



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

Write me a caption for this video I am posting on linkedin. It is a csa participant. Here is his transcript. – also mention he is a problem solver a quality we value in csa learners. Mention he noticed a problem of food delivery during the workshop and implemented an app to solve this problem. Mention the values associated with this and say more of his story to come.

**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** **10**

# 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?

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

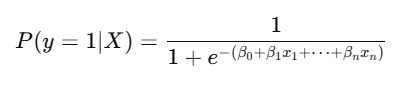
The key highlights achievements for this project are as follows:

# Chapter 2: Survey

## 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.

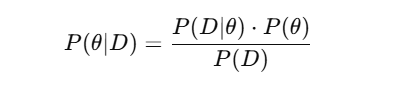
1 https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LogisticRegression.html

2 https://scikit-learn.org/stable/modules/linear\_model.html#linear-regression

## 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[1](#user-content-fn-4). 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[2](#user-content-fn-5). 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.1. 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. 2

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 3. 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 4.

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 called for the use of **causal inference** and **Bayesian methods** to model complex relationships and quantify the effects of socio-demographic, infrastructural, and psychological variables on computing skills development. However, such methods are rarely applied 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.

1 *Rikala, J., Skog, M., & Holmström, J. (2024).* A systematic review of the skill gap concept in digitalization research*.* Chalmers University of Technology.

2 *Bhorat, H., et al. (2023).* Digitalization and the Labor Market in Africa. Brookings Institution.

3 *Ndibalema, P. (2025).* A Systematic Review of ICT Competency in African Higher Education.

4 *Cloete, A., & De Villiers, C. (2021).* Youth Unemployment and Digital Skills in South Africa

# Chapter 3: Further chapters

The content of these chapters depends on the project and should be agreed with your super- visor (e.g. description of the solution, evaluation results, etc).

# 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] C. Baier and J.-P. Katoen. *Principles of Model Checking*. MIT Press, 2008.