



Title of project placed here



Name of author placed here

School of Computing Science Sir Alwyn Williams Building University of Glasgow

G12 8QQ

A dissertation presented in part fulfilment of the requirements of the Degree of Master of Science at The University of Glasgow

Date of submission placed here

**Abstract**

abstract goes here

## Education Use Consent

I hereby give my permission for this project to be shown to other University of Glasgow students and to be distributed in an electronic format. **Please note that you are under no obligation to sign this declaration, but doing so would help future students.**

Name: Signature:

## Acknowledgements

acknowledgements go here

# Contents

1. Introduction 4
   1. [A section . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4](#_TOC_250001)
      1. [A subsection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4](#_TOC_250000)

|  |  |  |  |
| --- | --- | --- | --- |
| **2** | **Survey** |  | **5** |
| **3** | **Further chapters** |  | **6** |
| **4** | **Conclusion** |  | **7** |
| **A** | **First appendix** |  | **8** |
|  | A.1 Section of first appendix | . . . . . . . . . . . . . . . . . . . . . . . . . . . | 8 |
| **B** | **Second appendix** |  | **9** |

**Bibliography** **1**

# Chapter 1: Introduction

## 1.1 Motivation

In an increasingly digital world and age, access to computing education and technical skills has become a critical factor of individual and national economic potential and success. For many African countries, the digital revolution presents both a tremendous opportunity and a significant challenge. While the global demand for software developers, data scientists, and digital professionals grows, many young Africans remain underserved in terms of access to quality training, infrastructure, and support systems necessary to enter these fields.

Organizations such as **CSA Africa, Moringa School, ALX** and Africa Code Academy are working to bridge this gap by providing digital skills training, mentorship, and community support. However, despite significant efforts and growing interest among learners, a large segment of the population continues to face barriers to acquiring core programming skills. These barriers are often multidimensional, ranging from socio-economic constraints and infrastructural limitations to motivational and psychological challenges. Yet, there remains a lack of rigorous, data-driven research, particularly causal studies, that not only quantifies the magnitude of the skills gap but also systematically and comprehensively examines its underlying drivers.

As digital inclusion grows more closely linked to social advancement and economic empowerment, it is critical to understand the nature of these barriers. Traditional analyses have largely focused on descriptive statistics, which, while informative, often fall short in providing the kind of elaborate insights needed for systemic change. A deeper, more technical analysis can uncover hidden patterns, correlations, and even potential causal relationships that are critical to crafting scalable, evidence-based solutions tailored to African contexts.

This research project is situated at the intersection of social impact and data science. It aims to make use of a large-scale dataset collected by CSA Africa from over 2,500 participants across six countries (Algeria, Nigeria, Kenya, Ghana, Malawi, Eswatini, Rwanda, South Sudan, Tanzania, Togo, Uganda, Zambia Zimbabwe, Central African Republic and Botswana). The dataset comprises detailed information on learners’ demographics, access to infrastructure (e.g., electricity, internet, computer), psychological perceptions, self-rated programming competence, motivation levels, and engagement frequency.

Using this dataset, the project seeks to leverage advanced data science methodologies to investigate, model, and explain the digital skills gap. Specifically, the goal is twofold: (1) to **evidence the existence** of a programming skills gap within the African context, and (2) to **identify and rank the cause or contributing elements** responsible for this gap. By doing so, the findings can support more targeted and effective interventions, helping CSA Africa and similar organizations optimize their programs to address the learners’ actual needs and requirements.

## 1.2 Purpose

This project is a research-driven data science investigation focused on uncovering the underlying causes of challenges in computing skills acquisition among young Africans. Its broader objective is to implement, apply, compare, and evaluate multiple analytical and modelling approaches, ranging from traditional machine learning to Bayesian and causal inference techniques, to quantify and explain the factors contributing to observed disparities in digital competencies.

The project’s central research questions are:

**Research Q1:** What evidence is there of a computing skills gap in CSA Africa’s target learner population?

**Research Q2:** What socio-demographic, infrastructural, and psychological factors causally contribute to this computing skills gap?

**Research Q3:** How do predictive, Bayesian, and causal inference models compare in explaining and quantifying the factors driving the computing skills gap in African learners?

To address these questions, the study will:

1. Construct a composite measure of computing skills level or "skill gap" using self-rated competence, programming frequency, and training exposure indicators.
2. Develop and compare three types of modelling frameworks:
   1. **Predictive modelling** using regression techniques to estimate the likelihood of computing skills deficiency and quantify computing skill levels based on observed features e.g. exposure to programming training.
   2. **Bayesian probabilistic modelling** using PyMC to estimate how different factors influence computing skill levels, quantify uncertainty in those effects, and update prior expert and CSA’s beliefs based on observed data.
   3. **Causal inference modelling** using the DoWhy package to model potential cause-effect relationships between various factors and skill gap outcomes.
3. Present a critical evaluation of these approaches in terms of performance, methodological robustness, interpretability, and practical applicability.
4. Deliver actionable insights and recommendations for CSA Africa to tailor its training and outreach strategies more effectively.

The outcome is expected to contribute both to **academic knowledge** in the field of Applied machine learning, Bayesian modelling and Causal inference and to **practical decision-making** in the design of inclusive digital learning programs.

## 1.3 Achievements

The key highlights achievements for this project are as follows

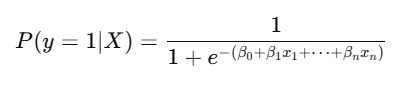
# Chapter 2: Survey

This chapter provides a review of the key analytical methods and literature that inform this study. It outlines the theoretical underpinnings and practical applications of traditional predictive modelling, Bayesian inference, and causal inference approaches. Each modelling paradigm is examined in relation to its strengths, limitations, and suitability for analyzing factors that influence programming skill acquisition. The chapter also highlights relevant research on computational skill development and methodological best practices, positioning this study within broader academic and applied contexts.

## 2.1: Predictive Modelling

Predictive modelling refers to the use of statistical and machine learning algorithms to predict an outcome variable based on input features. In the context of this research, traditional predictive modelling methods are applied in two complementary ways. First, a classification model to *estimate the likelihood* that an individual exhibits a computing skill gap, based on a range of socio-demographic, infrastructural, and psychological variables. This binary formulation helps to identify which features are most strongly associated with the presence or absence of a skill gap. Secondly, a regression model to estimate a continuous skill score that quantifies the degree of computing skills proficiency. This approach enables a more nuanced understanding of how different predictors influence skill levels along a spectrum, rather than as a simple yes-or-no outcome.

### 2.1.1 Logistic Regression - Interpretability and Usefulness in this Research Context

Logistic regression is a widely used classification algorithm that models the probability that a given input belongs to a particular category and in this case, whether a participant has a programming skill gap. It estimates a linear combination of the input features and applies the logistic (sigmoid) function to predict probabilities between 0 and 1[1](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html):

This model is particularly interpretable, allowing direct insights into how each variable influences the probability of skill deficiency. For instance, limited internet access or low perceived competence may significantly increase the likelihood of exhibiting a skills gap.

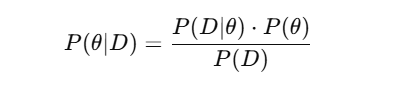
### 2.1.2 Linear Regression - Interpretability and Usefulness in this Research Context

Where the objective is to estimate a continuous skill score (i.e., one that we will calculate as a composite index of programming ability derived from self-assessments of competence, engagement frequency, and prior exposure), linear regression becomes a suitable modelling approach. This model assumes a linear relationship between the target variable and the input features [2](https://scikit-learn.org/stable/modules/linear_model.html#linear-regression) . This allows for a quantifiable interpretation of how features like lack of computer access or absence of prior training contribute to variations in computing skill levels.

## 2.2 Bayesian Modelling

Bayesian modelling is a statistical approach that combines prior beliefs with observed data to produce a **posterior distribution** over parameters of interest. Unlike traditional frequentist models that produce point estimates, Bayesian models provide full probability distributions, allowing researchers to quantify uncertainty and incorporate prior domain knowledge into the analysis. This feature is particularly valuable for social impact problems such as computing skills acquisition, where uncertainty and heterogeneity are common across populations.

Bayesian inference is based on **Bayes’ Theorem**, which updates beliefs about unknown parameters θ after observing data D:

Where:

* P(θ∣D) is the **posterior** probability distribution of the parameters given the data.
* P(D∣θ) is the **likelihood** of the data under the parameters.
* P(θ) the **prior** belief about the parameters before seeing the data.
* P(D) is the **evidence** (a normalizing constant).

### 2.2.1 Probabilistic Programming and PyMC

In this project, Bayesian models will be implemented using **PyMC**, a powerful probabilistic programming library for Python. PyMC allows for the flexible definition of custom probabilistic models and employs advanced sampling methods such as **Markov Chain Monte Carlo (MCMC)** to approximate the posterior distribution of parameters.

### 2.2.2 Bayesian Modelling- Interpretability and Usefulness in this Research Context

Bayesian models offer three key advantages in the context of this research:

1. **Uncertainty quantification** – Instead of producing a single estimate for each factor (e.g., “internet access reduces skill score by 2 units”), The Bayesian model will generate a distribution of likely values. This lets us report ranges (e.g., “we’re 95% confident the effect lies between 1.4 and 2.6”), adding nuance and caution to our findings which is especially useful for real-world decision-making.
2. **Prior integration** – The Bayesian models will allow the incorporation of existing beliefs or expert knowledge into the analysis through the use of priors. This means that if CSA Africa already has hypotheses or past experience about what matters most (e.g., mentorship, infrastructure), these will be formally included in the model and updated in light of the new data.
3. **Rich inference** – The Bayesian method will make it possible to ask and answer probabilistic, policy-relevant questions. For example: What is the probability that learners with unreliable electricity access are less likely to develop coding competence, given observed patterns? These insights are directly actionable for CSA and go beyond simply identifying correlations as they will help estimate **likelihoods and magnitudes under uncertainty.**

In summary, Bayesian modelling enhances this research by providing a deeper and more flexible framework for understanding the drivers of the computing skills gap. It complements the predictive methods done in the step 1 by offering a **more honest representation of uncertainty**, the ability to **incorporate domain expertise**, and a platform for **probabilistic, decision-focused insights**.

## 2.3 Causal Inference Modelling

Understanding what is correlated with a computing skills gap is valuable, but understanding what causes it is more powerful. Causal inference modelling enables us to go beyond associations and make principled, data-driven arguments about **cause-and-effect relationships**. This is especially important in policy and social impact contexts like this study, where the goal is not only to describe disparities but to identify which specific interventions are most likely to improve outcomes.

For example, the earlier predictive models might tell us that learners without internet access tend to have lower skill scores. But only causal models can answer: “If we gave learners internet access, would their skill levels actually improve?”.

### 2.3.1 Causal Modelling with DoWhy

This project employs the **DoWhy** Python library for causal inference, which formalizes the process of identifying causal effects using **structural causal models (SCMs)**. With DoWhy, causal relationships are encoded in a **directed acyclic graph (DAG)** that represents CSA’s assumptions about how variables interact.

The typical causal workflow involves:

* **Identifying** the causal question (e.g., “What is the effect of electricity access on computing skill levels?”)
* **Modelling** the assumptions via a causal graph
* **Estimating** causal effects using statistical techniques (e.g., matching, regression, inverse probability weighting)
* **Refuting** or validating the causal claims using robustness checks and simulated counterfactuals

### 2.3.2 Causal Inference - Interpretability and Usefulness in this Research Context

Causal inference is particularly well-suited to answering CSA Africa’s second core research question: *“What are the factors contributing to the skills gap, and which of these are actually driving it?”*

While predictive and Bayesian models can suggest relationships, they **cannot confirm whether changing a factor would actually impact the outcome**. Causal modelling fills this gap by:

* Differentiating **true causes** from mere correlations
* Enabling **counterfactual reasoning** (e.g., “What if learners had mentorship but everything else stayed the same?”)
* Simulating **interventions** to test which programmatic changes (e.g., providing devices, mentorship, or training) are likely to have the greatest impact

This makes causal inference a natural and essential next step in the research process, one that strengthens the policy relevance and actionability of the findings.

## 2.4 Related Work

Research on computing and digital skills gaps has gained global attention over the past 7 years, particularly as technological transformation accelerates under “Industry 4.0.” Skill gaps generally refer to the mismatch between the skills demanded by employers and those available in the workforce.1 Global reports warn that the widening skills gaps, especially in tech fields, threaten to slow innovation and economic growth if not addressed.3. However, much of the early research and theoretical work has focused on advanced economies, with far less empirical study in developing regions.2 Bhorat et al. (2023) observe that the majority of studies on digitalization concentrate on North America, Europe, and other developed contexts, examining impacts of new technologies (automation, AI, etc.) on employment. In contrast, developing regions like Africa, despite their rapidly growing, youthful labour force, have received far less empirical attention. This gap in the literature is notable because Africa’s population is young and expanding; by 2030, Africa will comprise one-fifth of the global labour force. 4

Descriptive studies that do exist in the African context, highlight significant deficits in digital competencies, often tied to infrastructural and institutional barriers. A systematic review by Ndibalema (2025), covering 14 studies across sub-Saharan Africa, attributes the gap to poor infrastructure, outdated curricula, and unprepared lecturers 5. Many universities still lack reliable access to computers, internet, and digital learning environments. Other research stresses the importance of aligning technical education with labour market needs, advocating for more hands-on ICT training, mentorship, and work-integrated learning opportunities to close the gap between academic training and workplace expectations 6.

Despite these contributions, key limitations remain. Much of the existing computing skill-gap research is descriptive, offering limited insight into **why** the skills gap persists or how specific factors contribute to it. There is a growing recognition of the need for deeper analytical approaches that go beyond correlations. Scholars have increasingly advocated for the use of causal inference and Bayesian methods to model complex relationships and quantify socio-demographic, infrastructural, and psychological effects on skill development 7. However, such methods remain rare in studies focused on African populations, representing a critical gap in the literature.

In summary, while research globally affirms the urgency of addressing computing skill gaps, there remains a scarcity of rigorous, data-driven investigations in contexts. Particularly lacking are studies that apply advanced statistical techniques to unpack the structural and behavioural drivers of the gap. This project aims to help fill that void by leveraging predictive, Bayesian, and causal inference methods to produce actionable insights grounded in empirical data.

# Chapter 3: Design and Implementation

This chapter presents the design rationale, system architecture, and implementation of predictive, Bayesian, and causal models used to characterise and forecast computing‑skill acquisition within the CSA Africa learner population. It begins with a high‑level overview of the end‑to‑end modelling pipeline that underpins all subsequent approaches, then specifies the software libraries, dependencies, and computational environment. Following the pipeline structure, we describe the dataset in detail, including the construction and processing of the composite programming skill score before delving into the architectures and implementation of the models classical machine‑learning baselines, Bayesian regression and causal inference models.

## 3.1 Overview of the Pipeline

The modelling pipeline developed for this research provides a structured and transparent framework that integrates all analytical components of the project. Its design is modular, ensuring consistency across the different methodological paradigms, traditional predictive models, Bayesian regression, and causal inference, while maintaining a clear progression from raw data to actionable insights. Each stage of the pipeline is deliberately formulated to safeguard data integrity, enhance model interpretability, and strengthen the robustness of results. The sequential stages are as follows:

1. **Data Ingestion and Cleaning:**  
   Raw data collected from a structured survey of CSA Africa learners is ingested into structured data frames. Data is then assessed for missingness, inconsistencies, and outliers, followed by targeted cleaning and filtering to ensure analytical reliability.
2. **Feature Engineering and Variable Construction:**  
   Numerical features are scaled to a common range or z-standardized, while categorical variables are encoded using one-hot representations. This step ensures consistency across heterogeneous data types and prepares the dataset for reliable modelling.
3. **Composite Skill Score Development:**  
   Core predictors i.e. self-rated competence, engagement frequency, training exposure, training quality, and confidence are aggregated into a unified, normalized composite programming skill score. This measure provides a continuous and interpretable representation of computing competence and serves as the principal outcome variable for subsequent analyses.
4. **Data Splitting:**  
   The processed dataset is partitioned into training and testing subsets to enable rigorous evaluation of model performance and guard against overfitting.
5. **Predictive Modelling:**  
   Initial predictive benchmarking is performed using established machine learning techniques including Linear Regression, Random Forest, and XGBoost. These models provide initial benchmarks and highlight the associative structure within the data.
6. **Bayesian Regression Modelling:**  
   Probabilistic models are implemented using PyMC, enabling the incorporation of prior knowledge, quantification of parameter uncertainty, and generation of full posterior distributions for detailed parameter inference.
7. **Causal Inference Analysis:**  
   Causal inference techniques are applied to investigate the directional effects of key factors on computing skill acquisition. By employing methods such as propensity score estimation and backdoor adjustment, the analysis moves beyond correlation towards understanding causal mechanisms which are relevant to policy interventions.
8. **Model Diagnostics and Evaluation:**  
   Each modelling paradigm is subjected to rigorous diagnostic checks. These include conventional performance metrics (e.g., R2, RMSE), Bayesian convergence statistics (i.e., R-hat, effective sample size), and robustness tests for causal claims (i.e., placebo and sensitivity analyses).
9. **Interpretation and Visualization:**  
   Final interpretation of results is facilitated through visualization of model outputs, including feature importance rankings, posterior distributions with credible intervals, and causal effect diagrams, to ensure clarity and practical applicability of results.

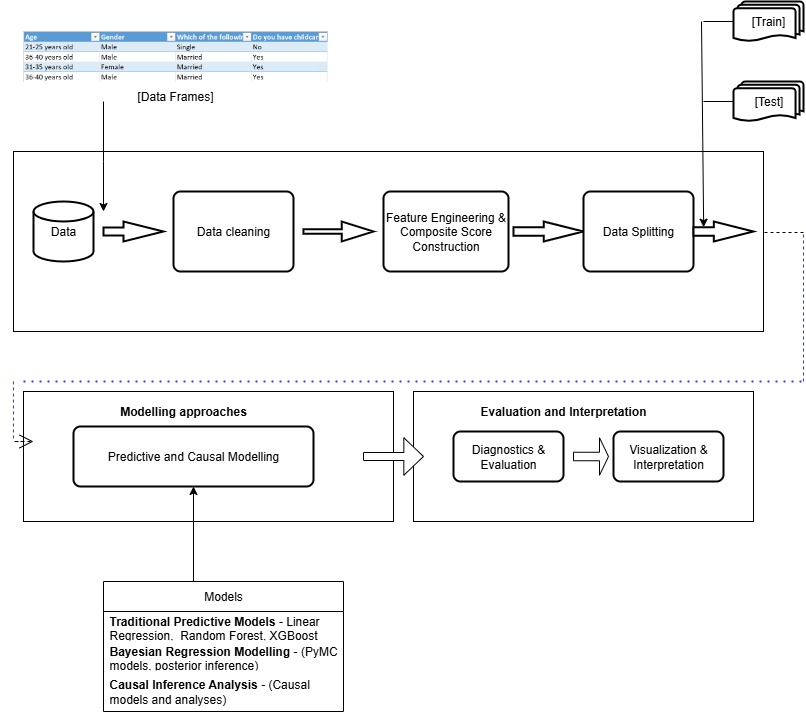


Figure 1: Modelling Pipeline

## 3.2 Libraries and Development Environment

This section provides details on the programming languages, software libraries, hardware and computational environment employed throughout the modelling and analysis processes. The modelling and analysis were implemented in Python within a dedicated conda environment (bayes-env), ensuring full reproducibility of results. All software dependencies were explicitly managed and configured with the necessary compiler toolchain (g++) and numerical libraries (i.e, BLAS) to support efficient Bayesian inference routines.

Core libraries and dependencies included:

**Software**

* **Data Handling and EDA:** Pandas, NumPy
* **Visualization:** Matplotlib, Seaborn
* **Pre-processing and Scaling:** scikit-learn.preprocessing (StandardScaler, MinMaxScaler)
* **Predictive Modelling:** scikit-learn (Linear Regression, Random Forest) and XGBoost (gradient boosting regression)
* **Bayesian Modelling:** PyMC (MCMC sampling), ArviZ
* **Causal Inference:** DoWhy, Causal Graphical Models

## Computational environment

## All experiments were executed on a local workstation equipped with a modern multi-core CPU. This hardware configuration allowed efficient parallelisation of Markov Chain Monte Carlo (MCMC) sampling and scalable training of ensemble learning models.

## 3.3 Dataset Description

This study is grounded in a rich dataset collected from the **Computer Science Academy (CSA) Africa** market research program, which engaged over **2,500 learners** across **15 African countries**. The dataset comprises structured self-reported responses, training feedback, and contextual indicators aimed at understanding the challenges and enablers of computing skill acquisition in diverse learning environments. The heterogeneity across countries including Nigeria, Kenya, Ghana, Malawi, Rwanda, Eswatini, South Sudan, and others provides valuable variation for studying computing skills development across different educational and infrastructural contexts.

### 3.3.1 Dataset Structure and Sources

The dataset originates from CSA Africa’s learner market research. It captures a wide spectrum of variables grouped into the following thematic areas:

* **Demographic Attributes**: age, gender, education level, and country of residence.
* **Self-reported Programming Indicators**: confidence in coding, self-rated competence, training participation, perceived quality of training received.
* **Engagement Metrics**: frequency of programming practice, number of projects completed, and weekly coding hours.
* **Infrastructure and Access**: access to internet, device availability, and infrastructural barriers such as electricity instability
* **Training Exposure**: course completion status, mode of delivery (online or in-person), and access to mentoring or community support

To support modelling, the data was structured into two key structured frames:

* **df\_model**: a focused dataset restricted to the five predictors used to construct the composite programming skill score: *competence\_score*, *engagement\_score*, *training\_score*, *training\_quality\_score*, and *confidence\_score*. This subset was primarily used to validate the measurement model and to implement initial Bayesian regressions.
* **df\_final**: A comprehensive dataset that includes the normalized composite skill score along with all remaining encoded predictors (37 features) not directly used in score construction. This frame enables broader exploratory analysis and supports both predictive and causal inference modelling phases.

### 3.3.2 Data Pre-processing and Cleaning

A structured pre-processing pipeline was applied to prepare the data for modelling:

* **Missing Data Handling**: Observations missing values in critical predictor or outcome fields were removed. Secondary variables were imputed using mean (for numerical) or mode (for categorical), where appropriate.
* **Standardization**: To ensure numerical stability in both traditional and Bayesian models, all predictors in df\_model were standardized using StandardScaler.
* **Score Normalization**: For each of the five core predictors, a MinMaxScaler transformation was applied. These normalized variables were then summed to create the normalized\_composite\_skill\_score, which served as the target variable.
* **Categorical Encoding**: Categorical variables in df\_final (e.g., country, gender, training source) were encoded using label encoding or one-hot encoding to prepare for use in machine learning models.

### 3.3.3 Exploratory Data Analysis (EDA)

EDA was conducted to identify initial patterns, detect anomalies, and validate feature relationships:

* **Correlations**: Moderate to strong correlations were observed among key variables:
  + Engagement and confidence (r ≈ 0.48)
  + Training quality and competence (r ≈ 0.51)

[Insert diagram]

* **Univariate Visualizations**: Histograms, bar plots and scatterplots were used to explore feature distributions by demographic groups and learning contexts.

[Insert diagram]

These insights informed the subsequent feature selection process, choice of priors for Bayesian models, and the construction of subgroup models for causal inference.

## 3.4 Composite Skill Score Construction

A central component of this study is the creation of a **composite programming skill score** that serves as the primary target variable in all predictive and inferential models. Given the absence of a single objective indicator of coding proficiency in the dataset, a derived metric was constructed to represent learners’ overall programming capability in a consistent and scalable way.

### 3.4.1 Motivation and Rationale

Programming ability is a multifaceted construct, particularly in the context of CSA Africa learners, who vary widely in exposure, access, confidence, and support. Rather than rely on one isolated measure, the composite score synthesizes several complementary indicators to provide a more holistic reflection of a learner’s skill trajectory. This approach aligns with best practices in educational measurement, where latent constructs are often represented using multiple aligned variables.

### 3.4.2 Core Components

The composite score was constructed from five key self-reported and behavioural indicators:

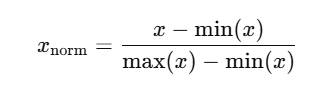
1. **Competence Score** – learner’s self-rated programming ability
2. **Engagement Score** – frequency of coding activity or practice
3. **Training Score** – exposure to programming courses or learning interventions
4. **Training Quality Score** – perceived quality or usefulness of the training received
5. **Confidence Score** – learner’s confidence in their problem-solving or technical skills

These components were selected based on domain relevance, internal consistency, and their collective ability to capture both **skill acquisition** and **readiness to apply skills**.

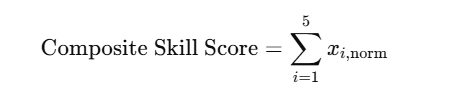
### 3.4.3 Normalization and Aggregation

To ensure comparability across variables measured on different scales, each of the five input features was normalized using **MinMax scaling**, transforming all values into the [0,1] range. This prevents any single variable from dominating the composite score due to scale differences.

The formula for the normalized version of each feature xx is:



Once normalized, the features were aggregated using a simple **unweighted sum**:



This produced a continuous score ranging from 0 to 5, with higher values representing stronger programming proficiency across multiple dimensions.

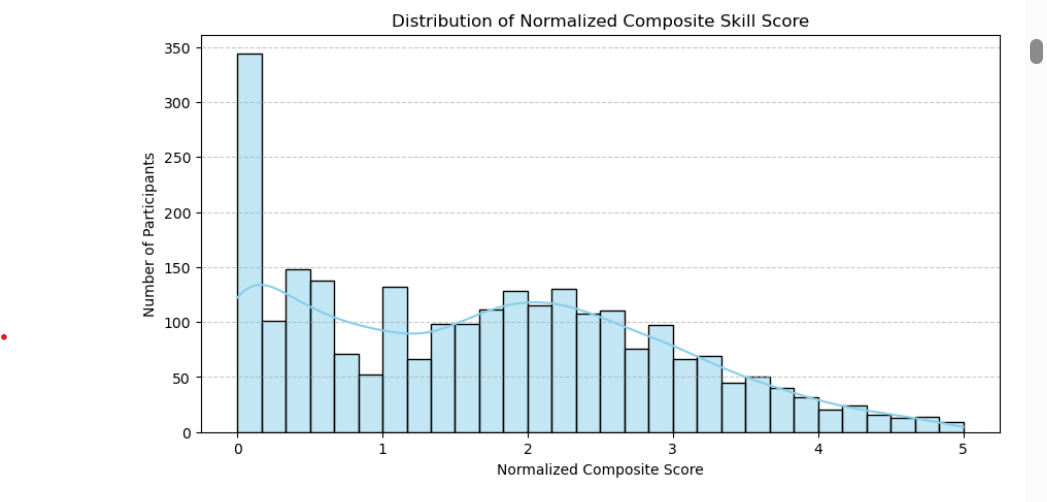


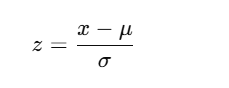
Figure 2: Distribution of Normalized Composite Skill Score

### 3.4.4 Justification for Equal Weights

Equal weighting was chosen for simplicity and interpretability, and to avoid introducing arbitrary bias in the absence of strong empirical evidence or expert priors regarding the relative importance of each factor. Later we use Bayesian and causal models to analyse the weighting of these features on the computing skill score and adjusting for relative influence statistically.

### 3.4.5 Post-processing for Modelling

For stability in regression modelling, especially under the Bayesian framework, the composite skill score was further standardized using z-score normalization:



This transformation ensured numerical stability during MCMC sampling and made coefficients more interpretable across models.

## 3.5 Traditional Predictive Modelling

Traditional machine learning models were implemented primarily to serve as baseline predictors for computing skill scores. These models helped establish reference performance levels before applying more advanced Bayesian and causal modelling techniques. The following section outlines each model's rationale, key observations, and performance metrics, with a focus on understanding their limitations and strengths in the context of the CSA Africa dataset.

### 3.5.1 Linear Regression: Baseline Modelling and Interpretability

Linear regression was selected as the initial model due to its simplicity, transparency, and ease of interpretation. This model estimates the linear relationship between each input feature and the target skill score. Its strengths lie in its transparency and its ability to provide clear coefficient estimates that indicate the direction and magnitude of each predictor’s influence.

**Performance Summary:**

* **RMSE:** ~0.87
* **R² Score:** 0.49

This performance indicates moderate predictive accuracy and shows that the model captures around 49% of the variance in skill scores. Key predictors with strong positive influence included:

* Country of residence (e.g., South Sudan, Malawi, Eswatini)
* Career aspirations
* Psychological perceptions and confidence
* Education level

**Limitations:** Despite its interpretability, the linear model fails to account for:

* Nonlinear interactions (e.g., infrastructure × mindset)
* Complex dependencies among features
* Saturation or threshold effects common in social data

These limitations motivated a shift toward more flexible and powerful models.

[Insert diagram]

### 3.5.2 Random Forest Regression: Capturing Nonlinearity and Feature Interactions

To address the above limitations, a Random Forest Regressor was implemented**.** This tree-based method is well-suited for modelling non-linear relationships and automatically accounts for feature interactions without the need for manual transformation.

**Performance Summary:**

* **RMSE:** 0.2839
* **R² Score:** 0.947

The model demonstrated a significant improvement over linear regression, explaining approximately 95% of the variance. Influential features identified by the Random Forest model included:

* Barriers to applying programming skills
* Learner perceptions and awareness
* Support systems (e.g., family or community)
* Electricity access and gender
* Motivation and interest levels

The ability of Random Forests to uncover such multidimensional patterns aligns well with CSA Africa’s objective of identifying structural and social barriers to digital skill development.

[Insert diagram]

### 3.5.3 XGBoost Regression: Performance Optimization and Regularization

XGBoost (Extreme Gradient Boosting) was subsequently used to optimize prediction accuracy and model interpretability. This gradient boosting framework is known for its speed, robustness, and effectiveness in handling feature interactions, missing values, and overfitting through regularization.

**Performance Summary:**

* **RMSE:** 0.2804
* **R² Score:** 0.9485

XGBoost delivered the highest performance among the tested models. Its results closely mirrored those of the Random Forest model while offering clearer differentiation in feature importance rankings. Prominent predictors included:

* Barriers to skill application
* Programming awareness and confidence
* Learning community support
* Country of residence and interest levels

The strong performance and enhanced interpretability made XGBoost the preferred traditional model prior to transitioning to Bayesian and causal frameworks.

[Insert diagram]

## 3.6 Bayesian Regression Modelling

### 3.6.1 Motivation and Context

Bayesian regression modelling was employed to improve interpretability, incorporate uncertainty, and enable probabilistic reasoning about the factors influencing programming skill acquisition and their relations. This framework is particularly valuable when attempting to move beyond point estimates provided by the frequentist approaches and instead draw inferences from full posterior distributions. Unlike traditional models that produce single-value coefficients, Bayesian models express uncertainty about parameters, offering richer insights into the relationships between predictors and outcomes.

The application of Bayesian regression was motivated by several advantages:

* **Uncertainty quantification:** Parameter estimates are represented as distributions, offering insight into the confidence of predictions.
* **Interpretability:** Posterior summaries (e.g., means, credible intervals) help assess the relative importance and direction of each feature.
* **Robustness to noise and overfitting:** Through prior specification, the model can regularize learning and adapt more conservatively to sparse or noisy signals.
* **Flexibility in model structure:** Alternative likelihoods (e.g., Student-t) and prior distributions make Bayesian methods adaptable to real-world data challenges.

This approach was particularly suitable for CSA Africa’s objectives, where capturing both average effects and variability across learner populations is critical for policy relevance.

### 3.6.2 Model Specification and Sampling (Base Model)

Bayesian models were implemented using the **PyMC** probabilistic programming library. The modelling process began with a baseline model using the five core predictors that constituted the composite skill score:

* competence\_score
* engagement\_score
* training\_score
* training\_quality\_score
* confidence\_score

Each predictor was standardized for numerical stability, and the target variable (normalized\_composite\_skill\_score) was z-scored. The core components of the model included:

Each model followed a consistent structure:

* **Likelihood:** A standard Gaussian likelihood was assumed with a continuous outcome (normalized composite skill score).
* **Priors:**
  + beta0 (intercept): Normal (0, 1)
  + beta (coefficient for the predictor): Normal (0, 1)
  + sigma (noise term): HalfNormal(1)
* **Sampling:** Models were fit using the No-U-Turn Sampler (NUTS) with 1,000 tuning steps and 1,000 posterior draws. All parameters were standardized to improve numerical stability. Convergence was assessed using trace plots and diagnostic statistics such as R-hat and Effective Sample Size (ESS).



**Figure 3**: Model Specification

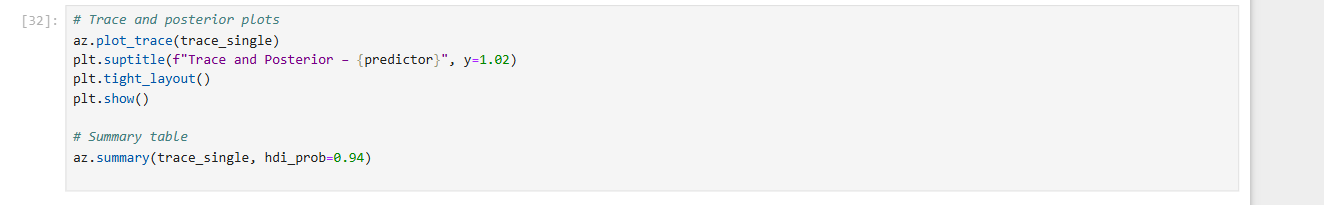
### 3.6.3 Diagnostics and Convergence Analysis

Robust diagnostics were conducted to validate the reliability of the posterior estimates:

* **Trace plots:** Used to visually assess mixing and convergence of chains.
* **Posterior distributions and HDIs:** Provided insight into effect sizes and uncertainty.
* **R-hat statistics:** Values close to 1.0 indicated satisfactory convergence.
* **Effective Sample Size (ESS):** Ensured sufficient sampling precision.

Initial runs revealed divergences and low ESS in some parameters, particularly for sigma. These were mitigated by:

* Increasing target\_accept
* Re-standardizing inputs and outputs to reduce gradient instability
* Running simplified single-predictor models for debugging



**Figure 4:** Trace and Posterior Plots specifications

## 3.7 Causal Inference – [To be written after implementation]

**[BELOW TO BE MOVED TO EVALUATION AND RESULTS CHAPTER]**

**3.6.4 Posterior Interpretation**

Posterior summaries indicated the relative strength and direction of influence for each predictor as listed below:

**Model: competence\_score**

* **Mean beta:** 0.922 (suggests a strong positive effect)
* **R-hat:** 1.00 (indicates convergence)
* **Trace plots:** Show good mixing across chains
* **Interpretation:** Higher self-rated competence is strongly associated with a higher programming skill score. The posterior is sharply peaked, indicating confidence in the estimate.

**Model: engagement\_score**

* **Mean beta:** 0.909
* **R-hat:** 1.00
* **Interpretation:** Frequency of coding engagement is a strong predictor of programming proficiency. Trace plots reflect well-behaved chains, and posterior distributions confirm a tightly concentrated estimate.

**Model: training\_score**

* **Mean beta:** 0.847
* **Interpretation:** Exposure to training is positively linked to skill outcomes, though slightly less than competence and engagement. Still, posterior uncertainty is low and sampling diagnostics are satisfactory.

**Model: training\_quality\_score**

* **Mean beta:** 0.858
* **Interpretation:** The quality of training attended by learners shows a consistent and strong association with skill acquisition, reinforcing the importance of training design and delivery.

**Model: confidence\_score**

* **Mean beta:** 0.792
* **Interpretation:** Learners’ confidence in problem-solving has a moderately strong effect. The slightly wider posterior suggests more variability in how confidence translates into actual skill.

**3.6.4 Summary of Bayesian Base Models**

Each of the five Bayesian base models demonstrated clean convergence, with effective sample sizes exceeding 1,000 and R-hat values at or near 1.00. Posterior means of the coefficients indicate consistently positive contributions from all predictors. These results validate the structure of the composite skill score and affirm the relevance of each component.

The decision to model each predictor independently provided clarity and robust diagnostics, and forms the foundation for more complex multivariate or hierarchical Bayesian models explored in the next phase.

# Chapter 4: Conclusion

Main conclusions of your project. Here you should also include suggestions for future work.

# Appendix A: First appendix

**A.1 Section of first appendix**

# Appendix B: Second appendix

**Bibliography**

[1] <https://scikitlearn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>

[2]  <https://scikit-learn.org/stable/modules/linear_model.html#linear-regression>

[3] *Braun, G., Rikala, P., Järvinen, M. et al (2024).* Bridging Skill Gaps - A Systematic Literature Review of Strategies for Industry.

[4] *Bhorat, H., et al. (2023).* Digitalization and digital skills gaps in Africa. Brookings Institution.

[5] *Ndibalema, P. (2025).* A Systematic Review Digital Literacy Gaps in Promoting 21 Century Skills in Sub-Saharan Africa.

[6] *Cloete, A., & De Villiers, C. (2021).* Youth Unemployment in South Africa

[7]  *Li, F., Ding, P., & Mealli, F. (2023).* Bayesian causal inference: a critical review. Philosophical Transactions of the Royal Society A, 381(20220153). https://doi.org/10.1098/rsta.2022.0153