable enrollment planning insights.

This methodology directly addresses limitations identified in existing studies, which often focus on individual employability prediction, rely heavily on academic performance indicators, and overlook structural labor market relationships (Alalwan, 2024; Al et al., 2025). By incorporating multi-dimensional employability features and network-based modeling, the proposed system supports evidence-based, demand-informed admission planning aligned with Kenya's evolving labor market needs.

2.4 Development of a Machine Learning Model for Forecasting Future Industry Employment Opportunities

The third objective of this study is to develop a machine learning model for forecasting future industry employment opportunities using historical labor market data. This section reviews existing literature related to employment forecasting, labor market analytics, and the application of machine learning techniques in predicting workforce demand.

In recent years, the need to forecast future employment opportunities has intensified due to rapid technological change, shifting industry structures, and persistent skills mismatches between education systems and labor market requirements. Advances in data availability, computational power, and machine learning techniques have enabled more sophisticated analysis of labor market trends beyond traditional econometric forecasting methods (World Economic Forum, 2024). These developments have encouraged the adoption of data-driven approaches capable of identifying complex, non-linear relationships within historical employment data.

Traditionally, labor market analysis has focused on retrospective assessment of employment outcomes or descriptive statistics of workforce participation. However, such approaches are limited in their ability to anticipate future employment demand across industries. As noted by the OECD (2024), proactive forecasting models are essential for guiding education and training systems toward emerging opportunities rather than reacting to labor market imbalances after they occur.

Several studies have applied machine learning techniques to employability-related problems, particularly at the individual graduate level. Al et al. (2025) utilized hybrid machine learning approaches combining regression and clustering techniques to predict employability outcomes among TVET graduates. Their findings indicate that hybrid models outperform single-algorithm approaches in predictive accuracy. Despite these strengths, the study focuses on post-training employability classification and does not extend to forecasting future industry-level employment demand.

Similarly, Wang (2025) demonstrated the effectiveness of machine learning in optimizing vocational education curricula based on observed labor market signals. While the framework improves alignment between training programs and current industry needs, it remains reactive in nature and does not incorporate temporal modeling of historical labor market data to predict future employment opportunities.

Emerging preprint studies such as *Unlocking Futures: A Machine Learning-Based Career and Employability Prediction System* (2024) and *Skill-Driven Certification Pathways and Their Impact on Graduate Employability* (2025) further highlight the potential of supervised learning and natural language processing techniques in modeling employability trajectories. However, these studies are largely experimental, lack rigorous validation, and do not address industry-wide employment forecasting within structured TVET systems.

At the policy level, global institutions such as the World Economic Forum (2024) and the OECD (2024) emphasize the importance of predictive analytics and skills-first labor market systems to anticipate future workforce needs. Despite these recommendations, there remains limited empirical research that operationalizes machine learning models to forecast future industry employment opportunities using longitudinal labor market data, particularly in developing economies such as Kenya.

In response to these gaps, this study develops and validates a machine learning model that explicitly forecasts future industry employment demand using historical labor market data. By shifting the analytical focus from reactive employability prediction to proactive employment opportunity forecasting, the proposed model aims to support evidence-based TVET admissions planning and enhance alignment between training pathways and Kenya's evolving labor market needs.

2.4 Model Testing

Testing the developed machine learning model constitutes the fourth objective of this study. Model testing is a fundamental phase in the machine learning lifecycle, as it assesses the ability of a trained model to perform accurately when applied to unseen data and real-world scenarios (Goodfellow, Bengio, & Courville, 2016). In this study, model testing is essential for evaluating the accuracy and reliability of the proposed system in predicting industry skills demand and aligning these demands with admissions decisions in Technical and Vocational Education and Training (TVET) institutions in Kenya.

Given the dynamic nature of labor market trends and the possibility of sparsity or imbalance in historical skills demand data, rigorous model testing is necessary to ensure effective generalization beyond the training dataset (Hastie, Tibshirani, & Friedman, 2017). The testing process is carried out using an independent hold-out test dataset that was not involved in the training or validation phases. This dataset closely reflects real-world industry demand patterns and TVET enrollment data, thereby enabling an objective evaluation of the model's predictive performance.

Since the developed system addresses a classification and prediction problem, appropriate performance evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrices are employed to assess the effectiveness of the model (Powers, 2020). These metrics provide detailed insights into the model's ability to correctly identify high-demand industry skills and recommend suitable TVET admission pathways. Through systematic testing, this study confirms the robustness, adaptability, and practical applicability of the proposed machine learning model in supporting data-driven TVET admissions and strengthening alignment with Kenya's labor market requirements.

CHAPTER THREE: RESEARCH METHODOLOGY

3.0 Introduction

This chapter describes the research methodology adopted for the study using the **Design Science Research (DSR)** approach. The study focuses on the design, development, and evaluation of a **machine learning decision-supportmodel** intended to forecast employment opportunities and guide admission decisions in Technical and Vocational Education and Training (TVET) institutions in Kenya.

The motivation for adopting DSR arises from persistent skills mismatches, graduate unemployment, and limited alignment between TVET program intake and labor market demand. Existing admission planning approaches are largely reactive and lack predictive, data-driven support. In response, this study applies DSR to systematically create and evaluate an artifact that integrates graduate attributes, industry demand data, and predictive analytics to support evidence-based admission planning.

The chapter outlines the DSR framework, research context, data sources, artifact design and development, model evaluation procedures, and ethical considerations.

3.1 Design Science Research (DSR) Approach

Design Science Research is a problem-solving research paradigm that focuses on the **creation and evaluation of artifacts** designed to address identified real-world problems. In this study, DSR is adopted to guide the systematic development of a **predictive employability and admission guidance model** for TVET institutions.

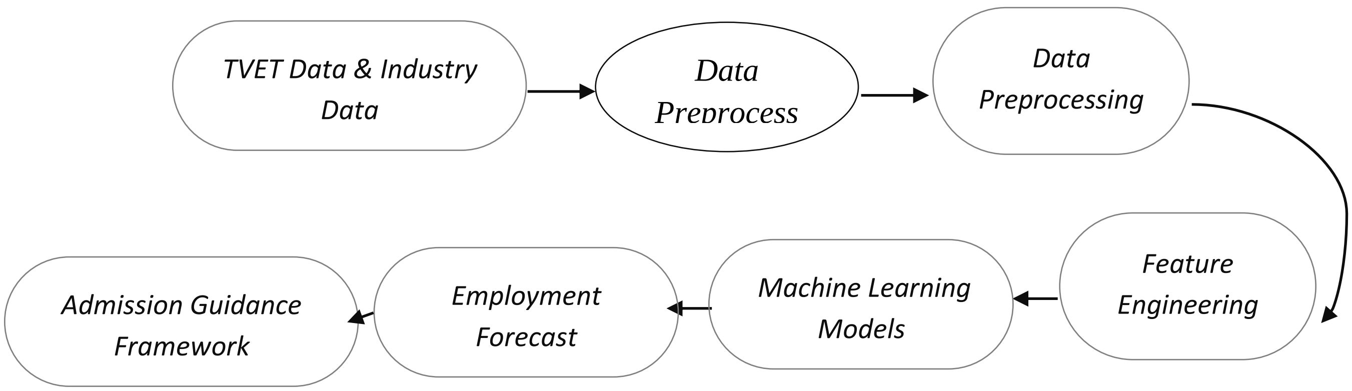
DSR is appropriate for this research because it emphasizes:

- Practical relevance through artifact creation
- Iterative design, development, and evaluation
- Integration of quantitative analytical techniques
- Generation of prescriptive knowledge applicable to real-world contexts

The primary artifact developed in this study consists of a **machine learning employment forecasting model** and a **data-driven admission guidance framework** tailored to the Kenyan TVET system.

(TVET Data & Industry Data → Data Preprocessing → Feature Engineering → Machine Learning Models → Employment Forecast → Admission Guidance Framework)

Figure 2: Research Design and Analytical Framework



3.2 Design Science Research Phases

Consistent with established DSR literature, the study follows six core phases:

1. Problem Identification and Motivation
2. Definition of Objectives for a Solution
3. Design and Development of the Artifact
4. Demonstration
5. Evaluation
6. Communication

Each phase is explicitly mapped to the TVET employability and admission planning context.

3.2.1 Problem Identification and Motivation

The problem identification phase focuses on understanding the misalignment between TVET graduate supply and industry skill demand in Kenya. Despite ongoing reforms, many TVET graduates face unemployment or underemployment, while industries report skills shortages in key sectors.

A review of institutional reports, graduate tracer studies, and labor market data reveals that admission decisions are often made without systematic forecasting of industry demand.

Furthermore, existing studies on employability prediction have largely overlooked Sub-Saharan Africa, limiting the availability of context-specific decision-support tools.

This gap motivates the development of a predictive, data-driven artifact capable of forecasting employment opportunities and aligning TVET admissions with labor market needs. The DSR approach ensures that the problem is grounded in both theoretical gaps and practical challenges faced by TVET institutions and policymakers.

3.2.2 Definition of Objectives for a Solution

Guided by the identified problem, the objectives of the solution are to:

- Design a machine learning–based model that predicts employment opportunities using graduate attributes and industry demand indicators
- Integrate historical TVET data and labor market information into a unified analytical framework
- Develop an admission guidance artifact that aligns predicted industry demand with TVET program intake
- Evaluate the predictive accuracy and practical usefulness of the proposed artifact

These objectives directly inform the design and development of the research artifact and ensure alignment with stakeholder needs.

3.2.3 Design and Development of the Artifact

The design and development phase focuses on constructing the core artifact of the study: a **predictive employability and admission guidance system**.

3.2.3.1 Study Context

The artifact is designed for application within Kenyan TVET institutions. Kenya provides a relevant context due to documented skills mismatches, graduate unemployment, and policy initiatives promoting competency-based and market-driven TVET systems.

3.2.3.2 Data Sources

The artifact relies on secondary data, supplemented by limited primary data for contextual validation.

Secondary data sources include:

- Online data sources like
    - Graduate tracer studies and employability reports
    - Industry and labor market reports detailing skill demand and sectoral growth
    - TVET institutional records on enrollment, program offerings, completion rates, and graduate outcomes

Primary data sources include:

- TVET institutional records on enrollment, program offerings, completion rates, and graduate outcomes
- Graduate tracer studies and employability reports
- Key informant interviews with policymakers and industry representatives

3.2.3.3 Data Collection and Preprocessing

Data collection follows a structured process:

- Mapping of TVET programs and enrollment capacities
- Compilation of graduate attributes including academic performance, technical skills, soft skills, and industrial attachment experience
- Collection of industry demand indicators and occupational trends
- Data integration, normalization, handling of missing values, and removal of duplicates
- Anonymization of records to ensure ethical compliance

3.2.3.4 Sampling Techniques

A stratified sampling approach is employed to ensure representativeness.

TVET institutions are stratified by:

- Geographic region
- Institutional size
- Program focus areas

Industry data are stratified by:

- Economic sector
- Skill level (artisan, certificate, diploma)

3.2.3.5 Feature Engineering

A multidimensional feature set is constructed comprising:

- Academic performance indicators
- Technical competencies
- Soft skills attributes
- Industrial attachment and work-based learning experience
- Certifications and micro-credentials
- Digital and technology adoption skills

A normalized feature vector for graduate iii is defined as:

Equation 1 A normalized feature vector for graduate

Xi=[xi1,xi2,xi3,…,xin]X_i = [x_{i1}, x_{i2}, x_{i3}, \ldots, x_{in}]Xi=[xi1,xi2,xi3,…,xin]

3.2.3.6 Machine Learning Model Design

The artifact incorporates supervised and hybrid machine learning techniques, including:

- Logistic Regression
- Random Forest
- Support Vector Machines
- Neural Networks
- Graph Neural Networks (GNNs) and clustering techniques

For Logistic Regression, employability probability is modeled as:

Equation 2Logistic Regression, employability probability

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

These models are selected based on their predictive capability and suitability for capturing complex employability patterns.

3.2.4 Demonstration

The demonstration phase involves applying the developed artifact to historical TVET and labor market data to generate employment forecasts and admission recommendations.

Employment demand is forecast as:

Equation 3 Employment demand

E^t+1=f(Xt,St,Gt)\hat{E}_{t+1} = f(X_t, S_t, G_t)E^t+1=f(Xt,St,Gt)

where E^t+1\hat{E}_{t+1}E^t+1 represents predicted employment demand, XtX_tXt graduate skills, StS_tSt sector indicators, and GtG_tGt growth variables.

3.2.5 Evaluation

The artifact is evaluated using train–test splits and k-fold cross-validation:

Equation 4 train–test splits and k-fold cross-validation

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

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$ROC\text{-}AUC = \int_0^1 TPR(FPR)\, d(FPR)$$

Model performance is assessed using:

- Accuracy
- Precision
- Recall
- F1-score
- ROC–AUC

The best-performing model is selected based on predictive accuracy and robustness.

An admission alignment score is computed as:

Equation 5 An admission alignment score

$S_p = \frac{D_p}{S_p}$where $\hat{E}_{t+1}$

where $D_p$represents predicted industry demand and $S_p$graduate supply for program p.

3.2.6 Communication

The final phase involves communicating the research findings and artifact to relevant stakeholders, including TVET administrators, policymakers, and academic audiences. Results are presented through analytical reports, visualizations, and an admission guidance framework that highlights programs with excess supply or unmet industry demand.

3.3 Ethical Considerations

Ethical standards are strictly observed. Personal identifiers are removed, informed consent is obtained for primary data collection, and data usage complies with institutional and national data protection regulations.

3.4 Limitations of the Study

The study acknowledges the following limitations:

- Dependence on secondary data quality and completeness
- Potential time lags in labor market data
- Limited generalizability beyond the Kenyan TVET context

Despite these limitations, the DSR-based methodology provides a robust, replicable, and practically relevant framework for data-driven admission planning and employability forecasting.

References

al, M. H. (2025). Predictive Analytics for Employability in Technical and Vocational Education and Training (TVET) Using Hybrid Machine Learning Approaches. *International Journal on Informatics Visualization.*, 15-21.

Alalwan, A. e. (2024). Machine Learning for Predicting Students' Employability Outcomes. *Education and Information Technologies.*, 26.

Forum, W. E. (2024). Putting Skills First: Improving Graduate Employability through Data-Driven Education Systems. *World Economic Forum Report.*, 34.

Muthoni, J. &. (2024). Machine Learning Model for Program Selection in Technical and Vocational Education and Training (TVET) in Kenya. *African Journal of Information Systems.*, 81.

Ochieng, D. (2024). Technology Acceptance and Employability among TVET Graduates in Kenya. *International Journal of Training Research.*, 19.

OECD. (2024). Skills Outlook 2024: Skills for a Digital Future. *OECD Publishing.*, 9.

Preprint, A. (2025). Skill-Driven Certification Pathways and Their Impact on Graduate Employability. *ArXiv Preprint*, 9.

Preprint, A. (2024). Unlocking Futures: A Machine Learning-Based Career and Employability Prediction System. *predict career and employability trajectories of graduates*, 34.

UNESCO. (2025). Artificial Intelligence in Technical and Vocational Education and Training: Implications for Skills and Employability. 41.

Wang, Y. (2025). Optimizing Vocational Education Curriculum Systems Using Machine Learning to Enhance Students' Employability. *Journal of Computational Methods in Computer Science*, 15.