



UGANDA CHRISTIAN
UNIVERSITY

A Centre of Excellence in the Heart of Africa

Department of Computing & Technology

Machine Learning Decomposition of Gender and Rural–Urban Wage
Gaps: Evidence from Uganda National Household Surveys

By

Bob Robert Tumushiime
Bachelor of Science in Computer Science

Reg No: J25M19/009, Student No: B31332
E-mail: b31332@students.ucu.ac.ug, Phone No:
0701007118/0778865111

A dissertation submitted to the Faculty of Engineering Design &
Technology for the study leading to the Degree in partial fulfillment
of the requirements for the award of **Master of Science in Data
Science and Analytics** of Uganda Christian University

September 2026

Abstract

UN Women puts Uganda’s monthly gender wage gap near 32 percent, with rural–urban differences similarly wide, yet the two available decomposition studies both analysed a single survey round from 2002/03 using a linear wage equation that assumes constant rates of return.

Both gaps were decomposed into explained and unexplained portions. Four Uganda National Household Survey rounds (2012/13–2023/24) were harmonised into a pooled file of 14,963 workers in formal paid employment aged 14 to 64. Ordinary least squares (OLS), LASSO, Random Forest, and XGBoost were estimated on a stratified 70/15/15 split, and SHapley Additive ex-Planations attributed predictions to characteristics. The best model’s predictions then replaced the linear fitted values in the Oaxaca–Blinder decomposition.

XGBoost predicted most accurately, with a test R^2 of 0.410 against 0.399 for Random Forest and 0.354 for both linear models; pooling all four rounds raised it from 0.286. Education was the leading predictor. The rural–urban gap was largely explained under both methods, 49 to 59 percent under OLS and 68 to 84 percent under the machine learning method. For the gender gap, OLS returned a negative explained share in three of four rounds, whereas the machine learning method attributed 14 to 57 percent to measured characteristics.

How much of the gender gap counts as explained therefore depends on the functional form permitted to the model, which bears on how existing single-round linear estimates should be read. The unexplained component is reported as a residual and is not a measure of discrimination.

Declaration

I, *Bob Robert Tumushiime*, declare to the best of my knowledge that the work in this dissertation is original and has not been submitted before to any university/institution. It is the result of my effort.

Signature:  _____

Date: September 14, 2026

Ethical Generative AI Usage Declaration

Students may use generative artificial intelligence (AI) ethically as a “study buddy” or assistant rather than a replacement for critical thinking. This includes citing AI use and fact-checking all outputs to maintain academic integrity.

For purposes of this assignment, the following shall guidelines shall apply.

- i. Generative AI may be used for brainstorming, and generating ideas to improve your work. You may also use AI for literature search, and summarizing journal papers for quick analysis and assessing relevance of publications. However, once the relevant publications have been identified, a student is expected to fully review the selected publications, identifying problems addressed, methods used, gaps identified and recommendations made.
- ii. Where AI has been used as guided in (i) above, full disclosure of how it has been used should be made in the appendix B, including how AI was used, the AI tools (including mode versions), and a compilation on major AI prompts used.
- iii. NO GENERATIVE AI CONTENT IS ALLOWED IN THE FINAL SUBMISSION. The final submission SHOULD be your original work.

A. Declaring the non-use of AI for content generation

I, *Bob Robert Tumushiime*, declare that no content generated by AI technologies has been used in this dissertation.

Signature:  Date: 14th September 2026

B. Acknowledging the use of AI as a study buddy

Please refer to Appendix B for the complete declaration of AI usage for generating ideas, brainstorming, literature search, and summarizing only to improve your work.

Approval

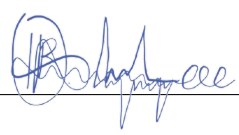
This dissertation has been submitted for examination with approval of the following advisors/supervisor.

Advisor: Dr Samuel Kakuba
Department of Computing & Technology
Faculty of Engineering Design & Technology
Uganda Christian University

Signature:  _____

Date:  _____

Co-Advisor: Eng. Ian Raymond Osolo
Department of Computing & Technology
Faculty of Engineering Design & Technology
Uganda Christian University

Signature:  _____

Date: 16 / 09 / 2026 _____

Acknowledgments

I wish to express my sincere gratitude to my supervisors, Dr. Samuel Kakuba, Eng. Ian Raymond Osolo, Dr. Daphine Bitalo and Dr. Ahile, for their invaluable guidance and constructive feedback throughout the thesis period.

I acknowledge UBOS for providing the datasets upon which this study relies.

I am also grateful to my friends, Aguta Danstan, Lordin Lutalo, Richard Wambede, Jude Mugimba and Isaac Omoding for their support during the thesis writing.

Finally, I thank my family for their unwavering support and encouragement throughout this journey.

Contents

Abstract	i
Declaration	ii
Ethical Generative AI Usage Declaration	iii
Approval	iv
Acknowledgments	v
List of Abbreviations and Acronyms	ix
1 Introduction	1
1.1 Background	1
1.2 Key Definitions and Disciplinary Positioning	1
1.2.1 How This Study Differed from a Social Economics Study	2
1.3 Problem Statement	4
1.4 Research Objectives	5
1.4.1 General Objective	5
1.4.2 Specific Objectives	5
1.5 Research Questions	5
1.6 Scope of the Study	5
1.7 Justification of the Study	6
1.8 Significance of the Study	6
1.9 Cultural and Policy Context	7
1.9.1 Traditional and Cultural Dimensions of Employment	7
1.9.2 Policy Landscape for Gender and Labour in Uganda	7
1.10 Theoretical Foundation	8
1.11 Definition of Key Terms	8
2 Literature Review	10
2.1 Introduction	10
2.2 Theoretical Foundations of Wage Determination	10
2.3 The Evolution of Wage Gap Decomposition, 1973–2024	11
2.3.1 The Classical Decomposition and Its Immediate Problems	12
2.3.2 Moving Beyond the Mean	13
2.3.3 Relaxing the Functional-Form Assumption	13
2.3.4 Recovering Interpretability from Opaque Models	14
2.4 Empirical Evidence on Wage Inequality in Uganda and Sub-Saharan Africa	14
2.4.1 Regional Evidence	14

2.4.2	Uganda-Specific Evidence	15
2.4.3	How Many Survey Rounds Have Existing Studies Used?	15
2.5	The UNHS as a Data Source for Multi-Round Wage Analysis	16
2.6	Harmonisation Practice in Multi-Round Survey Programmes	17
2.7	Explainability in Predictive Wage Models	18
2.8	Synthesis and Research Gaps	18
2.8.1	How the Reviewed Literature Provides the Rationale for This Study . . .	19
3	Methodology	21
3.1	Introduction	21
3.2	Research Design	21
3.2.1	Philosophical Stance: Pragmatism	22
3.3	Data Sources	22
3.3.1	Data Used	22
3.3.1.1	Uganda National Household Survey	22
3.3.1.2	Structure of the Data: Repeated Cross-Sections, Not a Panel .	22
3.3.2	Study Population and Sampling	23
3.3.3	Subgroup Sample Sizes and Data Sufficiency	23
3.3.4	Under-Representation of Women, and How It Was Handled	24
3.3.5	Variable Definitions	25
3.3.6	Data Integration and Cleaning	26
3.4	Model Development and Explainability	27
3.4.1	How Explainability Was Implemented	28
3.5	Wage Gap Decomposition	29
3.6	Quality Control and Error Management	30
3.7	Ethical Considerations	30
3.8	Methodological Constraints	30
3.9	Work Timeline	31
4	Results	32
4.1	Introduction	32
4.2	Results	32
4.2.1	Data Preprocessing Outcomes	32
4.2.2	Model Performance and Feature Importance	34
4.2.3	Wage Gap Decomposition	38
4.2.4	Subgroup Results Across Estimation Techniques	42
5	Discussion	44
5.1	Introduction	44
5.2	Multi-Round Coverage, Subject to a Cross-Sectional Limit	44
5.3	What Predicted Wages Best, and Why	45
5.3.1	Nonlinearity, Not Algorithm Choice, Drove the Predictive Gain	45
5.3.2	Why Sex and Residence Ranked Low in the SHAP Results	46
5.3.3	Why XGBoost Rather Than Random Forest	46
5.4	Divergent Findings for the Rural/Urban and Gender Gaps	47
5.5	Transferability, Subject to a Data-Sparsity Cost	49
5.6	Chapter Summary	50
6	Conclusion and Recommendations	51
6.1	Introduction	51
6.2	Summary of Findings	51

6.3	Conclusions	52
6.4	Limitations	53
6.5	Recommendations	54
6.5.1	For Policy and Practice	54
6.5.2	For Survey Practice at UBOS	54
6.6	Future Work	54
6.6.1	Directions for Further Research	54
6.6.2	Further Applications of the Framework	55
6.7	Contribution to Knowledge	55
REFERENCES		59
APPENDICES		59
A Research Project Timelines		60
B Declaration of AI Use		63
C List of 10 Major Journal Publications Reviewed		64
D Research Alignment with Programme, National, and Global Agenda		66
D.1	Alignment with MSc Data Science and Analytics Programme	66
D.2	Alignment with National Development Agenda	66
D.2.1	National Development Plan IV (NDP IV)	66
D.2.2	Uganda Vision 2040	67
D.3	Alignment with Global Development Agenda	67
D.3.1	Sustainable Development Goals (SDGs)	67
D.3.2	African Union Agenda 2063	67
E Research Poster Layout		69
F Research Ethics Committee (REC) Approval		70
G Data Collection Tools		72
H Student–Supervisor Meeting/Supervision Plan		73

List of Abbreviations and Acronyms

Acronym	Meaning
EA	Enumeration Area
EOC	Equal Opportunities Commission
ETL	Extract, Transform, Load
ICSE-18	International Classification of Status in Employment (2018)
ILO	International Labour Organization
ISCO	International Standard Classification of Occupations
ISIC	International Standard Industrial Classification of All Economic Activities
LASSO	Least Absolute Shrinkage and Selection Operator
MAE	Mean Absolute Error
ML	Machine Learning
NDP III	Third National Development Plan (2020/21–2024/25)
NLFS	National Labour Force Survey
OLS	Ordinary Least Squares
RF	Random Forest
RMSE	Root Mean Squared Error
SHAP	SHapley Additive exPlanations
UBOS	Uganda Bureau of Statistics
UGX	Uganda Shillings
UN Women	United Nations Entity for Gender Equality and the Empowerment of Women
UNHS	Uganda National Household Survey
UBOS	Uganda Bureau of Statistics
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting

Chapter 1

Introduction

1.1 Background

Labour markets in low-income countries are segmented along gender and geographic lines, and workers with comparable skills and experience often earn markedly different wages depending on who they are and where they work. How large those differences are, what sustains them, and whether they are closing matters for the design of labour market policy [1, 2, 3].

In Sub-Saharan Africa, wage differences are widened further by rapid structural transformation, uneven sectoral development, and occupational segregation by gender [4, 5]. Uganda illustrates the pattern well. The UNHS 2023/24 reports a national median monthly in-cash earning of UGX 200,000 for persons in paid employment, yet that single figure hides considerable variation: mean monthly consumption expenditure per adult stands at UGX 302,179 in Kampala against UGX 61,007 in Karamoja, a ratio of nearly five to one [6]. Women make up 44.2 percent of subsistence agriculture workers, which helps explain the UN Women estimate that women in Uganda earn about 32 percent less than men each month [7]. Rural poverty stands at 19.4 percent, more than three times the urban rate of 5.7 percent [6].

The standard tool for studying such differences is the Oaxaca–Blinder decomposition applied to a linear OLS wage equation. It measures the size of a gap and splits it into explained and unexplained parts, but it assumes that returns to schooling, experience, and sector are linear and additive, which is a strong assumption in a market as segmented as Uganda’s. Recent methodological work shows that ML models provide a more flexible basis for decomposition, and that explainability tools such as SHAP keep the results readable enough to be useful for policy [8, 9, 10, 11].

Two boundaries shaped the design of this study. The analysis covered workers in formal paid employment, meaning individuals who reported a wage or salary from an employer, since the Oaxaca–Blinder decomposition needs a well-defined continuous dependent variable. The wage data were also self-reported, and therefore carry the recall error and rounding that the measurement literature documents [12, 13].

1.2 Key Definitions and Disciplinary Positioning

Several of the concepts used in this study carry different meanings depending on the discipline in which they are used. Because the study sits within a Data Science programme while addressing a problem that belongs traditionally to social economics, the distinctions set out below are worth making at the outset.

Gender. In the social sciences, gender denotes the socially constructed roles, norms, and power relations that shape the opportunities available to men and women, spanning education access, household decision-making, land ownership, political participation, and labour force participation [14, 15, 16, 1, 2]. Work in this tradition normally covers the whole population, irrespective of age or employment status, and asks why women participate less in the labour force, why they are concentrated in unpaid care work, and how cultural expectations constrain their economic agency [15, 17].

However, this study used a narrower and more operational definition. The analysis was restricted to individuals aged 14–64 who reported monthly cash earnings from paid work. Within that subset, gender was operationalised as the binary sex variable recorded in the UNHS household roster. The study therefore did not examine gender dynamics in the general population; it examined wage differences between men and women who were already in formal paid work. Two consequences follow from restricting the sample this way:

1. **Selection into employment.** The sample excluded people who were unemployed, in subsistence agriculture, in unpaid domestic work, or outside the 14–64 age range. Since women are over-represented in the excluded categories, the gender wage gap reported here is the gap conditional on formal employment rather than the gap in total economic outcomes between all men and all women in Uganda. Gender-based economic inequality in the country is almost certainly wider than the figures reported in this study [17].
2. **Proxy, not construct.** The binary sex variable is a measurable stand-in for the social and structural differences that gender produces. Norms, attitudes, and power relations were not measured. What the Oaxaca–Blinder decomposition captured was whether men and women holding the same observable characteristics earned different wages. Those characteristics were education, occupation, sector, region, and experience. Where a difference remained, the decomposition established how much of it was attributable to differences in the characteristics themselves rather than to differences in the returns paid on them.

The study therefore treated gender as a wage predictor within the formal labour market for ages 14–64, and not as a broad social phenomenon across the whole population.

Socioeconomic factors refer to the observable characteristics that economic theory expects to influence earnings. Some of these are education level, years of experience, occupation, industry sector, geographic location, and household composition. They are the features that make up the Mincer human capital earnings function and its extensions [18]. Within the Oaxaca–Blinder framework, these factors form the explained component of a wage gap, that is the portion attributable to measurable differences in human capital and job characteristics between the groups being compared.

1.2.1 How This Study Differed from a Social Economics Study

A social economics study of Uganda’s wage gap would ask whether employer discrimination causes women to earn less than equally qualified men, and answering that question causally would require either a randomised audit study or a credible natural experiment. The question asked here was different: which observable features best predicted wages, how did the composition of the wage gap change when nonlinear ML models replaced linear OLS, and how had that composition shifted over twelve years? The contribution is methodological and computational, in that it builds a better measurement instrument, and it is not a contribution to social theory. The study neither advanced new sociological theory about gender nor claimed to identify the root causes of discrimination. What it produced instead is a more flexible, interpretable,

and reproducible quantitative framework that policymakers and researchers can use to decide where further causal investigation would be most worthwhile. Table 1.1 sets out the distinctions between the two approaches in more detail.

Table 1.1: Social economics vs. data science approach to wage gap analysis

Dimension	Social Economics Approach	Data Science Approach (This Study)
Core question	Why do wage gaps exist, and what social, institutional, and behavioural mechanisms cause them?	What observable features best predict and decompose the gap, and how does the gap’s composition change over time?
Primary method	Causal inference: instrumental variables, natural experiments, regression discontinuity, qualitative case studies	Predictive modelling: supervised ML, namely OLS, Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest (RF), and Extreme Gradient Boosting (XGBoost), with feature attribution via SHAP
Theory usage	Theory drives the model: hypotheses about discrimination, bargaining power, or social norms are tested directly	Theory informs the feature set (Mincer earnings function) and the decomposition structure (Oaxaca–Blinder), while the ML models remain data-driven and assumption-light
Key output	Causal estimates of the effect of specific interventions (for example, education expansion or anti-discrimination law) on wages	Feature importance rankings, nonlinear interaction effects, decomposition shares (explained vs. unexplained), and their temporal evolution
Claim about discrimination	Can estimate the causal effect of discrimination under identification assumptions	Cannot; the unexplained component is a residual that also contains unobserved ability, hours worked, job tenure, measurement error, and model misspecification
Data engineering	Typically uses a single, clean dataset, and harmonisation is not treated as a contribution	Multi-round Extract, Transform, Load (ETL) harmonisation across four surveys with changing variables, coding, and boundaries was itself a core contribution
Scalability	Findings are context-specific, and replication requires new causal designs	The pipeline was designed as a reusable, transferable framework applicable to other countries and survey programmes

Four data science contributions distinguished this work:

1. **Data engineering at scale:** designing and implementing a reproducible ETL pipeline that harmonised four survey rounds with differing variable schemas, classification systems, and geographic boundaries into a single analytical dataset.
2. **Model comparison and selection:** benchmarking four models of increasing complexity, namely OLS, LASSO, RF, and XGBoost, on cross-validated performance metrics, rather

than committing to a single pre-specified model as econometric studies typically do.

3. **Explainable Artificial Intelligence (XAI):** applying SHAP-based feature attribution so that predictions from otherwise opaque models could be interpreted at both the individual and the subgroup level, which connects predictive accuracy to policy-relevant transparency.
4. **ML-augmented decomposition:** substituting ML predictions into the Oaxaca–Blinder framework to capture nonlinear and interactive effects that OLS cannot represent, and tracking how the decomposition evolved across four temporal waves.

1.3 Problem Statement

Wage differences in Uganda are wide and have persisted for at least two decades. Working on UNHS 2002/03, Kagundu and Pavlova [19] reported a female-to-male hourly wage ratio of 0.73 in urban areas and 0.67 in rural areas. Sebaggala [20], using the same round, put the overall male–female gap at about 39 percent. More recently, UN Women [7], using UNHS 2019/20 microdata, placed the monthly gender wage gap at roughly 32 percent and found the penalty largest at the bottom of the wage distribution. On the spatial side, UBOS [6] records mean monthly consumption expenditure per adult of UGX 302,179 in Kampala against UGX 61,007 in Karamoja, and rural poverty of 19.4 percent against an urban rate of 5.7 percent.

What is far less clear is how these figures should be read together. The studies just cited differ in survey round, in sample definition, and in method, so the movement from about 39 percent in 2002/03 to about 32 percent in 2019/20 cannot be attributed to real improvement rather than to those differences. Table 2.2 records the survey coverage of every Ugandan and comparable regional wage gap study located in the review in Chapter Two. Each analyses a single survey round, and none pools two or more. Without one specification applied consistently to several rounds, no statement about direction of travel can be supported from the published evidence.

A second limitation is methodological. Both Ugandan decomposition studies fit a linear OLS wage equation [19, 20]. Evidence from a range of settings indicates that the pay return to schooling rises disproportionately at higher levels rather than at a constant rate per year [21, 22], and Kagundu and Pavlova’s own result, a female-to-male ratio of 1.04 in the urban public sector against 0.55 in the urban private sector [19], points to wage setting that differs sharply between segments of the same market. Where patterns of this kind are present, a linear specification cannot represent them without interaction terms the analyst must know to include in advance, and what it does not represent is carried into the unexplained component of the decomposition. Bonaccolto-Töpfer and Briel [9] showed on German data that replacing OLS fitted values with machine learning predictions moves part of that unexplained component into the explained one, and Strittmatter and Wunsch [8] reported the same direction of effect on Swiss administrative data.

Neither of those demonstrations used data from a low- or middle-income country, and neither covered more than a single point in time. Within the literature reviewed in Chapter Two, no study was located that applies a machine learning decomposition to Ugandan wage data, and none that tracks the explained and unexplained components across several UNHS rounds held in one harmonised file. This study addresses those two absences by applying explainable machine learning and Oaxaca–Blinder decomposition to four rounds of UNHS microdata covering twelve years, from 2012/13 to 2023/24.

1.4 Research Objectives

1.4.1 General Objective

The general objective of this study was to leverage machine learning approaches for modelling, decomposing, and interpreting wage disparities across gender and spatial groups in Uganda, using integrated multi-round survey data and XAI techniques.

1.4.2 Specific Objectives

1. To pre-process and harmonise four rounds of Uganda National Household Survey data spanning twelve years (2012/13–2023/24) into a consistent, analysis-ready pooled dataset.
2. To train and evaluate explainable machine learning models for predicting monthly wage differentials.
3. To decompose gender and urban–rural wage gaps into explained and unexplained components using ML-augmented Oaxaca–Blinder techniques, and to analyse how these components evolved across the four survey waves.

1.5 Research Questions

1. How were four rounds of UNHS microdata carrying different variable definitions, coding schemes, and classification systems pre-processed and harmonised into a unified dataset suitable for multi-round wage analysis?
2. Which machine learning model among OLS, LASSO, Random Forest, and XGBoost most accurately predicted monthly wage in Uganda, and what did SHAP-based explainability reveal about the dominant wage drivers across gender and spatial groups?
3. How did the explained and unexplained components of the gender and urban–rural wage gaps change across the four UNHS waves (2012/13 to 2023/24), and what did the ML-augmented Oaxaca–Blinder decomposition reveal about these trends that conventional OLS-based decomposition did not?

1.6 Scope of the Study

Thematic scope. The study covered wage determinants and wage gap decomposition for workers in formal paid employment. Unpaid domestic work, subsistence agriculture, and self-employment earnings were left out, since the Oaxaca–Blinder decomposition requires a continuous wage variable reported in a consistent format.

Geographic scope. The analysis covered all four statistical regions of Uganda, namely Central, Eastern, Northern, and Western, as defined by UBOS.

Temporal scope. The study spanned four UNHS rounds: 2012/13, 2016/17, 2019/20, and 2023/24.

Analytical scope. Four modelling approaches were compared, namely OLS, LASSO, RF, and XGBoost, alongside SHAP-based explainability and Oaxaca–Blinder decomposition. Causal inference methods such as instrumental variables or regression discontinuity were not attempted, because the UNHS design does not support causal identification.

1.7 Justification of the Study

The case for this study is primarily methodological, and its academic contribution lies in extending an established decomposition methodology into a setting where its assumptions have never been tested.

The first concerns the transferability of the method. ML-augmented Oaxaca–Blinder decomposition with SHAP explainability has been demonstrated in high-income countries, notably Germany [9] and Switzerland [8], but it had not been applied to sub-Saharan African household survey data, where labour markets are more segmented, informal employment dominates, and nonlinear wage dynamics are likely to be more pronounced. Applying the framework to four rounds of UNHS microdata tested a claim that the existing literature had assumed rather than demonstrated, namely that the accuracy and interpretability gains reported in high-income settings survive the move to a data-sparse, developing-country context. Establishing the conditions under which a method holds is itself an academic result, and the subgroup-sufficiency thresholds reported in Chapter Three formalise where those conditions break down.

The second concerns coverage across survey rounds. Existing ML-augmented decomposition studies have been conducted at single time points, which leaves the stability of their findings untested. Extending the approach across four survey waves spanning twelve years allowed the explained and unexplained components to be observed as a series rather than a snapshot, so that convergence or divergence in gap composition becomes an empirical finding rather than an assumption. No prior study of Ugandan wage data has produced such a series.

The third concerns the replicability of the framework. The end-to-end pipeline, running from multi-round survey harmonisation through ML-based earnings prediction with SHAP attribution to nonlinear Oaxaca–Blinder analysis, was built as a transferable template for other countries and survey programmes [23, 24]. Its academic value is that it makes the harmonisation decisions, which are usually buried in unpublished cleaning scripts, explicit and open to scrutiny and replication.

1.8 Significance of the Study

The significance of the study lies chiefly in what it establishes methodologically, with practical consequences following from that rather than standing alongside it.

For data science, the study tests a method outside the conditions in which it was developed. ML-augmented decomposition with SHAP attribution has been shown to work on large, clean administrative data from high-income economies; whether it holds up on a data-sparse survey from a segmented developing-country labour market was unknown before this work. Establishing where it holds, and where sample sizes force a retreat to simpler estimators, is a result in its own right.

Some care is needed over the word explain. The models predicted wages, and SHAP attributed the resulting gap to particular observable features, but neither operation identifies a causal mechanism. What the study measured is how much of the gap education, sector, region, experience, and occupation account for, and how much survives once they have been accounted for. That surviving portion is a residual containing the combined effect of everything the survey does not measure. It may include unequal treatment of similar workers, but it also includes unobserved ability, hours actually worked, job tenure, firm size, and error in the measurement of wages themselves. It is not, and is not treated in this thesis as, a measure of discrimination.

For policy research, the decomposition outputs supply something the Ugandan evidence base has lacked, which is a feature-level account that runs across four survey rounds rather than one. Equal-pay auditing needs to know which characteristics drive disparities in which sectors and regions, and the SHAP rankings answer that question directly. Turning those rankings into specific policy prescriptions is not attempted here and is left to subsequent work.

For other researchers, the pipeline itself is the deliverable. It is documented end to end and can be pointed at a different national survey, a different dimension of inequality, or a different dependent variable with the analytical logic unchanged.

1.9 Cultural and Policy Context

1.9.1 Traditional and Cultural Dimensions of Employment

Uganda’s labour market is shaped by economic conditions and also by long-standing cultural norms governing who works, where, and for what pay. In many communities, traditional gender roles direct women toward domestic labour, subsistence farming, and care work, while men are directed toward cash-crop agriculture, formal employment, and entrepreneurship [2, 3, 16]. Kinship-based inheritance systems, bride price practices, and patrilocal marriage norms reinforce these patterns and limit the economic autonomy and geographic mobility of women [1, 15].

The same divisions surface in the formal labour market as occupational segregation, with women concentrated in lower-paying sectors such as education, health, and domestic service, and men dominating construction, transport, manufacturing, and public administration. The UNHS captures these patterns through its International Standard Industrial Classification of All Economic Activities (ISIC) and International Standard Classification of Occupations (ISCO) variables. Within the Oaxaca–Blinder framework, cultural norms can register in two places at once: in the explained component, through occupation and sector variables that reflect gendered sorting, and in the unexplained component, through within-occupation pay differentials that may partly reflect employer attitudes shaped by those same norms. Cultural attitudes were not measured directly in this study, but the decomposition results identify the sectors and regions where unexplained gaps were widest, which offers a starting point for qualitative investigation into the role of tradition.

1.9.2 Policy Landscape for Gender and Labour in Uganda

Several policy instruments frame the context in which the findings of this study are interpreted.

The **Employment Act 2006** (Sections 6 and 75) prohibits discrimination in employment on the basis of sex and mandates equal remuneration for work of equal value, although enforcement remains limited, particularly in the informal sector.

The **National Gender Policy** (revised 2007) commits Uganda to mainstreaming gender across all sectors, including employment and economic participation, and identifies the gender wage gap as a key indicator of women’s economic empowerment.

The **Equal Opportunities Commission (EOC)**, established under the Equal Opportunities Commission Act 2007, monitors and enforces compliance with gender equity provisions. Its Annual Report on the State of Equal Opportunities treats wage disparities as a persistent area

of concern while noting the absence of rigorous quantitative evidence on the structure and drivers of those disparities.

The **Third National Development Plan (NDP III, 2020/21–2024/25)** sets targets for reducing gender disparities in labour force participation and earnings, with monitoring indicators drawn from the UNHS and the National Labour Force Survey (NLFS).

This study contributes directly to the evidence base that these instruments require, by providing feature-level decompositions across four survey rounds that show where interventions are most needed by sector, region, education level, and employment type.

1.10 Theoretical Foundation

The study was grounded in the Mincer human capital earnings function, which models wages as a function of education, experience, and individual characteristics [18], and this provided the feature set for the ML models. The Oaxaca–Blinder decomposition [25, 26] supplied the analytical structure for partitioning wage gaps into explained and unexplained components. The ML augmentation extended both by relaxing the linearity assumption, which allowed the models to capture the nonlinear interactions and threshold effects that Ugandan wage data are known to exhibit [19, 9]. SHAP values [10] then bridged the ML predictions and the economic theory, giving feature-level attribution that OLS coefficients cannot offer.

1.11 Definition of Key Terms

The following definitions were used throughout this study, aligned with the standards of the International Labour Organization (ILO), the Uganda Bureau of Statistics (UBOS), and established econometric practice.

Wage. Monetary compensation received by an individual in exchange for labour services rendered to an employer. In this study, wage refers specifically to self-reported gross monthly cash earnings from paid employment, as recorded in the UNHS labour force module [1, 6].

Formal Paid Employment.

An employment arrangement in which an individual works for a public or private employer and receives a regular wage or salary, following the ILO International Classification of Status in Employment (ICSE-18) [1]. In this study the category was identified by the presence of a reported monthly cash wage, which in practice excludes self-employment, own-account work, contributing family workers, and subsistence agriculture, since none of these report one. The operational detail is given in Section 3.3.6.

Gender Wage Gap.

The difference in average earnings between male and female workers, expressed as a proportion of male earnings. Following ILO convention, it is calculated as $(\bar{W}_M - \bar{W}_F)/\bar{W}_M \times 100$, where $\bar{W}_M$ and $\bar{W}_F$ denote mean earnings for males and females respectively [1].

Urban–Rural Wage Gap.

The difference in average earnings between workers residing in urban and rural areas, as classified by UBOS using Enumeration Area (EA) designations in the UNHS sampling frame [6].

Oaxaca–Blinder Decomposition.

A statistical technique that partitions the mean wage difference between two groups into an explained component, attributable to differences in observable characteristics such as education and experience, and an unexplained component, attributable to differences in returns to those characteristics together with everything the model does not observe [25, 26].

Explained Component.

The portion of the wage gap attributable to measurable differences in worker characteristics, such as education level, work experience, occupation, and sector, between the two groups being compared.

Unexplained Component.

The part of a measured wage gap that differences in observed characteristics do not account for. It is sometimes described in the econometric literature as an upper bound on discrimination [27], and that phrasing is easily misread. It is not a measure of discrimination. It holds every influence on pay that the survey does not record, including hours actually worked, job tenure, firm size and productivity, effort, professional networks, and error in reported earnings, together with any unequal treatment of otherwise similar workers [27, 17]. In this thesis it is reported as a residual and is not interpreted as discrimination at any point.

Human Capital.

The stock of skills, knowledge, and experience possessed by an individual that contributes to their productive capacity in the labour market. The Mincer earnings function operationalises human capital through years of schooling and potential experience [18].

Potential Experience.

A proxy for actual labour market experience, calculated as $\text{Age} - \text{Years of Schooling} - 6$, following the standard Mincer specification [18, 21]. The subtraction of 6 assumes school entry at age 6.

SHAP Values.

A model-agnostic explainability method based on cooperative game theory that assigns each feature an additive contribution to a model’s prediction for a given observation. SHAP values satisfy the properties of local accuracy, missingness, and consistency [10]. The implementation used in this study is set out in Section 3.4.1.

Survey Round.

One complete cycle of the UNHS. This study used four rounds: 2012/13, 2016/17, 2019/20, and 2023/24. Each round is a repeated cross-section rather than a panel, meaning that different households are sampled in each round [6]. The consequences of this structure are set out in Section 3.3.1.2.

Chapter 2

Literature Review

2.1 Introduction

This chapter reviews the literature on wage gap measurement and positions the present study within it. The review is organised chronologically rather than thematically, because the central argument of this thesis concerns a methodological sequence: each generation of decomposition methods solved a problem left open by the generation before it, and the problem left open by the current generation is the one this study addresses.

Section 2.2 sets out the theory of wage determination that all the studies reviewed share. Section 2.3 traces the development of wage gap decomposition from 1973 to the present, showing at each stage which limitation was resolved and which remained. Section 2.4 reviews empirical findings for sub-Saharan Africa and Uganda, and examines how many survey rounds each of the Ugandan studies actually used. Section 2.5 discusses the UNHS as a data source for multi-round wage analysis, and Section 2.6 reviews how other multi-round survey programmes have handled comparability across waves. Section 2.7 covers explainability methods for predictive models. Section 2.8 draws the strands together, states what the literature has settled and what it has not, and maps the remaining gaps onto the three objectives of this study.

The review is not written only to summarise what has been done. Each section closes by stating what the work reviewed in it leaves unresolved, and Section 2.8 collects those unresolved points and matches each one to an objective stated in Section 1.4. A reader should therefore be able to trace every objective of this study back to a specific limitation in the published work, rather than having to take the stated gap on trust.

2.2 Theoretical Foundations of Wage Determination

Almost every study reviewed in this chapter rests on the same theoretical base. Mincer [18] modelled the logarithm of earnings as a linear function of years of schooling and a quadratic in potential experience, on the reasoning that schooling is an investment whose return is realised as higher lifetime earnings. The specification proved durable because it is parsimonious, estimable from any survey carrying education and age, and interpretable, since the schooling coefficient reads directly as an average rate of return.

Its durability, however, has come at a cost that later work has repeatedly documented. Lemieux [21] showed that the log-linear form fits post-1980 data poorly, because returns rise disproportionately at higher levels of schooling, so that the true schooling profile is convex rather than linear. Heckman *et al.* [28] reached the same conclusion from a different direction, showing that

the coefficient on schooling in a Mincer equation can be read as a rate of return only under conditions that rarely hold in practice. Psacharopoulos and Patrinos [22], reviewing returns to education across 139 economies, found the same convexity in developing countries, where the tertiary premium far exceeds what a linear specification implies. The convexity matters for gap analysis in particular: if the schooling profile is nonlinear and men and women are distributed differently along it, a linear model will misallocate part of the gap.

A second body of theory questions the assumption that a single earnings function governs the whole labour market. Segmented labour market accounts hold that formal and informal work, and public and private employment, operate as distinct wage-setting regimes with different returns to identical characteristics. Kagundu and Pavlova [19] found direct evidence of this in Uganda: the female-to-male wage ratio was 1.04 in the urban public sector but 0.55 in the urban private sector. A pooled linear model cannot represent a difference of that magnitude without explicit interaction terms, and the number of interactions needed grows quickly with the number of segments.

Both strands of theory point the same way. The functional form imposed by the Mincer equation is a simplification whose cost is highest precisely in segmented, developing-country labour markets. This is the theoretical opening that the machine learning literature reviewed in Section 2.3 exploits.

2.3 The Evolution of Wage Gap Decomposition, 1973–2024

The methodological problem addressed in this thesis has a long history, and the sequence in which it has been tackled is the clearest way to see what remains outstanding. Table 2.1 summarises the sequence, and the subsections that follow explain each stage.

Table 2.1: Chronological development of wage gap decomposition methods

Period	Development	Problem solved and problem left open
1973	Oaxaca; Blinder [25, 26], building on Kitagawa [29]	Solved: separating a mean wage gap into endowment and returns components. Left open: which group’s coefficients serve as the non-discriminatory benchmark.
1979	Heckman selection correction [30]	Solved: bias from non-random selection into wage employment. Left open: valid exclusion restrictions are rarely available in household surveys.
1988	Neumark pooled-coefficient reference [27]; Cotton share-weighted reference [31]	Solved: the index-number problem in choosing a reference wage structure. Left open: two defensible references exist, and the decomposition still describes the mean only.
1996–2005	DiNardo <i>et al.</i> ; Machado and Mata [32, 33]	Solved: decomposition across the whole wage distribution rather than at the mean. Left open: linearity within each quantile regression is retained.
2008–2011	Ñopo matching; Jann’s implementation; Fortin <i>et al.</i> synthesis [34, 35, 36]	Solved: common-support problems, and standardised software that made the method routine. Left open: functional form is still specified by the analyst.
2017–2019	Machine learning enters econometrics [37, 38, 39]	Solved: flexible estimation of conditional expectations without pre-specified functional form. Left open: the resulting models are opaque, so decomposition results cannot be attributed to features.
2017–2020	SHAP and tree explainers [10, 40]	Solved: additive, theoretically grounded feature attribution for opaque models. Left open: attribution had not been connected to the decomposition framework.
2021–2022	ML-augmented decomposition [8, 9]	Solved: substituting ML predictions into Oaxaca–Blinder, shifting weight from the unexplained to the explained component. Left open: demonstrated only in high-income countries, at single time points.

2.3.1 The Classical Decomposition and Its Immediate Problems

Oaxaca [25] and Blinder [26], publishing independently in the same year, proposed splitting the difference in mean log wages between two groups into a portion attributable to differences in average characteristics and a portion attributable to differences in the coefficients those characteristics attract. The first portion became known as the explained component and the second as the unexplained component. Oaxaca and Sierminska [29] note that the underlying algebra had in fact been set out by Kitagawa nearly two decades earlier in a demographic context, and argue for crediting the method accordingly. Every study reviewed below is a descendant of this idea.

Two difficulties surfaced almost at once. The first is the index-number problem: the split depends on whose coefficients are treated as the non-discriminatory benchmark, and using male coefficients gives a different answer from using female ones. Neumark [27] resolved it by estimating a pooled wage structure over both groups and using that as the reference, while Cotton [31], in the same year, proposed weighting the two groups’ coefficients by their population

shares instead. The Neumark reference is the convention most subsequent work has followed, including both Ugandan studies reviewed in Section 2.4.2, but the existence of two defensible answers is a reminder that the explained/unexplained split is partly a modelling choice rather than a property of the data. The second difficulty is selection. Wages are observed only for people in paid employment, and if entry into paid employment is systematically related to unobserved earning capacity, the estimated wage equations are biased. Heckman [30] provided the standard two-step correction, though its credibility rests on an exclusion restriction, a variable affecting participation but not wages, and household surveys seldom supply a convincing one.

2.3.2 Moving Beyond the Mean

The classical decomposition describes the average worker, which is unsatisfying when a gap is concentrated at one end of the distribution. DiNardo *et al.* [32] introduced a reweighting approach that constructs counterfactual densities, and Machado and Mata [33] combined quantile regression with simulation to decompose changes at any point in the distribution. Firpo *et al.* [41] then developed unconditional quantile regression, which allows a decomposition to be performed at a chosen quantile with the same interpretive logic as at the mean. Fortin *et al.* [36] consolidated these developments into the reference treatment of decomposition methods in economics.

This line of work matters here because it established that gaps are rarely uniform across the distribution. The sticky-floor pattern reported for Uganda by UN Women [7], in which the gender penalty is widest among the lowest earners, is exactly the kind of feature that a mean decomposition conceals. What the distributional methods did not relax, however, is the assumption of linearity within each estimated equation. They ask the model to describe a different part of the distribution, not to adopt a different functional form.

Ñopo [34] approached the problem from another direction, using matching to restrict the comparison to men and women with genuinely comparable characteristics, and showing that part of what earlier studies had called unexplained arises from comparing groups whose characteristics do not overlap. Jann [35] contributed the standard software implementation, which turned decomposition into a routine procedure and accounts for much of its subsequent ubiquity in applied work.

2.3.3 Relaxing the Functional-Form Assumption

The functional-form assumption remained untouched until machine learning methods were adopted in econometrics. Mullainathan and Spiess [37] argued that a class of economic problems is fundamentally about prediction rather than parameter estimation, and that for these, regularised and tree-based methods dominate ordinary least squares. Athey and Imbens [38] surveyed the methods most relevant to economists, and Chernozhukov *et al.* [39] showed how flexible first-stage predictions can be embedded in a structural estimation without contaminating inference.

Applied to wages, the appeal is direct. Random Forest [42] accommodates mixed feature types, tolerates outliers, and recovers interactions without their being specified. XGBoost [43] builds trees sequentially under L_1 and L_2 regularisation and generally sets the standard on tabular data. Cengiz *et al.* [44] demonstrated the value of such methods in labour economics specifically. The obstacle was interpretability: a model that predicts well but cannot say why is of limited use in a decomposition, whose entire purpose is attribution.

2.3.4 Recovering Interpretability from Opaque Models

Ribeiro *et al.* [45] offered an early solution through local surrogate models, and Lundberg and Lee [10] then unified the competing approaches under Shapley values, producing attributions that are additive, locally accurate, and consistent. Lundberg *et al.* [40] subsequently made exact computation tractable for tree ensembles, which is what renders the method practical on survey-scale data.

With flexible prediction and credible attribution both available, the two could be combined. Strittmatter and Wunsch [8] applied machine learning to the gender wage gap using Swiss administrative data, and Bonaccolto-Töpfer and Briel [9] carried out an ML-augmented Oaxaca–Blinder decomposition on German data. Their central finding is the one that motivates this thesis: relative to linear decomposition, the ML-augmented version shifts explanatory weight from the unexplained component to the explained component. Part of what linear models had been attributing to discrimination turns out to be nonlinear structure among observable characteristics that those models could not represent. Lombardía *et al.* [23] pursued a different route away from the classical linear framework, proposing a bias-corrected, model-based decomposition for small-area gender pay gap estimation; their work is evidence that the limitations of the classical approach were being addressed outside the machine learning literature as well. Mora and Herrera [24] extended decomposition to repeated cross-sections, which is the data structure the UNHS provides.

Two limitations of this most recent generation are worth stating plainly, because they define what this study attempts. Both demonstrations used high-income countries with large administrative datasets, and neither examined more than a single time point.

2.4 Empirical Evidence on Wage Inequality in Uganda and Sub-Saharan Africa

2.4.1 Regional Evidence

Cross-country work provides the backdrop against which the Ugandan findings should be read. Weichselbaumer and Winter-Ebmer [46], in a meta-analysis of 263 published studies, found that raw gender wage gaps have narrowed considerably over time while the unexplained portion has proved far more resistant, which suggests that equalising characteristics is easier than equalising the returns paid on them. Blau and Kahn [17] documented the same asymmetry for the United States, where convergence in education and experience has largely run its course while the residual gap persists.

For sub-Saharan Africa the evidence is thinner but consistent. Appleton *et al.* [47] compared Uganda, Côte d’Ivoire, and Ethiopia, and found substantial gender earnings gaps in all three, with public sector employment narrowing the gap in each case. Nordman *et al.* [48] examined seven West African cities and found that occupational sorting accounted for a large share of the gap, though the share varied widely by city. Brixiová and Ncube [5] found comparable gaps in Eswatini. Van den Broeck and Swinnen [4] showed that structural transformation across the region has not closed gender wage gaps uniformly, and in some sectors has widened them. The International Labour Organization [1] situates these findings in the global pattern of stagnant real wage growth following the pandemic.

2.4.2 Uganda-Specific Evidence

The Ugandan evidence base is narrow and largely dated. Two studies, both published in 2007 and both analysing the UNHS 2002/03, remain the standard references.

Kagundu and Pavlova [19] separated the sample into urban and rural sub-samples and applied the Oaxaca decomposition [25] with the Neumark adjustment [27] and the Heckman two-step correction [30]. They reported a female-to-male hourly wage ratio of 0.73 in urban areas and 0.67 in rural areas, and a sharp public–private divide, with an urban public sector ratio of 1.04 against 0.55 in the private sector. Between 41 and 68 percent of the urban gender wage gap was unexplained. Returns to education were large, with university-educated men earning roughly 122 percent more and women 165 percent more than those with no education, and occupational segregation was pronounced, with 25.9 percent of employed urban women working as domestic workers.

Sebaggala [20], working from the same dataset, estimated an overall male–female wage gap of approximately 39 percent. Applying the Oaxaca decomposition with the Neumark reference structure, he found that most of the unexplained differential took the form of a wage penalty imposed on women rather than a premium accruing to men. The distinction matters for interpretation, since a penalty and a premium of equal size have different implications for where remedial policy should be directed.

Nothing of comparable methodological depth appeared for over a decade. The Economic Policy Research Centre [49] examined labour market transitions among young Ugandan women and men, documenting gendered differences in the move from education into work, but the study is descriptive and does not decompose earnings. The UN Women Gender Pay Gap Report [7], using UNHS 2019/20 microdata, put the monthly gender wage gap at approximately 32 percent and identified a sticky-floor pattern in which the penalty is largest at the bottom of the distribution. The UNHS 2023/24 report itself [6] documents the spatial dimension: mean monthly consumption expenditure per adult differs by nearly five to one between Kampala and Karamoja, and rural poverty stands at 19.4 percent against 5.7 percent in urban areas. These are descriptive statistics rather than decompositions, and no attempt is made in the report to separate composition from returns.

2.4.3 How Many Survey Rounds Have Existing Studies Used?

Since this study’s first objective concerns multi-round harmonisation, the survey coverage of the existing literature is directly relevant. Table 2.2 records, for each Ugandan and comparable regional study reviewed, the data used and the number of survey rounds analysed.

Table 2.2: Survey coverage of reviewed wage gap studies relevant to Uganda

Study	Data	Rounds	Method
Appleton <i>et al.</i> (1999) [47]	Uganda IHS 1992; Côte d’Ivoire; Ethiopia	One per country	Oaxaca decomposition with selection correction
Kagundu and Pavlova (2007) [19]	UNHS 2002/03	One	Oaxaca–Neumark with Heckman correction, urban and rural sub-samples
Sebaggala (2007) [20]	UNHS 2002/03	One	Oaxaca–Neumark decomposition
Nordman <i>et al.</i> (2011) [48]	Seven West African city surveys	One per city	Quantile decomposition
Brixiová and Ncube (2021) [5]	Eswatini labour force survey	One	Oaxaca–Blinder decomposition
UN Women (2024) [7]	UNHS 2019/20	One	Descriptive and quantile analysis
UBOS (2025) [6]	UNHS 2023/24	One	Descriptive statistics only
This study	UNHS 2012/13, 2016/17, 2019/20, 2023/24	Four	ML-augmented Oaxaca–Blinder with SHAP attribution

The pattern in the table is the point. Every wage gap study located for Uganda, and every regional comparator, analyses a single survey round. None pools two rounds, and none pools all available rounds. The consequence is that the Ugandan literature contains a series of disconnected snapshots taken with different methods at different times, from which no statement about direction of travel can be derived. Whether the 39 percent gap Sebaggala measured in 2002/03 and the 32 percent gap UN Women measured in 2019/20 represent genuine improvement, or merely differences in sample definition and method, cannot be established from the published work. Only a harmonised multi-round analysis using a single consistent specification can settle it, and Mora and Herrera [24] show that repeated cross-sections are adequate for the purpose, provided the harmonisation is done carefully.

2.5 The UNHS as a Data Source for Multi-Round Wage Analysis

The UNHS is a nationally representative, multi-topic household survey conducted by UBOS at roughly four-year intervals under a two-stage stratified sampling design, covering demographics, education, employment, earnings, consumption expenditure, and housing conditions [6]. Four rounds are relevant to this study: 2012/13, 2016/17, 2019/20, and 2023/24 [50, 51, 52, 6].

Two features of the design constrain what can be asked of the data. The rounds are repeated cross-sections rather than a panel, so different households are sampled each time and individual wage trajectories cannot be followed; what can be compared is the population distribution at successive points. Mora and Herrera [24] address precisely this structure and show that gap composition can be tracked across repeated cross-sections even though individuals cannot. The

rounds are also not identical instruments. Variable names, occupation and industry coding schemes, and district boundaries all changed over the twelve years spanned, which is the reason harmonisation is treated as a substantive objective in this study rather than as preparatory work.

Drawing on the specifications used by Kagundu and Pavlova [19] and Seabagala [20], and on the wider returns-to-education literature [22, 18], the feature set for this study comprises education level, gender, urban or rural location, sector of employment, public or private employment, age and experience, marital status, and household characteristics. The difference from the earlier studies lies not in which variables are used but in what the model is permitted to do with them: the tree-based estimators recover nonlinear effects and interactions without the analyst specifying them in advance.

The bearing of this section on the purpose of the study is direct. The UNHS carries the variables the Mincer specification requires and covers twelve years, so it can in principle answer the question this study asks. It does not deliver that capability ready-made, because variable names, classification systems, and district boundaries all change between rounds. That is the reason harmonisation is stated as Objective 1 and treated as a substantive contribution in this thesis, rather than handled quietly in a cleaning script as it is in studies that use a single round.

2.6 Harmonisation Practice in Multi-Round Survey Programmes

Table 2.2 pointed to a specific gap: within the literature reviewed in this chapter, no Ugandan wage study pools more than a single survey round, so no domestic precedent exists for the harmonisation problem this study had to solve. The relevant precedent instead comes from international survey programmes that routinely combine multiple rounds, and it is their accumulated practice, not any single Ugandan antecedent, that this study’s pipeline draws on. Chapter Three sets out the procedures actually implemented; what follows here is an account of where that practice originates.

Ex-post harmonisation, meaning the reconciliation of surveys that were never designed to be comparable, was formalised by Fortier *et al.* [53] in the Maelstrom protocol. The protocol proceeds through four stages: an inventory of the contributing studies, definition of the target variables, formal documentation of every harmonisation decision, and evaluation of the resulting quality. Its distinguishing feature is the insistence that each decision be recorded and justified rather than left implicit in cleaning code. The Luxembourg Income Study [54] and the Integrated Public Use Microdata Series [55] pioneered the underlying approach, and between them demonstrate that comparability across decades and countries is achievable when the decisions are made explicit. Kołczyńska [56] takes the argument further, showing that a harmonisation is only as credible as the workflow that produced it, and that the combination step should be released as executable code rather than described in prose.

Classification change is the recurring obstacle. Industry codes migrate between ISIC revisions and occupation codes between ISCO versions, so that identical jobs receive different codes in different rounds. The United Nations Statistics Division [57] publishes concordance tables for exactly this purpose, and the Living Standards Measurement Study – Integrated Surveys on Agriculture programme [58] has developed comparable practice for African household surveys, including the treatment of administrative boundary changes, which are frequent in Uganda as districts subdivide.

Missing data is the other standing problem. Van Buuren [59] provides the reference treatment, distinguishing missingness mechanisms and setting out when deletion, single imputation, or multiple imputation is defensible. The distinction matters for wage analysis because missingness in reported earnings is unlikely to be random, and because the dependent variable and the predictors call for different treatment.

What this literature does not supply is a worked application to Ugandan survey data. The programmes above operate on their own harmonised collections; no published pipeline harmonises the UNHS rounds for wage analysis, and the harmonisation decisions taken in the studies reviewed in Section 2.4 are not documented, since each analysed a single round and therefore never faced the problem.

What this body of practice supplies to the present study is a standard to work to rather than a solution to reuse. The procedures implemented in Section 3.3.6 therefore follow the Maelstrom sequence and Kołczyńska’s argument that the combination step should be released as executable code, but the target schema, the crosswalks, and the missing-data rules all had to be built for these four rounds specifically. That construction is what Objective 1 delivers.

2.7 Explainability in Predictive Wage Models

Predictive accuracy alone does not make a model useful for decomposition, which requires knowing how much each characteristic contributed to a predicted wage. SHAP values [10] supply this through a model-agnostic attribution grounded in cooperative game theory, satisfying local accuracy, missingness, and consistency, and permitting questions that an OLS coefficient cannot answer, such as how much sector of employment contributed to a particular worker’s predicted wage relative to the population average. Saeed and Omlin [60] survey the broader explainable artificial intelligence field and its open problems, while Molnar [11] provides the standard practical treatment, including the conditions under which attributions can mislead when features are strongly correlated. That caveat is pertinent here, since education, occupation, and sector are correlated in the UNHS data.

Evaluation practice in this literature is settled. Root Mean Squared Error, Mean Absolute Error, and R^2 on a held-out test set are standard [11], with hyperparameters tuned on a separate validation split for tree ensembles and by cross-validation for regularised linear models. The Ugandan literature offers a natural benchmark: Kagundu and Pavlova [19] reported R^2 values of 43.6 percent for urban males and 59.7 percent for urban females, figures against which the models in this study can be compared using a similar feature set.

This section supplies the second half of the argument behind Objective 2. Section 2.3 established that flexible models remove the functional-form restriction but lose the attribution a decomposition depends on. The SHAP literature reviewed here shows that the attribution can be recovered, exactly and additively, for tree ensembles. Objective 2 is therefore not simply to find the most accurate wage model. It is to establish whether accuracy and attribution can be held together on UNHS data, because the decomposition in Objective 3 cannot proceed unless they can.

2.8 Synthesis and Research Gaps

Read in sequence, the literature tells a coherent story. The problem of separating a wage gap into composition and returns was posed and solved in principle in 1973. The reference-structure and selection problems that followed were addressed by 1988. The restriction to the

mean was lifted between 1996 and 2011. The restriction to analyst-specified functional form was lifted after 2017, once machine learning methods entered econometrics, and the interpretability cost of that move was paid off by SHAP between 2017 and 2020. By 2022 the two had been combined, and ML-augmented decomposition had been shown to move explanatory weight from the unexplained component to the explained one.

That sequence has nonetheless left three matters unresolved. First, every demonstration of ML-augmented decomposition located in this review comes from a high-income country with large, clean administrative data. Whether the gains survive in a segmented, data-sparse, developing-country labour market is an open empirical question, not a settled one. Second, every such demonstration is a single snapshot, so nothing is known about whether the shift from unexplained to explained is stable over time or an artefact of one particular year. Third, for Uganda specifically, the substantive literature consists of two studies from 2007 using data from 2002/03, plus more recent descriptive reporting. As Table 2.2 shows, no study located here has pooled survey rounds, and consequently there is no consistent series from which the direction of travel could be read.

2.8.1 How the Reviewed Literature Provides the Rationale for This Study

The three gaps above are statements about what is missing. The rationale for conducting this particular study rests on four specific findings in the work reviewed, each of which points directly to a decision taken in Chapter Three.

The literature shows the functional form is the binding constraint, not the data. Lemieux [21] and Psacharopoulos and Patrinos [22] both find that returns to schooling are convex rather than linear, and Kagundu and Pavlova [19] find returns differing by a factor of nearly two between the urban public and urban private sectors of the same economy. Taken together, these say that a linear wage equation is misspecified in exactly the setting this study examines. That is the reason the study compares estimators of increasing flexibility rather than adopting one in advance, and the reason the comparison is worth making on Ugandan data specifically.

The literature shows what happens when that constraint is relaxed, and gives a direction to test against. Bonaccolto-Töpfer and Briel [9] and Strittmatter and Wunsch [8] both report that ML-augmented decomposition shifts weight from the unexplained to the explained component. This is a prediction and not merely an observation, and it can fail. The present study is a test of it in a setting where the authors of those papers did not claim it would hold.

The literature shows that attribution can be recovered, so the test is feasible. Lundberg and Lee [10] and Lundberg *et al.* [40] establish that feature attribution for tree ensembles is exact and additive. Without that result, substituting a tree model into Oaxaca–Blinder would produce a number with no interpretable breakdown, and the exercise would answer nothing that a prediction contest could not. This is why explainability is an objective in its own right here rather than a presentational extra.

The literature shows the data structure is adequate, and states on what conditions. Mora and Herrera [24] set out how gap decomposition behaves on repeated cross-sections, which is what the UNHS provides, and Fortier *et al.* [53] together with Kolczyńska [56] set out how surveys never designed to be comparable can be reconciled after the fact. Between them they establish that the multi-round question is answerable with the data that already exist, provided

the harmonisation is documented. That is the condition Objective 1 is written to satisfy.

Table 2.3 maps the three gaps onto the objectives stated in Section 1.4.

Table 2.3: Research gaps and the objectives that address them

Gap	What the literature has not done	Objective addressing it
1	No harmonised multi-round UNHS dataset exists for wage analysis; every Ugandan study uses a single round, so harmonisation decisions have never had to be made or documented.	Objective 1: pre-process and harmonise four UNHS rounds into a consistent analysis-ready dataset.
2	ML-augmented decomposition with SHAP attribution has been demonstrated only in high-income settings; its performance on segmented, data-sparse labour markets is untested.	Objective 2: train and evaluate explainable ML models for predicting monthly wage differentials in Uganda.
3	Existing ML-augmented decompositions are single-snapshot, and the Ugandan evidence consists of disconnected estimates from which no trend can be inferred.	Objective 3: decompose gender and urban–rural gaps across four waves and analyse how the components evolved.

The chapter that follows sets out the methodology through which these three objectives were pursued.

Chapter 3

Methodology

3.1 Introduction

This chapter sets out how the study was carried out. It covers the research design and the philosophical stance behind it, the data and the study population, the variables used, how the four survey rounds were cleaned and combined, how the models were built and explained, and how the wage gap was decomposed. The chapter follows the order in which the work was actually done: data preparation first, then modelling and explainability, then decomposition and the checks that establish how far the results can be trusted.

3.2 Research Design

The study used a quantitative, non-experimental design built on secondary survey microdata. Four rounds of nationally representative household survey data were harmonised into a single pooled multi-round dataset, supervised regression models were fitted to it, and the resulting wage gaps were decomposed. No primary data were collected and no treatment was assigned to anyone, so the design is descriptive and comparative rather than experimental. All work was carried out in Python 3.10, using Pandas, Scikit-learn [61], XGBoost [43], SHAP [10], Statsmodels, and Optuna [62].

The design is comparative in two senses, and both are what allow the objectives to be answered. Across models, four estimators of increasing flexibility, namely OLS, LASSO, Random Forest, and XGBoost, were fitted to the same data and scored on the same held-out test set, so that a difference in accuracy can be attributed to the estimator rather than to the sample. Across rounds, one specification was applied to each of the four survey waves in turn, so that a difference in the decomposition can be attributed to the period rather than to a change of method. The single-round, single-model Ugandan studies reviewed in Chapter Two could make neither comparison.

An earlier version of this work described the design as Design Science Research, on the reasoning that the harmonisation pipeline is a designed object evaluated against real data. That label has been withdrawn. Design Science Research, as set out by Hevner *et al.* [63] and Peffers *et al.* [64], centres on building an artefact and then demonstrating and evaluating it against the problem it was built for, normally with intended users and in a setting of use. This study produces analytical results together with a documented and reusable pipeline, but it does not build, deploy, or evaluate a system in that sense, and no demonstration or evaluation stage of the kind those authors describe was carried out. Describing the work as design science would

therefore claim a form of contribution the study does not make. The pipeline is reported instead as a research output in Section 6.7, alongside the methodological and empirical contributions, and the design of the study is stated as above.

3.2.1 Philosophical Stance: Pragmatism

The study’s philosophical stance is pragmatist. Pragmatism holds that the choice of method should follow from the research question rather than from prior commitment to a single epistemological position [65]. That stance is what permits the study to combine, without contradiction, a positivist measurement instrument, the Oaxaca–Blinder decomposition, with a machine learning component whose justification is functional rather than axiomatic: Random Forest and XGBoost were adopted because they predicted wages more accurately than OLS on this data, not because of a theoretical commitment to nonlinear models. Feilzer [66] argues that pragmatism is the natural fit for applied, problem-driven research that draws on quantitative technique in the service of a practical question, which is the posture this study takes throughout: the guiding question was whether a more flexible decomposition method changed what is known about Uganda’s wage gaps, not whether either method was correct in principle.

3.3 Data Sources

3.3.1 Data Used

The study drew on secondary microdata from UBOS. No primary data were collected.

3.3.1.1 Uganda National Household Survey

The UNHS is a nationally representative, multi-topic household survey using a two-stage stratified cluster sample, run by UBOS at roughly four-year intervals. Four rounds were used: UNHS 2012/13, 2016/17, 2019/20, and 2023/24. The variables of interest were drawn from two modules of each round: the household roster, which supplies demographic characteristics, and the labour module, which supplies employment status, occupation, industry, and monthly cash earnings. Each round also uses its own variable names, coding schemes, and geographic identifiers, which is what makes the harmonisation procedure described in Section 3.3.6 a necessary step rather than a formality.

3.3.1.2 Structure of the Data: Repeated Cross-Sections, Not a Panel

The difference between a panel, a longitudinal design, and a repeated cross-section governs what this study can and cannot claim, so it is set out plainly here.

A panel follows the same individuals or households over time, so that a change measured between two waves is a change within a person. The UNHS is not built that way. Each round draws a fresh two-stage stratified cluster sample from the national sampling frame, so the households interviewed in 2016/17 are not the households interviewed in 2012/13, and no person in this study appears more than once. The four rounds are four independent cross-sections of the same population, taken at four points in time.

What was built from them is a *pooled repeated cross-section*: the four cleaned rounds stacked into a single file, with a round identifier retained on every record. This structure supports two things. It supports comparing a population quantity across rounds, for example the mean gender wage gap among wage earners in 2012/13 against the same quantity in 2023/24. It also supports fitting one model to all four rounds with round held as a control variable, which

is what the pooled models in Section 3.4 do. It does not support any statement about an individual trajectory, because the data contain no individual trajectories to observe.

The word *longitudinal* is used in the survey literature for both structures, which is a common source of confusion. In this thesis it is used only in the second sense, meaning repeated measurement of the same population over time, and the phrase *pooled multi-round dataset* is used wherever precision matters. Mora and Herrera [24] set out the conditions under which wage gap decomposition is valid on repeated cross-sections, and their treatment is what this study follows. Every temporal statement in Chapters Four and Five should accordingly be read as a statement about the composition of Uganda’s wage-earning workforce at four points in time, and not as a record of what happened to any worker within it.

3.3.2 Study Population and Sampling

The study population comprised individuals aged 14 to 64 in Uganda who reported positive monthly cash earnings from paid work in any of the four UNHS rounds. The 14–64 age band matches the UBOS working-age definition used in the UNHS and the NLFS. Employment status was not filtered separately: a person entered the analysis sample by virtue of reporting a wage, and the exclusion of self-employed workers, unpaid family workers, and subsistence farmers follows from this directly, since none of these categories report a continuous cash wage in the survey instrument. This is a narrower operation than applying a formal ICSE-18 employment-status filter, and the distinction matters for Constraint 6 in Section 3.8: a small number of workers who would be excluded under a strict status-based definition may remain in the sample if they happened to report a wage figure, and the Oaxaca–Blinder decomposition therefore requires only that a wage was reported, not that a formal status test was passed.

No additional sampling was performed. The study used the full analysis sample available from each UNHS round and retained the original UBOS two-stage stratified cluster sampling weights, so that estimates remain nationally representative of the population they describe.

3.3.3 Subgroup Sample Sizes and Data Sufficiency

Once the decomposition was broken down beyond the two main groupings, for example by gender, region, and sector at the same time, individual cells could become too small to estimate reliably, so the sample available at each level of disaggregation had to be established before the results could be trusted. Table 3.1 was the starting point for that assessment: it reported, for each UNHS round, the number of unique households, the number of workers reporting a wage, and the share of those workers who were female.

Table 3.1: Sample sizes of workers reporting a wage, by UNHS round

UNHS Round	Unique HH	Wage Reports	Female %
2012/13	2,095	2,749	34.70%
2016/17	4,008	4,978	37.20%
2019/20	1,910	2,319	34.11%
2023/24	5,483	6,979	38.62%
Pooled total	13,496	17,025	36.96%

Two features of the table shaped everything that followed. The rounds were markedly uneven

in size, ranging from 2,319 wage reports in 2019/20 to 6,979 in 2023/24, so a subgroup that was comfortably large in one round could be too sparse to estimate in another. The female share was also consistently around a third, between 34 and 39 percent, which meant that any cut of the data by gender started from a smaller female base than male base, and that imbalance widened quickly once the data was divided further by region or sector. These two facts were what made the sufficiency measures described below necessary, and the consequences of the second are taken up in full in Section 3.3.4.

Applying the completeness filter and the 14–64 age restriction described in Section 3.3.6 reduced this pooled figure from 17,025 to 14,963 records, an overall retention of 87.9 percent, distributed as 2,283 records for 2012/13, 4,500 for 2016/17, 2,041 for 2019/20, and 6,139 for 2023/24. It is this reduced figure, not the 17,025 in Table 3.1, that entered model training and decomposition.

At the two levels used for the main decompositions, gender pooled across regions within each round, and urban against rural pooled across gender, sample sizes ranged from 2,319 to 6,979 per round before the completeness filter, which is comfortably above what a stable Oaxaca–Blinder estimate requires. The difficulty arises at finer cuts: the gender gap within urban or rural areas considered separately can fall below 500 observations in the smaller rounds, namely 2012/13 and 2019/20.

Five measures kept these finer-grained estimates stable. Where a within-round subgroup fell short, results from neighbouring rounds, for instance 2012/13 and 2016/17, were pooled to restore statistical power, with round fixed effects included to absorb the resulting time difference; this pooling trigger activated below 500 observations in a cell. Categories were also collapsed where the data could not support finer distinctions: education was grouped into four levels, namely none, primary, secondary, and tertiary, rather than finer bands, and sector was analysed at the nine-group ISCO major-group level rather than at a more detailed occupational level. A hard minimum of 200 observations in either the male or the female subsample was enforced as the threshold below which a cell could no longer support the full modelling pipeline. Cells at or above this threshold but still small were flagged and reported with appropriate caveats rather than dropped. Cells falling below it, after pooling adjacent rounds had already been attempted, fell back to LASSO for that cell specifically, since LASSO converges on fewer observations than Random Forest or XGBoost [67, 68]. An absolute floor of 20 observations per group was treated as a hard stop below which a regression is not numerically identified, at which point the cell was skipped entirely rather than estimated. Every decomposition estimate, regardless of cell size, was accompanied by a stratified bootstrap confidence interval, described in Section 3.6, so that the precision behind each figure is stated rather than assumed.

3.3.4 Under-Representation of Women, and How It Was Handled

Women are a minority of this sample. As Table 3.1 shows, the female share of wage reports ranged from 34.1 percent in 2019/20 to 38.6 percent in 2023/24, and stood at 37.0 percent across the pooled file. After the completeness filter and the age restriction, the analysis sample of 14,963 records held 9,728 men and 5,235 women, a female share of 35.0 percent. The imbalance is present in every round and widens at every further cut of the data, so it had to be addressed directly rather than noted and passed over.

The first thing to establish is what the imbalance is. It is not a defect in the UNHS sample design, and it is not a coverage failure that weighting alone could put right. Uganda’s wage-employment sector is genuinely male-dominated, because a large share of working women are in subsistence agriculture, unpaid family work, or own-account trade, none of which reports

a continuous monthly cash wage [69]. A sample of wage earners that was half female would therefore not describe Uganda’s wage earners. The imbalance is a property of the population under study, and the correct response is to carry it and account for it, not to remove it by resampling.

Four measures were used so that the imbalance limited precision rather than distorting the estimates.

Sampling weights were retained. The original UBOS two-stage stratified cluster sampling weights were carried through the pipeline and into estimation, so that population quantities remain nationally representative of wage earners rather than of the achieved sample.

The data split was stratified by sex as well as by round. The 70/15/15 train–validation–test split described in Section 3.4 was stratified jointly by survey round and sex, so each of the three splits holds approximately the same female share as the full file. Without this, a random split could leave the smaller group thinly represented in the test set, and the reported accuracy would then depend partly on the draw.

The decomposition does not require equal group sizes. Oaxaca–Blinder compares group means against a reference wage structure. The Neumark [27] pooled reference used here is estimated on both groups together, which limits the influence of the smaller group’s own coefficients on the benchmark. What unequal group sizes do affect is precision, because the female-side means and predictions rest on fewer observations than the male-side ones.

The sufficiency rules were written around the smaller group. The 500-observation pooling trigger and the 200-observation hard minimum in Section 3.3.3 are applied to the smaller of the two subsamples in a cell, not to the cell total, so a cell with plenty of men and few women cannot pass the test on the strength of the men alone.

The effect on the results is one of precision rather than of level. In practice the female subsample cleared the 200-observation hard minimum in every cell estimated, the smallest being 607 women in the rural gender split for 2016/17, so the LASSO fallback stayed available but was never used. Pooling with an adjacent round was needed for four of the sixteen round-level cells, and all four are gender-by-residence cells for 2012/13 and 2019/20, which are exactly the cells where the female subsample is smallest. Estimates involving the female subsample carry wider bootstrap confidence intervals than those for the two primary splits, and the widening is clearest in those intersectional cells, which is why Tables 4.6 and 4.7 carry pooling flags. The point estimates are not biased by the group-size difference; they are less precisely determined, and the intervals reported alongside them state how much less.

A separate and larger consequence cannot be repaired by weighting, stratification, or pooling, and it is taken up at Constraint 6 in Section 3.8. The women in this sample are those who had already entered wage employment. They are plausibly more educated and more urban than Uganda’s female workforce as a whole, since the women excluded by the wage-reporting criterion are concentrated in the least secure and lowest-paid forms of work. The gender wage gap reported in this thesis is therefore the gap conditional on entry into wage employment, and the economy-wide gap is very likely wider than the figures given here.

3.3.5 Variable Definitions

The dependent variable is the natural logarithm of monthly cash earnings. Taking the log is standard practice in wage analysis: it flattens the right-skewed earnings distribution and lets coefficients be read as approximate percentage effects [18].

Table 3.2 lists the independent variables, each drawn from the UNHS microdata and made consistent across the four rounds.

Table 3.2: Independent variables used in the wage models

Variable	Definition and source
Sex	Male or female, taken from the household roster.
Age; age squared	Age in years, 14–64, plus its square to capture the curved experience–earnings profile predicted by human capital theory.
Education level	Four harmonised tiers: none, primary, secondary, tertiary.
Occupation	ISCO-08 major group (nine categories), aggregated from national occupation codes via correspondence tables built for this study; see Section 3.3.6.
Industry sector	ISIC major division, aggregated from national industry codes via the same correspondence approach.
Urban/rural residence	Binary, taken from the household identification module.
Region	Central, Eastern, Northern, or Western, the four UBOS statistical regions.
Sub-region	A finer geographic split within each region, carried into the model as a label-encoded feature.
Survey round	2012/13, 2016/17, 2019/20, or 2023/24, retained as a control variable in the pooled models and as the time dimension for the cross-round decomposition.

3.3.6 Data Integration and Cleaning

The integration pipeline followed established ex-post harmonisation practice: an inventory of the four contributing rounds, definition of a common target schema, documented harmonisation decisions, and evaluation of the resulting data quality, following the shape of the Maelstrom protocol of Fortier *et al.* [53] and the principle, argued by Kołczyńska [56], that the combination step should be released as executable code rather than described only in prose. Figure 3.1 shows the four stages, each of which is described below.

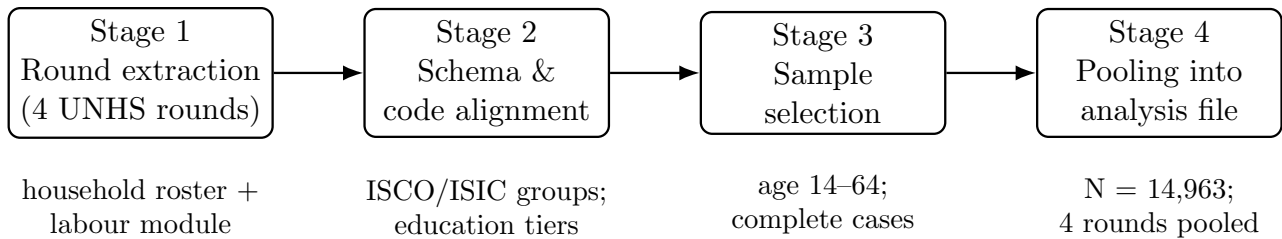


Figure 3.1: The four-stage data integration pipeline

Each of the four UNHS rounds was first loaded and processed independently. The target variables were extracted from the household roster and labour modules using each round’s own variable names, which differ from round to round, and a round identifier was attached to every record so that observations could be tracked through the remaining stages.

Once loaded, each round was mapped onto a common schema. Education levels were standardised to the four tiers in Table 3.2. Occupation and industry codes were aggregated to ISCO-08

major groups and ISIC major divisions respectively, using correspondence tables built for this study that follow the ISCO-08 and ISIC Rev. 4 major-group structure published by the UN Statistics Division [57]. Where a round already supplied a one-digit major-group code, it was used directly. Where only detailed four-digit codes were available, they were aggregated upward through the intermediate classification levels. Aggregating to the major-group level, rather than attempting a revision-by-revision crosswalk, is what makes the four rounds comparable, since detailed occupation and industry codes are not stable across the twelve-year window [58].

Geographic identifiers were retained at the region and sub-region level. District-level identifiers were not usable for harmonisation: they were missing for a majority of records and are not used as model features. Region and sub-region codes that could not be mapped onto the reference frame, which affected a meaningful share of the 2016/17 and 2019/20 rounds, were assigned an explicit **Unknown** category rather than triggering deletion of the record, since dropping on that basis would have removed most of both rounds. The consequence of this choice for interpretation is carried forward to Constraint 5 in Section 3.8.

The analysis sample was then restricted to individuals aged 14 to 64 who reported positive monthly cash earnings, as set out in Section 3.3.2. Two further steps were applied to the dependent variable before the four rounds were pooled. Wages reported below an assumed monthly floor were raised to that floor rather than discarded, using UGX 1,000 for the 2012/13, 2016/17, and 2019/20 rounds and UGX 1,500 for 2023/24, before the log transformation was applied.

Missing values in the four variables central to the analysis, namely wage, sex, age, and education, were handled by complete-case deletion rather than imputation: records missing any of these four were dropped rather than filled in, on the reasoning that imputing a wage determinant risks distorting the wage structure the study set out to estimate [59]. This deletion, together with the age restriction, is what reduced the pooled sample from 17,025 to 14,963 records. Missingness in variables outside this core set was handled differently, through the **Unknown** category described above rather than through deletion.

Finally, the four standardised, cleaned round-level datasets were combined into a single pooled analysis file of 14,963 records, retaining the round identifiers and the original UBOS sampling weights, along with the derived features described next. The log wage and the harmonised occupation and industry group labels were constructed at this stage; interaction effects among the predictors were left for the tree-based models to discover during training rather than specified in advance [18, 42, 43].

3.4 Model Development and Explainability

Four models were trained and compared, ordered from simplest to most flexible. Ordinary Least Squares (OLS) served as the standard benchmark in labour economics, with directly readable coefficients. LASSO, a regularised linear model, performs automatic feature selection and is useful for identifying the most informative predictors among correlated ones [67]. Random Forest, an ensemble of decision trees, captures nonlinear patterns and feature interactions without the analyst specifying them. XGBoost, a gradient-boosted tree ensemble, typically performs well on tabular data [43, 9].

The pooled dataset was split 70/15/15 into training, validation, and test sets, stratified jointly by survey round and sex so that each split preserved roughly the same round and gender composition as the full data; the three resulting index sets were confirmed mutually disjoint before any model was fit. Random Forest and XGBoost were tuned using Bayesian optimisation

via Optuna [62], with 50 trials per model fit on the training set and scored on the held-out validation set, so that the test set played no part in tuning. The LASSO penalty was chosen separately, using 5-fold cross-validation on the training set alone.

Models were compared on RMSE, MAE, and R^2 , reported for all three splits so that any overfitting would be visible rather than assumed away. The selection rule was fixed before the comparison was run: the model with the lowest error on the held-out test set would be the one carried forward into the explainability and decomposition work. That rule, and the question of why it selected XGBoost rather than Random Forest when the two performed so similarly, is discussed in Section 5.3.3.

As a further check on the value of pooling survey rounds, the best-performing model was retrained on 2023/24 alone and on 2019/20 together with 2023/24, and both versions were compared against the model trained on all four pooled rounds, to establish whether pooling more rounds improves generalisation or simply adds noise. This comparison is reported in Chapter Four and speaks directly to Objective 1: it is the empirical test of whether the harmonisation effort was worth undertaking.

For each trained model, SHAP values [10] were computed to quantify how much each feature moved a given prediction up or down. Summary, dependence, and force plots were produced for the full sample and separately for the gender and urban–rural subgroups, which made it possible to see which features drove wages differently for men and women and for urban and rural workers, including nonlinear and interactive effects that a table of OLS coefficients cannot show. Section 3.4.1 sets out how this was implemented and how it connects to the decomposition.

3.4.1 How Explainability Was Implemented

Objective 2 required the models to be explainable as well as accurate. Because the word explainable is used loosely in the wider literature, this section states exactly what it means in this study, what was computed, and how the result connects to the decomposition that follows.

Explainability here is supplied by SHAP and by nothing else. No claim is made that Random Forest or XGBoost can be read directly. SHAP treats a single prediction as the outcome of a cooperative game in which the features are the players, and gives each feature the payoff it earns by joining the coalition, averaged over every order in which the features could have joined [10]. For an individual worker i the prediction then breaks down exactly as

$$f(x_i) = \phi_0 + \sum_{k=1}^K \phi_{ik}, \quad (3.1)$$

where ϕ_0 is the average prediction over the data and ϕ_{ik} is the contribution of feature k to worker i 's prediction, measured in log wage points. Because the parts sum to the whole with no remainder, the attributions can be averaged across workers or across groups without the accounting breaking, and that property is what the decomposition in Section 3.5 relies on.

SHAP values were computed with TreeExplainer [40], which returns exact values for tree ensembles rather than the sampled approximation used for models of arbitrary form. Four outputs were produced:

1. A ranking of average influence, the mean of $|\phi_{ik}|$ over all workers, computed for all four models so that the linear and the tree specifications could be compared on the same scale

(Table 4.3).

2. A global attribution plot for the selected model, in which each worker is one point, horizontal position gives ϕ_{ik} and colour gives the value of the feature itself, so that the direction of an effect is visible and not only its size (Figure 4.2).
3. Dependence plots, which place ϕ_{ik} against the value of feature k and colour by a second feature, so that curvature and interaction can be read off directly (Figure 4.3).
4. Subgroup attributions, in which the same fitted model is explained separately on the female and the male subsample, which shows whether a feature operates differently inside each group (Figure 4.4).

The link between explainability and the decomposition is Equation 3.1 itself. Averaging that identity within each of two groups A and B and subtracting gives

$$\text{explained}_k = \overline{\phi_k | A} - \overline{\phi_k | B}, \quad (3.2)$$

with ϕ_0 cancelling, since it is the same constant for both groups. Summing Equation 3.2 over k returns the explained component of the decomposition exactly. The feature-level breakdown reported for each gap is therefore not a separate exercise carried out alongside the decomposition; it is the same quantity read one feature at a time. Where a cell fell back to LASSO under the rule in Section 3.3.3, the equivalent breakdown has a closed form, $(\bar{x}_A - \bar{x}_B)$ multiplied by the coefficient, and no SHAP computation was needed.

Two limits of the method should be stated plainly. First, SHAP attributes a prediction and not an outcome. A large attribution means the model leaned on that feature, and not that the feature causes wages to move. Second, attributions become unstable when features are strongly correlated, because a contribution shared between two correlated features can be assigned to either of them [11]. Education, occupation, and industry are correlated in the UNHS data, so the ordering of those three in Table 4.3 should be read as a group rather than as a strict ranking.

3.5 Wage Gap Decomposition

The Oaxaca–Blinder decomposition splits the average wage gap between two groups into an explained part, attributable to differences in observable characteristics, and an unexplained part, attributable to differences in the returns to those characteristics together with whatever the model does not observe:

$$\bar{Y}_A - \bar{Y}_B = \underbrace{[\bar{X}_A - \bar{X}_B]\hat{\beta}^*}_{\text{Explained}} + \underbrace{\bar{X}_A[\hat{\beta}_A - \hat{\beta}^*] + \bar{X}_B[\hat{\beta}^* - \hat{\beta}_B]}_{\text{Unexplained}} \quad (3.3)$$

where $\bar{Y}_A$ and $\bar{Y}_B$ are mean log wages for groups A and B , $\bar{X}$ are mean characteristics, $\hat{\beta}$ are estimated returns, and $\hat{\beta}^*$ is the reference wage structure.

Following Bonaccolto–Töpfer and Briel [9], the OLS estimates $\hat{\beta}$ in Equation 3.3 were replaced with predictions from the best-performing ML model for that cell. SHAP values then supplied a feature-level breakdown of the explained part, through the identity derived in Equation 3.2, making it possible to see which specific characteristics mattered most for the gap, and in which direction, at each survey round. The decomposition was run separately for each of the four UNHS rounds, and the resulting path of the explained and unexplained shares over 2012–2024

was plotted and examined to assess whether Uganda’s wage gaps were narrowing, widening, or holding steady.

3.6 Quality Control and Error Management

Three checks were built directly into the pipeline rather than performed as a separate audit afterwards. Every train, validation, and test split was confirmed mutually disjoint by index before model fitting began, which rules out the specific and common error of a record leaking across splits. The Oaxaca–Blinder identity, namely that the explained and unexplained components sum exactly to the observed gap, was checked programmatically for every cell in both the OLS and the ML-augmented versions, with the residual required to fall below a fixed numerical tolerance; this is a direct check that the decomposition arithmetic is correct, not merely plausible. The final saved model was reloaded from disk and its predictions on the test set compared against the predictions of the live, in-memory model, to confirm that the model used for reporting is the one that was actually trained.

Bootstrap confidence intervals were produced for every decomposition estimate, using 200 stratified replications per cell. Two hundred replications, rather than the 1,000 the pipeline is configured to target, were used for the results reported in this thesis, as a deliberate trade-off to keep the full four-round, multi-subgroup run computationally tractable; the code supports increasing this to 1,000 without any other change, and doing so is noted as a direction for future refinement in Chapter Six.

Extreme values in reported wages were addressed only through the minimum-wage flooring described in Section 3.3.6; no separate outlier-fencing procedure was applied to the upper tail of the earnings distribution.

3.7 Ethical Considerations

The study used publicly available UNHS microdata obtained from UBOS. No primary data were collected, and UBOS anonymises the microdata before release, so no individual-level identifiers were present in the data used. The study followed Uganda Christian University (UCU) Research Ethics Committee requirements, and research clearance was obtained from the UCU Faculty of Engineering, Design and Technology before data analysis began.

3.8 Methodological Constraints

Several constraints follow directly from choices made earlier in this chapter, and each is revisited where it becomes relevant to interpreting the results. Two of them, concerning self-reported wages and the meaning of the unexplained component, bear on how the Chapter Four figures should be read, while three others, concerning employment coverage, geographic resolution, and gendered selection into wage employment, are taken up again in the Chapter Five discussion of what the estimated gaps do and do not represent.

The first constraint concerns the wage data themselves. The earnings figures came from what respondents told the interviewer, and are therefore subject to recall error and rounding, particularly for people paid irregularly [12, 13]. A second and related limitation is that the UNHS is a series of repeated cross-sections rather than a panel, as set out in Section 3.3.1.2, so it does not follow the same individuals across rounds; this study could therefore track how the gap

changed in aggregate across the population, but not how any single person’s wage moved over time.

A third constraint arises from who enters the sample at all. The wage-reporting criterion described in Section 3.3.2 leaves out most of Uganda’s workforce, including the self-employed, unpaid family workers, and subsistence farmers, so the true extent of economic inequality in Uganda is almost certainly wider than the results reported here suggest. A fourth constraint concerns how the results are interpreted rather than how they were produced: the unexplained component of the Oaxaca–Blinder decomposition absorbs everything the model does not observe, including ability, motivation, networks, hours worked, job tenure, and measurement error alongside any unequal treatment, so it should not be read as a measure of labour market discrimination.

A fifth constraint concerns geographic resolution. District identifiers could not be harmonised across rounds and were not used, and region and sub-region codes that could not be mapped were assigned an **Unknown** category rather than dropped, as described in Section 3.3.6. Findings at the region level should accordingly be treated as more reliable than any inference drawn at a finer geographic resolution than this study attempted.

The sixth constraint is the most consequential for how the headline result should be understood, because the employment-coverage restriction just described is not gender-neutral. A large share of Uganda’s female workforce works in the informal economy, as self-employed traders, unpaid family labourers, and subsistence farmers, and these are precisely the categories excluded by the wage-reporting criterion, since they do not report a continuous, market-priced cash wage of the kind the Oaxaca–Blinder method requires [69]. Wage employment, by contrast, skews male, which is the imbalance quantified in Section 3.3.4. The female workers captured in this study are therefore not representative of Ugandan women as a whole: they are the subset who had already entered wage employment, plausibly more educated and more urban than the female workforce overall. The gender wage gap reported in this thesis should accordingly be read as the gap conditional on wage-sector entry, not as an estimate of the gap across the whole female workforce. Because the women excluded by this restriction are disproportionately those in the least secure, lowest-paid forms of work, the true economy-wide gender earnings gap is almost certainly larger than what is reported here. Addressing this in future work would require a different outcome measure for informal and self-employment earnings, together with a selection-correction approach, such as a Heckman-style two-step model, to account for the fact that entry into wage employment is itself correlated with sex, education, and location, the very variables central to this decomposition.

3.9 Work Timeline

The research was carried out between May and July 2026. Data acquisition and harmonisation, corresponding to Objective 1, took place in May 2026. Model development, tuning, and SHAP analysis, corresponding to Objective 2, ran from May to June 2026. Oaxaca–Blinder decomposition and temporal analysis, corresponding to Objective 3, were completed in June 2026. Report writing and submission followed in June and July 2026. A detailed Gantt chart is given in Appendix A.

Chapter 4

Results

4.1 Introduction

This chapter reports what the study found, organised around the three objectives set out in Chapter One: what the harmonisation pipeline produced, how the four candidate models performed and what SHAP revealed about wage determination, and what the Oaxaca–Blinder decomposition showed about the gender and rural/urban wage gaps.

4.2 Results

4.2.1 Data Preprocessing Outcomes

The pipeline described in Section 3.3.6 produced a harmonised pooled multi-round dataset spanning the four UNHS rounds, 2012/13, 2016/17, 2019/20, and 2023/24. Alongside the pooled analysis file, the study delivered a harmonisation log covering variable standardisation, the ISCO and ISIC correspondence tables, the treatment of unmappable geographic codes, and the handling of missing values. Table 4.1 and Figure 4.1 show what that process actually did to the data: how many people in each round told the survey what they earned, and how many of those people ended up in the group actually used for the analysis, once the age rule and the completeness check described below were both applied.

Table 4.1: How many people were kept for analysis, by survey round			
Round	Reported a wage	Kept for analysis	Percent kept
2012/13	2,749	2,283	83.1%
2016/17	4,978	4,500	90.4%
2019/20	2,319	2,041	88.0%
2023/24	6,979	6,139	88.0%
Pooled total	17,025	14,963	87.9%

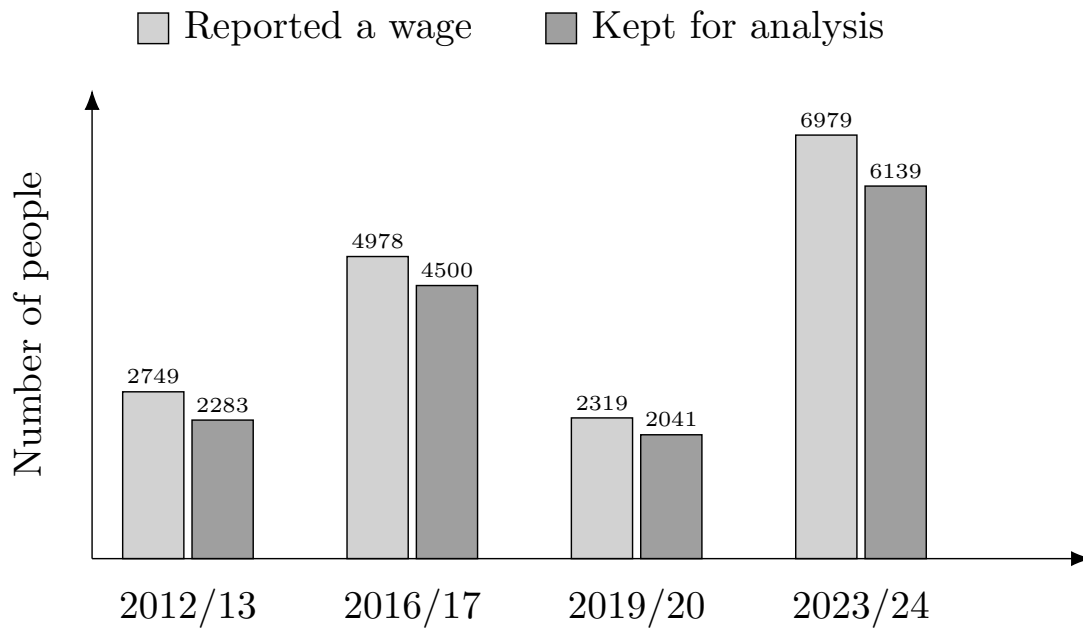


Figure 4.1: How many people were kept for analysis, by survey round

Getting into the "reported a wage" column only needed one thing: could a monthly wage be worked out from what the person told the survey? The UNHS asks how much someone was paid and how often, daily, weekly, monthly, or by the hour, and the pipeline turns that into one monthly figure. If either the amount or how often they were paid was left blank, no monthly wage could be worked out, and that person is not counted here. Nothing else mattered at this stage. A person of any age, even with other details missing, still counted as having reported a wage as long as a monthly figure could be built for them. This is the same group counted as wage reports in Table 3.1 in Chapter Three.

Getting into the "kept for analysis" column required two further conditions, applied in this order. First, the person needed their sex, age, and education all filled in; if any one was missing, they were dropped, since guessing a missing detail like education could distort the very wage patterns the study was trying to measure. Second, only people aged 14 to 64 were kept, matching the working-age definition used throughout this study (Section 3.3.2). Checking these in the other order would not have changed who ended up in the final count, since a person had to pass both either way, but this is the order the pipeline actually used.

How many people were kept varied by round, from 83.1 percent in 2012/13 to 90.4 percent in 2016/17. Both checks were applied together and only the combined result was recorded, so it is not possible to say how much of that difference came from missing details as opposed to people falling outside the 14–64 age range. What can be said is that some rounds started out more complete than others. Overall, 87.9 percent of everyone who reported a wage across all four rounds made it into the final dataset used to train the models and run the decomposition, 14,963 people out of 17,025.

Read against Objective 1, these retention figures carry two meanings. The pipeline kept 87.9 percent of all wage reports, so the harmonised file is not a small, heavily filtered residue of the original surveys, and the four rounds enter the pooled analysis in roughly the proportions in which UBOS collected them. That matters because a harmonisation that discarded a third of its input would raise a question about which third was lost, and whether it was lost at random. The second meaning lies in the variation: retention ranges from 83.1 percent in 2012/13 to 90.4

percent in 2016/17, so the rounds did not arrive equally complete. Ex-post harmonisation work in other survey programmes reports the same pattern, with earlier rounds of a series generally less complete than later ones [53, 56]. The practical consequence here is that the smaller and older rounds carry less weight in the pooled model, which is part of why the round-level decompositions for 2012/13 and 2019/20 needed the sample-sufficiency safeguards described in Section 3.3.3.

4.2.2 Model Performance and Feature Importance

Table 4.2 reports RMSE, MAE, and R^2 for all four candidate models across the training, validation, and test splits. These three metrics were chosen because they answer a question that matters directly for the decomposition work in Section 4.2.3. RMSE and MAE are expressed in log-wage units, so they translate approximately into the typical percentage error in a predicted wage, while R^2 reports how much of the variation in wages the observable characteristics account for. This matters because the ML-augmented decomposition substitutes model predictions for OLS fitted values, so the credibility of that decomposition rests on how accurately the underlying model predicts wages in the first place. A model that fits poorly would carry that error directly into the explained and unexplained shares reported later in this chapter.

Table 4.2: Model comparison across train, validation, and test splits

Model	Split	RMSE	MAE	R^2
OLS	Train	0.9485	0.6324	0.343
OLS	Validation	1.0162	0.6519	0.298
OLS	Test	0.9526	0.6463	0.354
LASSO	Train	0.9485	0.6325	0.343
LASSO	Validation	1.0163	0.6519	0.298
LASSO	Test	0.9528	0.6464	0.354
Random Forest	Train	0.8361	0.5391	0.489
Random Forest	Validation	0.9748	0.6109	0.354
Random Forest	Test	0.9194	0.6059	0.399
XGBoost	Train	0.8138	0.5253	0.516
XGBoost	Validation	0.9689	0.6046	0.362
XGBoost	Test	0.9110	0.6016	0.410

XGBoost had the best test-set score on every metric and was carried forward into the explainability and decomposition work, under the selection rule fixed in Section 3.4. OLS and LASSO performed almost identically across every split; with only ten features and little correlation between them, LASSO’s penalty had almost nothing to shrink. The tree models fit the training data more closely: XGBoost reached a training R^2 of 0.516 against OLS’s 0.343. This gap did not signal overfitting. XGBoost’s test RMSE was only 0.097 above its training RMSE, and its validation and test scores sat close together (RMSE 0.969 against 0.911, R^2 0.362 against 0.410). A genuinely overfit model would have performed noticeably worse on both splits rather than only slightly worse than on training. The tree models’ larger train-to-test gap therefore reflects the expected cost of their extra flexibility rather than poor generalisation, while the

linear models’ near-zero train-to-test gap reflects the opposite problem: a specification too rigid to fit the training data closely in the first place. Across all four models, the ten available features explained roughly 35 to 41 percent of the variance in log wages; the remainder reflects factors the survey does not capture, such as firm size, informal-sector productivity, or exact hours worked.

Whether pooling the four survey rounds actually improved the model, rather than simply adding more rows of similar information, was tested directly rather than assumed. The same tuned XGBoost hyperparameters were used to train three versions of the model, one on 2023/24 alone, one on 2019/20 and 2023/24 combined, and one on all four pooled rounds, and each version was then scored against the same untouched test set spanning all four rounds. This comparison is the direct empirical check on the value of the harmonisation effort reported in Section 4.2.1: if pooling more rounds had not improved test-set accuracy, the twelve years of harmonisation work would have bought nothing beyond convenience. The full comparison and what it means for the case behind pooling the survey rounds is presented in Chapter Five, alongside the discussion of Research Question 1.

SHAP values were then used to rank how much each feature moved XGBoost’s wage predictions, shown in Table 4.3. The method behind these values, and the four outputs produced from them, is set out in Section 3.4.1.

Table 4.3: Mean absolute SHAP value by feature and model

Feature	OLS	LASSO	Random Forest	XGBoost
Education	0.4119	0.4129	0.3216	0.3105
Age	0.6103	0.5797	0.1023	0.1486
Sex	0.1869	0.1865	0.1346	0.1405
Occupation	0.0755	0.0747	0.1330	0.1402
Region	0.1129	0.1124	0.1202	0.1152
Industry sector	0.0341	0.0338	0.0992	0.1024
Urban/rural	0.1273	0.1271	0.1018	0.0951
Survey round	0.1130	0.1126	0.0601	0.0706
Sub-region	0.0128	0.0125	0.0329	0.0591
Age squared	0.4330	0.4049	0.1123	0.0498

Education carried by far the largest average influence, followed by age, sex, and occupation, which is broadly consistent with human capital theory and with the labour market patterns UBOS itself reports [6]. The ranking looked different for the linear models: OLS and LASSO gave age and age squared the largest average influence, with education second. Both tree models pushed education to the top, but they did not treat age itself the same way; Random Forest pushed it down to sixth place, while XGBoost kept it in second, just behind education. Where the two tree models agreed was in demoting age squared, which ranked second or third for the linear models but fell to fifth for Random Forest and last for XGBoost. A table of OLS coefficients could not have shown this: it suggested that age’s relationship with wages was not well captured by a single quadratic term, and instead interacts with other characteristics, most notably education, in a way that only a nonlinear model recovers.

A ranking of average influence establishes which features mattered, but not how they mattered. The remainder of this section reports the SHAP outputs that address the latter question, since

the direction and shape of each feature’s effect is what distinguishes an explainable model from one that merely predicts accurately. Figure 4.2 shows the global attribution for XGBoost, with every point representing one worker in the test sample.

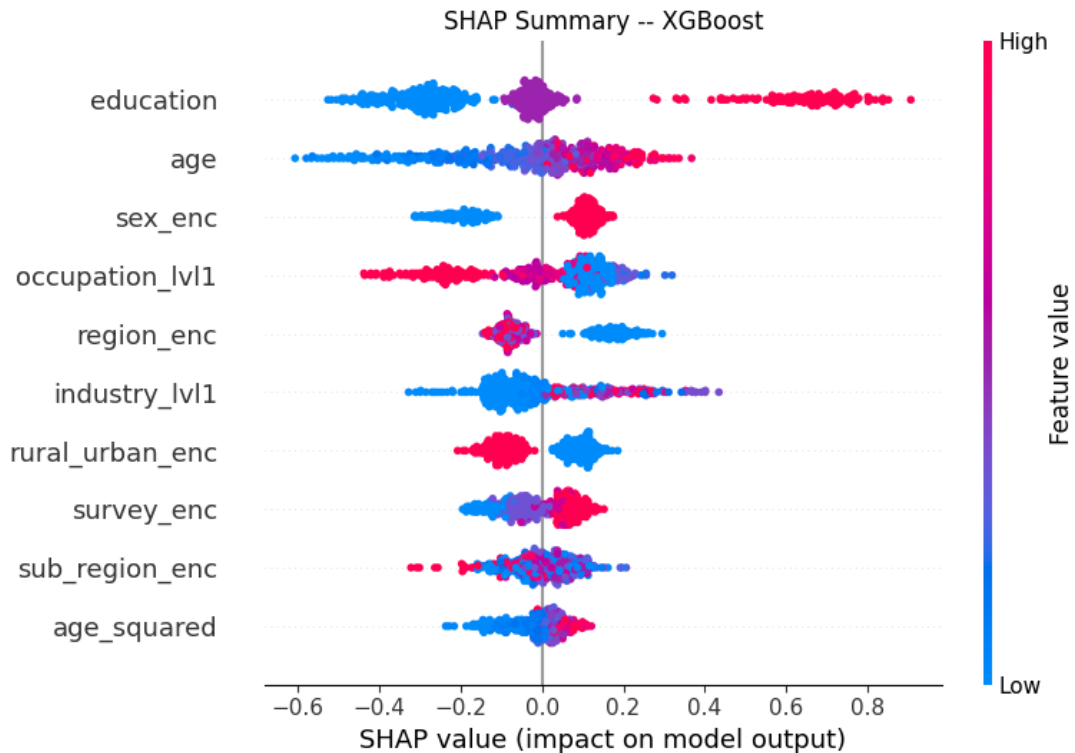


Figure 4.2: Global SHAP attribution for XGBoost. Each point is one worker; position on the horizontal axis gives the effect of that feature on the predicted log wage, and colour gives the value of the feature itself.

The direction of each effect is now visible in a way the summary table could not convey. High education values, shown in red, sit almost entirely to the right of zero and extend furthest from it, meaning that higher education raised predicted wages and did so more strongly than any other feature. The sex variable separates into two distinct clusters on either side of zero, one for each category, which is the visual signature of a systematic group difference rather than a diffuse effect. Urban residence shows the same two-cluster pattern, though the clusters sit closer together.

The dependence plots in Figure 4.3 take this further by showing the shape of two relationships across their whole range.

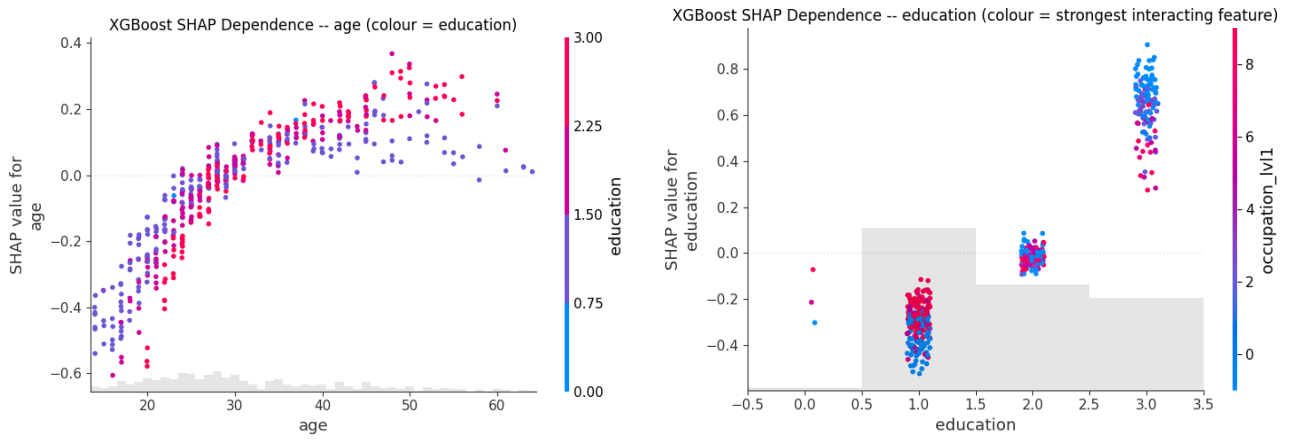


Figure 4.3: SHAP dependence for age, coloured by education (left), and for education, coloured by occupation (right), both for XGBoost.

The left panel recovers the concave age–earnings profile that the Mincer specification assumes: the effect of age on predicted wages rises steeply through the twenties, flattens through the late thirties and forties, and declines slightly thereafter. The model was not told to expect this shape; it emerged from the data. The colouring adds what the quadratic term cannot represent, since at any given age the higher-education points sit consistently above the lower-education ones, indicating that the return to experience was larger for better-educated workers. This is the age–education interaction referred to earlier, shown directly rather than inferred from a ranking.

The right panel shows why the linear specification struggled with education. The effect of education on predicted wages was not proportional to years of schooling: the three lower education tiers cluster near or below zero with little separation between them, while the highest tier jumps sharply upward, contributing between roughly 0.4 and 0.9 log points. A linear term forces a constant increment per tier and cannot represent a threshold of this kind. The colouring by occupation also shows meaningful vertical spread within the highest tier, meaning that the same level of education translated into different wage effects depending on the occupation held, which is the nonlinear interaction structure that the decomposition results in Section 4.2.3 reclassify from unexplained to explained.

Finally, the subgroup plots in Figure 4.4 repeat the global attribution separately for female and male workers, to establish whether the same features operated in the same way within each group.

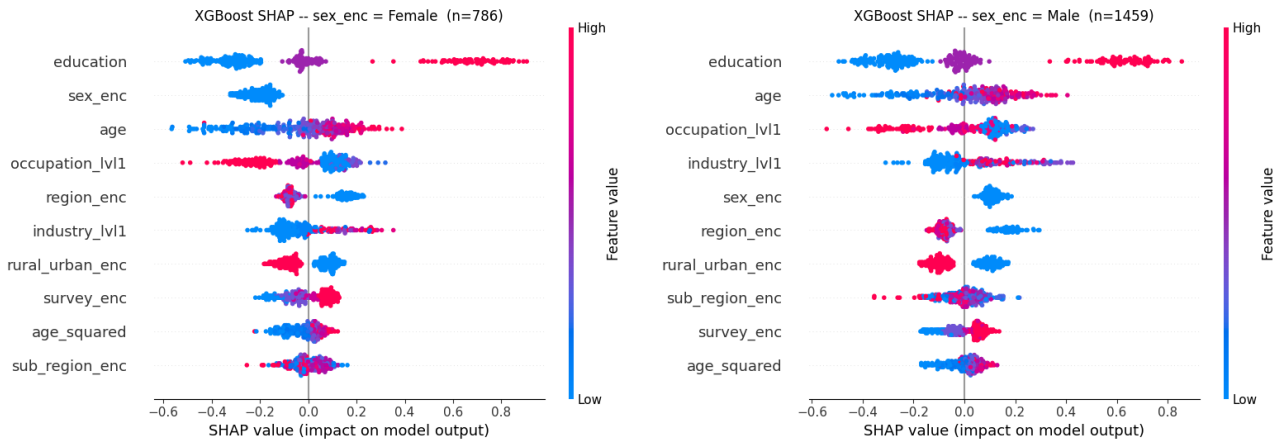


Figure 4.4: Global SHAP attribution for XGBoost, computed separately for female workers (left, $n = 786$) and male workers (right, $n = 1,459$).

Education ranked first for both groups, so the dominant wage driver was common to men and women alike. The ordering below it differed. For female workers, sex ranked second and its points sit almost entirely to the left of zero, meaning that being female reduced the predicted wage relative to the sample average. For male workers, sex fell to fifth place, and industry sector rose to fourth, ahead of where it sat for women. The practical consequence is that a single pooled feature ranking, of the kind reported in Table 4.3, averages over two subgroups whose wage structures were not identical, which is precisely the situation the Oaxaca–Blinder decomposition is designed to handle by estimating separate wage equations for each group.

Two readings follow from Tables 4.2 and 4.3, and both bear on Objective 2.

The first concerns where the gain in accuracy came from. The distance between the two tree models is small, a test R^2 of 0.410 against 0.399, while the distance between either of them and the two linear models is about four times larger. The gain therefore came from allowing curved and interacting relationships at all, and not from the particular boosting algorithm used. That reading is consistent with the argument of Mullainathan and Spiess [37], who hold that the advantage of machine learning in applied economic work lies in flexible estimation of a conditional expectation rather than in any one estimator. It also sets a limit on what this study should claim: the result is about model class, not about XGBoost in particular, and Section 5.3.3 returns to that point.

The second concerns what the models are leaning on. Education ranked first for both tree models, which matches the returns-to-education evidence for developing countries reported by Psacharopoulos and Patrinos [22]. The disagreement between the model families over age squared is the more informative finding. A quadratic in experience is a fixed part of the Mincer specification [18], and both linear models leaned on it heavily, yet both tree models pushed it down the ranking while keeping or raising age itself. Lemieux [21] argued on other data that the Mincer functional form fits modern earnings poorly, and the ranking here is consistent with that argument. Figure 4.3 shows the mechanism: the age profile the models recover is curved, but its height depends on education, and a single squared term applied to the whole sample cannot represent a curve whose shape differs by education level.

4.2.3 Wage Gap Decomposition

The gender and rural/urban wage gaps were decomposed for each of the four survey rounds, twice over: once with the classical OLS method, and once with the ML-augmented method,

using XGBoost predictions in place of OLS fitted values. The gap itself is reported in log points; a gap of g means the higher-earning group earned roughly $(e^g - 1) \times 100$ percent more. The 2012/13 gender gap of 0.450 log points, for example, works out to men earning about 57 percent more than women that round.

For the gender gap, the raw gap ranged from 0.37 to 0.45 log points across the four rounds, broadly consistent with the UN Women estimate of around 32 percent for Uganda [7], as Table 4.4 shows.

Table 4.4: Gender wage gap decomposition by survey round

Round	Gap (log points)	OLS explained	ML explained
2012/13	0.450	−19.8%	52.0%
2016/17	0.388	−5.6%	14.4%
2019/20	0.446	−4.6%	20.9%
2023/24	0.372	5.5%	56.6%
All four rounds together	0.389	−4.1%	29.6%

Two decomposition methods were used, and they told different stories about what was driving the gap. The classical OLS method found a negative explained share in three of the four rounds. In plain terms, observable differences between men and women, such as education or occupation, would on their own predict that women should earn *more* than men, so OLS pushed the entire gap, and then some, into the unexplained part, as low as −19.8 percent explained in 2012/13. The ML-augmented method, run on the exact same data, attributed between 14 and 57 percent of the gap to observable characteristics instead. The reason is that XGBoost can pick up patterns OLS cannot, such as education paying off differently depending on someone’s occupation, and it credits those patterns to the explained part rather than dumping them into the residual. This difference matters: how much of the gender gap looks “explained” depends on which method is used to measure it, not just on the data itself.

The last row of Table 4.4 is different from the four rows above it. It is not an average of those four numbers. Instead, the decomposition was run once more, this time using all 9,728 men and 5,235 women from every round together in a single calculation, with survey round included as a control so the model still knows which year each record came from. Across the whole study period, taken together, education, occupation, and the other observed characteristics accounted for about 30 percent of the gender wage gap under the ML method, and essentially none of it under OLS.

For the rural/urban gap, the two methods agreed far more closely, as Table 4.5 shows.

Table 4.5: Rural/urban wage gap decomposition by survey round

Round	Gap (log points)	OLS explained	ML explained
2012/13	0.799	51.1%	84.4%
2016/17	0.489	55.9%	68.2%
2019/20	0.559	49.4%	74.5%
2023/24	0.462	58.6%	82.6%
All four rounds together	0.520	51.7%	74.1%

The gap itself was large, ranging from 0.46 to 0.80 log points, and OLS already explained 49 to 59 percent of it through observable differences, while the ML-augmented method pushed that up to 68 to 84 percent. Rural and urban workers differ substantially in what jobs they hold, not only in what they are paid for similar jobs, so this gap is well explained by observable characteristics under either method. As with the gender gap, the last row of Table 4.5 is a single calculation run on all 7,275 urban and 7,688 rural workers together, not an average of the four rounds above it; together, those two groups add up to the full 14,963-person dataset used throughout this chapter.

The gender gap was also split further by urban and rural residence, producing two more decompositions: the gender gap measured only among people living in urban areas, and the gender gap measured only among people living in rural areas. The 2012/13 and 2019/20 estimates in both needed round-pooling with an adjacent survey round to reach a stable sample size (see Section 3.3.3), shown as roundsUsed in the tables below.

Table 4.6: Gender wage gap among urban workers, by survey round

Round	Gap (log points)	OLS explained	ML explained
2012/13*	0.534	14.7%	36.8%
2016/17	0.544	22.4%	33.9%
2019/20*	0.538	21.0%	31.6%
2023/24	0.398	9.0%	59.9%
All four rounds together	0.478	11.0%	40.4%

Table 4.7: Gender wage gap among rural workers, by survey round

Round	Gap (log points)	OLS explained	ML explained
2012/13*	0.382	−7.0%	30.0%
2016/17	0.332	−11.7%	14.5%
2019/20*	0.390	−8.7%	15.3%
2023/24	0.384	11.2%	59.0%
All four rounds together	0.388	0.5%	34.7%

*2012/13 was pooled with 2016/17, and 2019/20 was pooled with 2016/17, to reach a stable sample size.

Two patterns stand out from Tables 4.6 and 4.7. First, OLS struggled with the rural gender gap in much the same way it struggled with the gender gap overall. It found a negative explained share in three of the four rural rounds, meaning that, going purely by observable characteristics, women in rural areas should have been paid about the same as men or slightly more. OLS again pushed almost the whole gap into the unexplained part. It did not have this problem in urban areas, where the explained share stayed positive throughout. Second, the ML-explained share was consistently higher among urban workers than rural ones for most of the period, though the gap between the two narrowed sharply over time: in 2016/17 it was 33.9 percent explained among urban workers against 14.5 percent among rural workers, and by 2023/24 the two had converged almost completely, 59.9 percent urban against 59.0 percent rural. Run once on all four rounds together, the same pattern holds at a smaller scale: 40.4 percent explained among urban workers against 34.7 percent among rural workers over the whole period. Taken together, this suggests that whatever kept the rural gender gap harder to explain in the earlier rounds had mostly faded out by the most recent survey.

Every decomposition estimate in the round-by-round rows above came with a bootstrap confidence interval, and the sample-sufficiency diagnostic confirmed that no cell fell below the 200-observation hard minimum even before round-pooling was applied, so the LASSO fallback described in Section 3.3.3 was available as a safeguard but was never actually needed. Only the 2012/13 and 2019/20 estimates of the two gender-by-residence splits, four cells out of sixteen in total, required pooling with an adjacent round to reach the 500-observation pooling trigger.

Each of these confidence intervals was built from 200 stratified bootstrap replications, the runtime-capped figure explained in Section 3.6, rather than the 1,000 the pipeline is configured to target.

Taken together, these results are consistent with the broader pattern reported by Bonaccolto-Töpfer and Briel [9]: allowing for nonlinear interactions reclassifies part of what a linear model calls unexplained as, in fact, explainable structure in the observable characteristics. Kagundu and Pavlova [19] had put Uganda’s unexplained gender gap at 41 to 68 percent using linear methods; the ML-augmented decomposition here reclassified a substantial part of that as explained, though a sizeable unexplained residual remained in every round. Whether that residual reflects unequal treatment, unobserved ability, or something else cannot be settled by this decomposition alone, and is left as an avenue for future research building on these outputs.

Three points follow from Tables 4.4 to 4.7, and they speak to Objective 3.

First, the two gaps are of different kinds. The rural/urban gap behaves as a composition gap: both methods agree that most of it comes from rural and urban workers holding different jobs with different schooling behind them. That is what the near fivefold difference in mean consumption expenditure between Kampala and Karamoja reported by UBOS [6] would lead one to expect, since a difference of that size is more plausibly about where the better-paying work is located than about identical workers being paid differently in two places. The gender gap does not behave that way under OLS, and this is a result about the method rather than about Ugandan women. A negative explained share means that, judged only on the characteristics the linear model can use, women in the sample should have been paid at least as much as men. Taken at face value that is not credible. It is better read as the linear model failing to represent the wage structure than as a statement about what workers bring to the labour market.

Second, the change when the functional form is relaxed is large enough to alter the empirical

conclusion, not merely to refine it. The explained share of the pooled gender gap moves from -4.1 percent to 29.6 percent on identical data. Bonaccolto-Töpfer and Briel [9] reported the same direction of movement on German data; the present results show the same direction on Ugandan data but with a larger swing, which is what one would expect in a market where returns differ more sharply between segments [19]. This is the point at which the premise stated in Section 1.7 is either supported or not, and it is supported.

Third, the shift does not dispose of the unexplained component. It remained the larger part of the gender gap in three of the four rounds, and it was concentrated in rural areas until 2023/24. Kagundu and Pavlova [19] put Uganda’s unexplained urban gender gap at 41 to 68 percent using linear methods on 2002/03 data. The estimates here indicate that part of a figure of that kind is functional form rather than substance, but they do not support the stronger claim that the unexplained component is an artefact. What it contains cannot be settled by this decomposition, and Section 5.4 returns to that question.

4.2.4 Subgroup Results Across Estimation Techniques

The decomposition was not run once on the full dataset. It was run on four subgroup splits, each estimated twice, once with the classical linear method and once with the machine learning augmented method, and each estimated separately for every survey round and once more on all four rounds taken together. Tables 4.4 to 4.7 give the round-level detail. Table 4.8 brings the pooled estimates together so that the two techniques can be compared across subgroups in a single view.

Table 4.8: Explained share of each wage gap by subgroup and estimation technique, all four rounds pooled

Subgroup split	n (A / B)	Gap (log pts)	OLS explained	ML explained
Gender (all workers)	9,728 / 5,235	0.389	-4.1%	29.6%
Urban vs rural (all)	7,275 / 7,688	0.520	51.7%	74.1%
Gender, urban workers only	4,455 / 2,820	0.478	11.0%	40.4%
Gender, rural workers only	5,273 / 2,415	0.388	0.5%	34.7%

Group A is men for the three gender splits and urban workers for the urban/rural split, so a positive gap means men earned more, or urban workers earned more, than the comparison group. The two gender-by-residence rows sum to the urban/rural row: $4,455 + 2,820 = 7,275$ urban workers and $5,273 + 2,415 = 7,688$ rural workers, together the full 14,963-record file. ML estimates use XGBoost, the best-performing model in Table 4.2. Explained shares are the round-pooled estimates, that is a single decomposition run on all four rounds with survey round held as a control, and not an average of the four round-level figures.

Two things stand out when the splits are read side by side rather than one at a time. The technique makes little difference where a gap is driven by composition, and a large difference where it is not. For the urban/rural split the two techniques sit about 22 percentage points apart and agree on the direction and on the broad story. For the gender split they sit about 34 points apart and fall on opposite sides of zero, which is not a difference of degree. Within the gender gap the pattern also depends on where the worker lives: the linear method returns a usable positive explained share among urban workers, at 11.0 percent, but effectively nothing among rural workers, at 0.5 percent, while the machine learning method returns 40.4 and 34.7 percent for the same two groups. The linear method’s difficulty is therefore concentrated in the

subgroup where this study’s most policy-relevant finding sits, which is a reason to treat earlier linear estimates of Uganda’s rural gender gap with particular caution.

One limit of this comparison should be stated rather than left for a reader to discover. The two techniques compared within each subgroup are OLS and the best-performing tree ensemble, not all four estimators from Table 4.2. LASSO was kept in the pipeline as a fallback for any cell falling below the 200-observation minimum in Section 3.3.3, and since no cell reached that threshold, no subgroup was estimated with LASSO. Random Forest was not run through the full four-round, four-split, bootstrapped decomposition, because doing so would have multiplied a runtime already constrained to 200 bootstrap replications in Section 3.6. Section 5.3.3 sets out why substituting Random Forest for XGBoost would be unlikely to change the picture materially, and a complete estimator-by-subgroup grid is recorded as a refinement in Section 6.6.1.

Chapter 5

Discussion

5.1 Introduction

Chapter Four reported what the analysis found. This chapter asks what those findings mean. It follows the three research questions in order: what pooling four rounds of UNHS data made possible that no earlier Ugandan wage study could attempt, what the model comparison and SHAP results say about how wages in Uganda are actually determined, and what the decomposition results say about the gender and rural/urban gaps once the method is allowed to bend. It then asks whether the central premise of this thesis held up, that ML-augmented decomposition transfers from the high-income settings where it was first demonstrated to a data-sparse developing-country survey, and closes by returning to the study’s own methodological limits now that there are results to weigh them against.

5.2 Multi-Round Coverage, Subject to a Cross-Sectional Limit

Research Question 1 asked how four rounds of UNHS microdata, each collected with its own variable definitions, coding schemes, and classification systems, could be brought into one dataset fit for multi-round wage analysis. The answer mattered for more than convenience. Every existing Ugandan wage study reviewed in Chapter Two, Kagundu and Pavlova [19], Sebagala [20], and the UN Women report [7], worked with a single survey round. None of them could say whether the gaps they measured were a snapshot of one particular year or a stable feature of Uganda’s labour market. By building a single pooled dataset spanning 2012/13 to 2023/24, this study made it possible to ask a question the earlier literature simply could not: not just how large the gap was, but whether it was narrowing, widening, or holding steady over time.

Pooling the four rounds turned out to matter for a second, more direct reason as well: it produced a more accurate model. The best model was retrained on progressively fewer rounds and scored each time against the same untouched four-round test set, and Table 5.1 shows what that comparison found.

Table 5.1: Pooling all four survey rounds versus training on fewer rounds (XGBoost, scored on the full four-round test set)

Training rounds	Train rows	Test RMSE	Test R^2
2023/24 only	4,297	1.0020	0.286
2019/20 + 2023/24	5,726	0.9350	0.378
All four rounds (pooled)	10,473	0.9110	0.410

The model trained on 2023/24 alone did reasonably well on its own round, but its test R^2 dropped to 0.286 once it was asked to predict wages from years it had never seen. Adding 2019/20 closed some of that gap, and using all four pooled rounds closed it further still, giving both the lowest RMSE and the highest R^2 of the three. This supports the central premise behind pooling the survey rounds from a second angle: a model trained on more years of data learns wage relationships that hold up across different labour-market conditions, rather than fitting the quirks of whichever round it happened to see. Twelve years of harmonisation work bought both a longer view of the gap and a more accurate model for measuring it.

That answer comes with a genuine limit, and it is worth stating plainly rather than glossing over. The four rounds are repeated cross-sections, not a panel: a different set of households was sampled each time, as set out in Section 3.3.1.2. The pooled dataset can show that the explained share of the gender gap was higher in 2023/24 than in 2016/17 or 2019/20, but it cannot show that any specific person’s circumstances changed over that period. The temporal story told in Chapter Four describes the composition of Uganda’s formal labour force at four points in time; it is not a record of individual trajectories through it.

5.3 What Predicted Wages Best, and Why

Research Question 2 asked which model predicted wages best, and what SHAP revealed about the drivers behind that prediction.

5.3.1 Nonlinearity, Not Algorithm Choice, Drove the Predictive Gain

XGBoost won on every split of the data, but the margin over Random Forest was modest, a test R^2 of 0.410 against 0.399, while the margin over the two linear models was larger, 0.354 for both OLS and LASSO. That pattern is informative in its own right: the real gain from moving to machine learning came from allowing nonlinear relationships and interactions at all, not from any one tree-based algorithm being substantially more capable than another. Once a model was permitted to bend, education’s relationship with wages, and its interaction with age, appeared in a form no linear specification could represent, and that pattern held consistently across both tree-based models.

The R^2 figure also deserves to be read against the existing literature rather than cited only where it flatters the model. Kagundu and Pavlova [19] reported OLS R^2 values of 43.6 percent for urban males and 59.7 percent for urban females using the 2002/03 round. The pooled XGBoost model here reached 41.0 percent, below both figures. The two are not directly comparable, however. Kagundu and Pavlova modelled a single, relatively homogeneous subgroup, urban residents split by sex, in one survey round, while this study pooled the entire working-age formal labour force, urban and rural, across four rounds spanning twelve years of considerable

economic change. A model fitted to a more heterogeneous population over a longer and more turbulent period should be expected to explain a smaller share of the variance even when it is given a more flexible functional form to work with, since the additional variation the model must accommodate is real variation in the underlying population, not noise the model has failed to capture. The lower ceiling observed here is therefore a property of the wider, pooled population this study analysed, not evidence that the ML approach underperformed the earlier OLS work on a comparable task.

5.3.2 Why Sex and Residence Ranked Low in the SHAP Results

The SHAP results complicate a natural assumption about which observable characteristics ought to matter most for a wage-gap study specifically. Sex and rural/urban residence, the two variables of direct interest to Research Question 3, ranked in the middle of the feature list for XGBoost (Table 4.3), well behind education and, for the tree-based models, behind age. This is not a contradiction. A feature can have a large main effect on the predicted wage level, as education and age do, while still carrying most of the information relevant to a group gap through how it is distributed differently across the two groups being compared, rather than through the size of its own SHAP value. That distinction, between a feature’s importance for predicting wages in general and its importance for explaining a gap between two specific groups, is exactly what the Oaxaca–Blinder decomposition in Section 5.4 is designed to isolate, and the two sets of results should be read together rather than treated as competing rankings of the same underlying quantity.

The subgroup attribution plots reported in Section 4.2.2 bear this out directly. When the same model is explained separately for female and male workers, sex rises from fifth position among men to second among women, and its contribution sits almost entirely on the negative side of the scale for the female subsample. The pooled ranking conceals this because it averages the two groups together. The practical implication is that feature importance computed on a pooled sample is the wrong diagnostic for a group-gap question, and the separate wage equations estimated within the decomposition are not merely a technical convention but a response to a difference the explainability results make visible.

The dependence plots in the same section serve a second purpose in this argument. The concave age profile they recover, and the threshold shape of the education effect, are the specific nonlinear structures that the linear specification cannot represent. They therefore supply the mechanism behind the decomposition finding discussed in Section 5.4: the explanatory weight that shifts from the unexplained to the explained component under the ML-augmented method corresponds to structure that is visible in these plots and absent from a linear model by construction.

5.3.3 Why XGBoost Rather Than Random Forest

The two tree ensembles performed almost identically, which raises a fair question: since Random Forest is often described as the more interpretable of the two, why was XGBoost carried forward into the explainability and decomposition work? Three points answer it, and a fourth states what was not tested.

The first is that the selection rule was fixed before the models were compared. Section 3.4 states that the model carried forward would be the one with the lowest error on the held-out test set. XGBoost was better on all three metrics and on all three splits: test RMSE of 0.9110 against 0.9194, test MAE of 0.6016 against 0.6059, and test R^2 of 0.410 against 0.399, with the same ordering on training and on validation. The margin is small, but it is consistent rather

than one metric turning one way and another the other way. Changing the rule after seeing the results would have been the weaker choice, and it would have left the study open to the charge that the reported model was selected to suit the conclusion.

The second is that in this pipeline the two models are equally interpretable, because the interpretation does not come from the model. Random Forest’s reputation for transparency rests on its being an average of decision trees, each of which can be read as a set of if-then rules. That advantage is real for a single shallow tree. It does not survive an ensemble of several hundred tuned trees, which no reader inspects rule by rule. Explainability in this study is supplied by SHAP, computed with TreeExplainer, which returns exact attributions for Random Forest and XGBoost alike and produces the same outputs for each: a ranking of average influence, a signed global attribution, and dependence and subgroup plots. Section 3.4.1 sets this out in full. Judged on the criterion that actually delivered the interpretation in this study, the two models are interchangeable.

The third is that the substantive conclusions do not depend on which of the two was selected. Table 4.3 reports mean absolute SHAP values for all four models. Random Forest and XGBoost agree on the feature that matters most, education, at 0.3216 and 0.3105. They agree in demoting age squared, which both linear models ranked near the top and which falls to fifth for Random Forest and last for XGBoost. They place occupation, sex, region, and industry within a narrow band of one another. The conclusion drawn in Section 5.3.1, that the predictive gain came from allowing nonlinearity rather than from any particular algorithm, rests on precisely this agreement. Had Random Forest been carried forward instead, the model comparison would have named a different winner by 0.011 of R^2 , and the decomposition would have started from predictions that differ little from the ones used here.

What cannot be said is that this reasoning was verified by re-running the full decomposition under Random Forest. It was not. The argument above rests on the closeness of the two models’ test-set performance and on the agreement of their SHAP rankings, both of which are reported in Chapter Four, rather than on a direct comparison of decomposition outputs. That check is a small and inexpensive piece of further work, and it is recorded as such in Section 6.6.1.

5.4 Divergent Findings for the Rural/Urban and Gender Gaps

Research Question 3 asked how the explained and unexplained components of the gender and rural/urban gaps changed across the four waves, and what the ML-augmented method revealed that classical OLS did not. The two gaps told almost opposite stories, and each is worth discussing on its own terms.

The rural/urban gap behaved the way a composition-driven gap should. OLS and the ML-augmented method mostly agreed, 49 to 59 percent explained under OLS against 68 to 84 percent under XGBoost, and both methods agreed the gap itself is large, between 0.46 and 0.80 log points across the four rounds. Rural and urban workers in Uganda hold different jobs, in different sectors, with different education profiles; the decomposition confirmed that this compositional difference, rather than a difference in what the same job pays depending on location, accounts for most of the raw gap. Run once on all four rounds together, the same pattern holds: 51.7 percent explained under OLS and 74.1 percent under the ML-augmented method. This is broadly consistent with the structural picture in the UNHS 2023/24 report itself, which recorded a near-fivefold gap in mean consumption expenditure between Kampala and Karamoja [6]. A gap of that size is far more plausibly a story about where the higher-

paying jobs and sectors are physically located than about identical rural and urban workers being paid differently for identical work.

The gender gap was a different case, and the more analytically significant one in this study. Classical OLS found a negative explained share in three of the four rounds, as low as -19.8 percent in 2012/13, which taken literally would mean that women in the sample held, on average, more of the characteristics OLS associates with higher pay than men did, so the entire raw gap landed in the unexplained residual and then some. This is not a plausible economic finding on its own terms, and it points to a specification problem rather than a substantive result. OLS forces every characteristic's return to follow a single straight line across the whole sample. Where the true relationship bends, for instance if education pays off differently in occupations where men and women are unevenly represented, OLS has no way to represent that bend and pushes it into the unexplained term instead. The ML-augmented method, using the exact same underlying data, attributed between 14 and 57 percent of the same gap to observable characteristics; run once on all four rounds together, it attributed 29.6 percent, against -4.1 percent for OLS. Nothing about the workers or their wages changed between the two decompositions; only the functional form allowed for the returns did. This is the clearest illustration in the whole study of why Bonaccolto-Töpfer and Briel [9] argued that ML-augmented decomposition reclassifies part of what a linear model calls unexplained as, in fact, nonlinear structure in observable characteristics: here, it did so dramatically, reversing the sign of the explained share entirely in three of four rounds.

Figure 5.1 sets the two series side by side across all four rounds, and the shape of the gap between them is itself informative: it does not narrow steadily, but widens and narrows again, tracking economic conditions in each round rather than moving in one direction. OLS and the ML-augmented method are closest in the two middle rounds and furthest apart at the two ends of the study period.

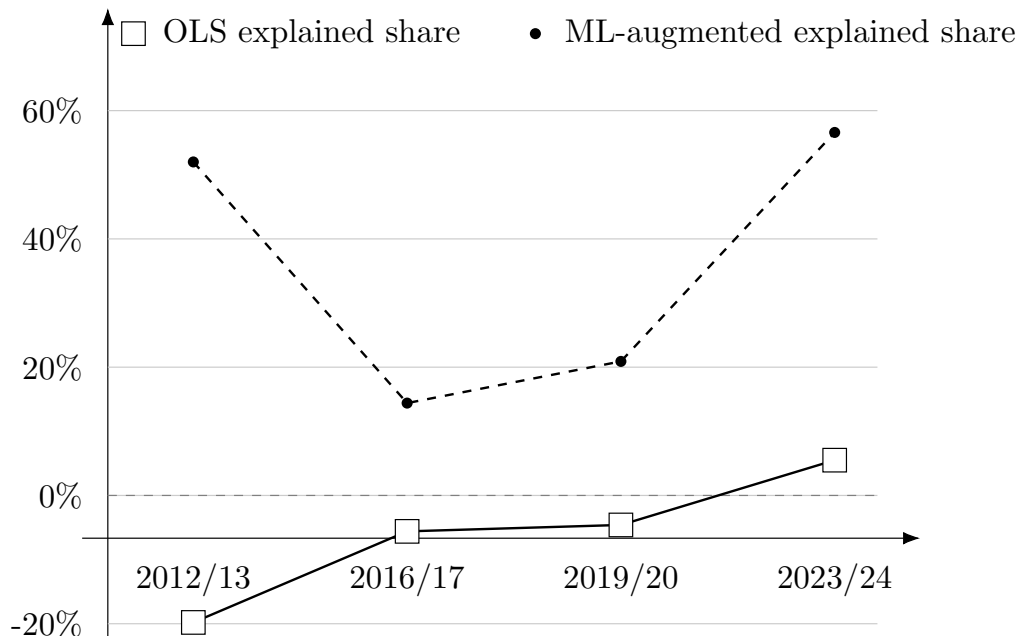


Figure 5.1: Explained share of the gender wage gap, OLS versus ML-augmented method, by survey round

A substantial unexplained residual remained in every round even under the ML-augmented method, and most of it was concentrated in rural areas. The urban/rural split of the gen-

der gap, reported in Section 4.2.3, showed the ML-explained share sitting far higher among urban workers than rural ones for most of the study period, 33.9 percent urban against 14.5 percent rural in 2016/17, converging only by 2023/24. Chapter One’s discussion of Uganda’s cultural and policy context offers a plausible reading of that pattern: kinship-based land and inheritance norms, bride price practices, and patrilocal marriage customs bear most directly on rural women’s economic mobility, and those norms are exactly the kind of thing no household survey variable captures directly. If that reading is correct, the large, persistent rural unexplained residual is not a data limitation to be modelled away with additional features; it may be measuring something real that falls outside what education, occupation, and sector can represent.

None of this licenses calling the unexplained component discrimination outright. As Section 3.8 already established, the residual mixes together ability, motivation, networks, measurement error, and possibly unequal treatment, and this decomposition has no way to separate those out. What the ML-augmented method can responsibly claim is narrower, and for that reason more defensible: a meaningful share of what earlier Ugandan studies using linear methods labelled unexplained, including Kagundu and Pavlova’s 41 to 68 percent estimate and Sebaggala’s roughly 39 percent gap [20], was arguably an artefact of forcing a straight-line relationship onto a wage structure that does not actually behave that way.

What the unexplained component is not. Nothing in this study identifies discrimination, and no finding reported here should be presented as evidence of it. The decomposition separates the part of a gap that tracks measured characteristics from the part that does not, and the second part is a residual rather than a quantity with a single cause. A residual of a given size is consistent with unequal pay for equal work. It is equally consistent with differences in hours worked, job tenure, firm size, bargaining position, career interruption, or accuracy of reporting, none of which the UNHS records in a form this study could use. Establishing discrimination would require a design built for that purpose, such as a correspondence or audit study or a credible natural experiment, and Section 6.6.1 recommends exactly that for the rural gender residual. Statements in this thesis about the gender gap are statements about measurement, not about employer behaviour.

5.5 Transferability, Subject to a Data-Sparsity Cost

Chapter One’s justification for this study rested on a specific, testable premise: that ML-augmented Oaxaca–Blinder decomposition, demonstrated previously in Germany and Switzerland, would transfer to a data-sparse, sub-Saharan African household survey context. The results in Chapter Four support that premise, with a caveat worth stating plainly. The method did transfer: XGBoost outperformed OLS on every split, and the ML-augmented decomposition told a materially different, and arguably more defensible, story about the gender gap than the linear method did. The subgroup-sufficiency safeguards built into Section 3.3.3, the round-pooling trigger, the 200-observation hard minimum, the LASSO fallback, were not a formality. Four of the sixteen round-level decomposition cells, specifically the 2012/13 and 2019/20 estimates for both the gender-within-urban and gender-within-rural splits, needed round-pooling with an adjacent survey round to reach a stable sample size, and those same intersectional splits carried noticeably wider bootstrap confidence intervals than the two primary splits taken on their own. These are the cells in which the female subsample is smallest, as Section 3.3.4 sets out. A data-sparse setting is data-sparse precisely where a study most wants to cut the data finely, by gender, by region, and by sector at once, and that tension did not disappear just because the headline method performed well.

5.6 Chapter Summary

The harmonisation pipeline made a genuinely multi-round Uganda wage analysis possible for the first time. Allowing for nonlinear relationships changed which model won, but it changed the story of the gender gap far more: the explained share of that gap depended heavily on whether the model was permitted to bend, a methodological finding with direct consequences for how earlier single-round, OLS-based Ugandan wage studies should now be read. The rural/urban gap was large but well explained by observable differences under either method. The unexplained share of the gender gap, though reduced by the ML-augmented method, did not disappear, and remained concentrated in rural areas in a pattern consistent with, though not proof of, the cultural and structural barriers described in Chapter One. Chapter Six turns these points into specific recommendations for policy, for the research community, and revisits the study's own limitations in light of what was found.

Chapter 6

Conclusion and Recommendations

6.1 Introduction

This chapter closes the thesis. It summarises what was found against each of the three specific objectives set out in Chapter One, states what the research can and cannot claim to have established, and turns the discussion in Chapter Five into concrete recommendations for policy, for future work, and for the practice of multi-round survey work at UBOS itself.

6.2 Summary of Findings

Objective 1 was to pre-process and harmonise four rounds of UNHS microdata (2012/13 to 2023/24) into a single analysis-ready dataset. This was achieved in four stages. Each round was extracted independently, then its own variable names and codes were mapped onto a common schema. Occupation and industry codes were aggregated to ISCO and ISIC major groups, since the detailed codes were not stable across rounds. A consistent completeness and age filter was then applied before the rounds were pooled.

Three pre-processing decisions shaped the resulting dataset and are worth restating, because each carries a consequence for how the findings should be read. First, records were kept only where all four core variables, wage, sex, age, and education, were present, rather than imputing missing values, on the reasoning that filling in a missing wage determinant would distort the very wage structure the study set out to measure. Applying that rule together with the 14–64 age restriction reduced the pooled file from 17,025 wage reports to 14,963 records, a retention rate of 87.9 percent. Second, wages reported below an assumed monthly floor were raised to that floor rather than discarded, at UGX 1,000 for the first three rounds and UGX 1,500 for 2023/24, which compresses the lower tail of the wage distribution. Third, geographic codes that could not be mapped onto a common frame, affecting a substantial share of the 2016/17 and 2019/20 rounds, were assigned an explicit **Unknown** category rather than triggering deletion, since dropping on that basis would have removed most of both rounds; district-level identifiers proved unusable for harmonisation and were excluded from the feature set entirely.

The result was a dataset of 14,963 formal paid workers drawn from four survey rounds that each used different coding schemes for education, occupation, industry, and geography, accompanied by a harmonisation log documenting every crosswalk and coding decision.

Objective 2 was to train and evaluate explainable machine learning models for predicting monthly wages. XGBoost was the best-performing of the four models tested, reaching a test R^2 of 0.410 against 0.399 for Random Forest, and 0.354 for both OLS and LASSO. Pooling

all four survey rounds measurably improved how well the model generalised compared with training on one or two rounds alone, and SHAP analysis showed that education, followed by age, sex, and occupation, drove most of the model’s predictions, with the tree-based models picking up an age-education interaction that the linear models could not represent.

Objective 3 was to decompose the gender and rural/urban wage gaps using ML-augmented Oaxaca–Blinder techniques across the four survey waves. The rural/urban gap was large (0.46–0.80 log points) and well explained by observable characteristics under both classical and ML-augmented methods (49–59 percent under OLS, 68–84 percent under XGBoost). The gender gap told a sharper story: classical OLS found a negative explained share in three of the four rounds, while the ML-augmented method, applied to the exact same data, attributed 14 to 57 percent of the gap to observable characteristics. The gender gap’s unexplained share remained substantial throughout, and was markedly harder to explain in rural areas than urban ones until the two converged by 2023/24.

6.3 Conclusions

The general objective of this research was to leverage machine learning approaches for modelling, decomposing, and interpreting wage disparities across gender and spatial groups in Uganda, using integrated multi-round survey data and explainable artificial intelligence techniques. That objective was met. The research built a multi-round, ML-augmented wage decomposition for Uganda, extending a method previously demonstrated in Germany and Switzerland to a sub-Saharan African setting, and showed, concretely, that the conclusions a researcher draws about Uganda’s gender wage gap depend heavily on whether the model estimating that gap is allowed to represent nonlinear relationships. That last point is this research’s central finding: it is not simply that XGBoost predicted wages more accurately than OLS, but that the two methods, applied to identical data, produced decomposition results different enough to change the empirical conclusion about how much of the gender gap is explained by observable characteristics at all.

Consistent with the distinction drawn in Section 1.2.1, this is a contribution to data science methodology rather than to social science theory: it demonstrates a better way to measure a wage gap, not a new account of why that gap exists. The research did not identify the causal mechanisms behind Uganda’s wage gaps, and the unexplained share of the gender gap is not read here as a measure of discrimination. What it delivered instead was a better measurement tool, one that showed a substantial part of what earlier linear-method studies attributed to discrimination or unmeasured factors is better read as nonlinear structure in ordinary, observable characteristics. Whether the residual that remains after that reclassification reflects unequal treatment, unmeasured ability, or something else again is a question for the causal research recommended in Section 6.6, not one this decomposition can answer on its own.

What the unexplained component is not. Nothing in this research identifies discrimination, and no finding reported here should be presented as evidence of it. The decomposition separates the part of a gap that tracks measured characteristics from the part that does not, and the second part is a residual rather than a quantity with a single cause. A residual of a given size is consistent with unequal pay for equal work. It is equally consistent with differences in hours worked, job tenure, firm size, bargaining position, career interruption, or accuracy of reporting, none of which the UNHS records in a form this research could use. Establishing discrimination would require a design built for that purpose, such as a correspondence or audit study or a credible natural experiment, and Section 6.6.1 recommends exactly that for the

rural gender residual. Statements in this thesis about the gender gap are statements about measurement, not about employer behaviour.

Uganda's gender and rural/urban wage gaps are real, large, and, by the evidence gathered here, only partly a matter of who holds which job. Allowing the analysis to bend where the data actually bends changed how much of the gender gap could be called explained at all, a result that should give pause to any future Uganda wage study that reaches for a linear model by default. The rural residual that persisted throughout this research, largest exactly where the cultural and structural barriers described in Chapter One would predict it, is the clearest signal in this thesis of where the next piece of research needs to look.

6.4 Limitations

Section 3.8 set out the study's methodological constraints before the results were known. Three of those constraints deserved a second look once the results were in, because of what the study actually found.

The first was the formal-employment restriction, and it cut differently for the two gaps this study measured. Women were disproportionately concentrated in the unpaid, subsistence, and informal work this study excluded, so the gender gap reported here, already large and only partly explained, almost certainly understated the true gender gap in economic outcomes across the whole population. The rural/urban gap was less affected by this restriction, since both excluded and included categories existed in both areas. It was not unaffected either: informal and subsistence work was itself more common in rural areas, so the rural/urban gap measured among formal workers may understate the true gap between rural and urban livelihoods more broadly.

The second was the trade-off between subgroup detail and time resolution, flagged as a general risk in Chapter Three, which turned out to matter most exactly where the results were most interesting: the gender-by-residence intersection. This is also where the female subsample is smallest, as Section 3.3.4 records. The cells that needed round-pooling were the ones showing the clearest and most policy-relevant pattern, a large, converging urban/rural gap in how well the gender gap could be explained, and pooling adjacent rounds to reach a stable sample size necessarily smoothed over whatever happened within that pooled window. The finding that urban and rural gender gaps converged by 2023/24 should be read as a real pattern, but a coarser one than the round-by-round gender and rural/urban results reported separately.

The third was a consequence of the data structure itself rather than of any choice made during analysis. The four UNHS rounds are repeated cross-sections, not a panel: a different set of households was sampled each time. The pooled dataset could show that the explained share of the gender gap was higher in 2023/24 than in 2016/17 or 2019/20, but it could not show that any specific person's circumstances changed over that period. The temporal story this study told is a story about the composition of Uganda's formal labour force at four points in time, not about individual trajectories through it, and the two should not be conflated when the findings are applied.

6.5 Recommendations

6.5.1 For Policy and Practice

The Equal Opportunities Commission and the institutions responsible for monitoring the Employment Act 2006 and the National Gender Policy would gain most from the finding that the rural gender gap was harder to explain by observable characteristics than the urban one for most of the study period. That pattern points enforcement and monitoring attention toward rural labour markets specifically, rather than treating the national gender gap as a single, uniform target. The persistence of a large unexplained gender-gap residual in every round, even after the ML-augmented reclassification, also supports continuing to treat wage-gap monitoring as an ongoing exercise under the Third National Development Plan rather than a problem that recent progress has resolved. That residual should be read as a signal of where measurement is currently weakest and where further investigation is warranted, and not as a quantity of discrimination to be reported.

6.5.2 For Survey Practice at UBOS

Building the harmonisation pipeline surfaced practical friction that is worth recording for anyone extending this work with future survey rounds. Occupation and industry coding changed enough between rounds, ISIC Rev. 3.1 to Rev. 4, and shifting ISCO detail, that a full-time cross-walking effort was needed before any modelling could begin, and geographic identifiers below the four-region level were not consistently comparable across rounds because of repeated district boundary changes. Maintaining a single, versioned crosswalk table across survey rounds, and preserving a stable sub-national geographic identifier alongside whatever administrative boundary is current at the time of each round, would substantially reduce the harmonisation burden facing the next researcher who wants to build a multi-round dataset from UNHS microdata.

6.6 Future Work

6.6.1 Directions for Further Research

Three limitations set out in Section 6.4 point directly to what would strengthen this line of work next. First, because the decomposition here cannot distinguish discrimination from unobserved ability, networks, or omitted variables, a natural next step is a study designed for causal identification, an audit study or a natural experiment, aimed specifically at the rural gender-gap residual this study could not explain. Second, because this study worked with repeated cross-sections, a genuine panel, even a smaller matched sample tracking the same households across two or three rounds, would make it possible to ask whether individual wages moved the way the aggregate composition suggests, rather than inferring individual dynamics from population-level trends. Third, the formal-employment restriction used here excludes the subsistence, informal, and unpaid work in which women are disproportionately concentrated. Extending this framework to informal earnings, where a continuous wage variable is harder to define but not impossible to construct, would give a fuller picture of the true gender gap in economic outcomes, rather than the gap conditional on formal employment reported here.

Fourth, the decomposition reported here was estimated with one nonlinear model. Re-running the full four-round, four-split decomposition under Random Forest as well as XGBoost would establish directly, rather than by inference from test-set performance and SHAP agreement, how far the decomposition results depend on the choice of tree ensemble. The same applies

to a complete estimator-by-subgroup grid covering all four models. Neither requires new data, and both would place the argument made in Section 5.3.3 on a firmer footing.

Fifth, the bootstrap confidence intervals reported in Chapter Four rest on 200 stratified replications rather than the 1,000 the pipeline is configured to target, a trade-off made to keep the full multi-subgroup run tractable. Re-running at 1,000 replications would tighten the reported intervals, particularly for the intersectional cells where the female subsample is smallest.

6.6.2 Further Applications of the Framework

Beyond further research, the outputs of this study open up several downstream applications that sit outside the scope of the thesis itself. The SHAP decompositions point to which observable characteristics, and in which sectors and regions, contribute most to the wage gap. Follow-up work could turn these findings into an empirical basis for targeting enforcement and monitoring under the Employment Act 2006, giving the Equal Opportunities Commission an evidence base more granular than the national headline figure. The temporal decomposition across the four waves showed whether Uganda’s wage gaps narrowed over the study period, and later work could line these trends up against Third National Development Plan targets for reducing gender disparities in labour force participation and earnings, giving national development monitoring a concrete baseline to track against. Finally, this study worked at the national level using survey data, but the SHAP-based decomposition can, in principle, be applied to firm-level payroll data; employers or HR auditors could adapt the pipeline to spot within-organisation pay gaps that need a closer look. These directions are noted here to place the framework within its wider potential uses, but they are not deliverables of this thesis.

6.7 Contribution to Knowledge

This study contributed a measurement method rather than a new explanation of Uganda’s wage gaps. It does not account for why those gaps exist; it establishes a more accurate way of determining how much of them observable characteristics can account for. Three specific contributions follow from that.

The first is methodological. The study demonstrated that ML-augmented, SHAP-explained Oaxaca–Blinder decomposition is viable in a data-sparse, developing-country household survey setting, provided the sample-sufficiency safeguards set out in Section 3.3.3 are built in from the start. Previous demonstrations of the method had been confined to high-income countries with larger and more complete datasets, so its performance under these conditions had not previously been established.

The second is empirical. The study produced a novel multi-round account of how Uganda’s gender and rural/urban wage gaps, and their explained and unexplained components, moved across twelve years and four survey rounds. No previous Ugandan wage study had covered more than a single survey round, so no comparable account of change over time existed.

The third is practical. The study documented a reusable analytical pipeline running from raw UNHS microdata through harmonisation, model training, SHAP explainability, and decomposition. Other researchers can adapt this pipeline to other countries, other survey programmes, or other dimensions of labour market inequality, and the harmonisation decisions it records are the ones most likely to otherwise remain undocumented in a cleaning script. The pipeline is reported here as a research output, and not as a designed artefact under a design science framework; the reasons for that distinction are given in Section 3.2.

Bibliography

- [1] International Labour Organization, “Global wage report 2022–23: The impact of inflation and COVID-19 on wages and purchasing power,” 2022.
- [2] UN Women, “Gender snapshot 2023: Progress on the sustainable development goals,” 2023.
- [3] World Bank, “World development report 2022: Finance for an equitable recovery,” 2022.
- [4] G. Van den Broeck and J. Swinnen, “Structural transformation and gender wage gaps in sub-Saharan africa,” *World Development*, vol. 161, p. 106097, 2023.
- [5] Z. Brixiová and M. Ncube, “Gender gaps in the labour market in Eswatini,” *African Development Review*, vol. 33, no. 2, pp. 323–335, 2021.
- [6] Uganda Bureau of Statistics, “Uganda national household survey 2023/24,” tech. rep., UBOS, Kampala, Uganda, 2025.
- [7] UN Women, “Gender pay gap and labour-market inequalities in Uganda,” tech. rep., UN Women, Nairobi, 2024. Part of the regional study: Why Women Earn Less: Gender Pay Gap and Labour-Market Inequalities in East and Southern Africa.
- [8] A. Strittmatter and C. Wunsch, “The gender pay gap revisited with big data: Do methodological choices matter?,” *IZA Discussion Paper*, no. 14673, 2021.
- [9] M. Bonaccolto-Töpfer and S. Briel, “The gender pay gap and machine learning: New evidence from Germany,” *Labour Economics*, vol. 78, p. 102209, 2022.
- [10] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.
- [11] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Lulu.com, 2nd ed., 2022.
- [12] J. Bound and A. B. Krueger, “The extent of measurement error in longitudinal earnings data: Do two wrongs make a right?,” *Journal of Labor Economics*, vol. 9, no. 1, pp. 1–24, 1991.
- [13] J. C. Moore, L. L. Stinson, and E. J. Welniak, “Income measurement error in surveys: A review,” *Journal of Official Statistics*, vol. 16, no. 4, pp. 331–361, 2000.
- [14] J. W. Scott, “Gender: A useful category of historical analysis,” *The American Historical Review*, vol. 91, no. 5, pp. 1053–1075, 1986.
- [15] N. Kabeer, “Resources, agency, achievements: Reflections on the measurement of women’s empowerment,” *Development and Change*, vol. 30, no. 3, pp. 435–464, 1999.

- [16] World Bank, *World Development Report 2012: Gender Equality and Development*. Washington, DC: World Bank, 2011.
- [17] F. D. Blau and L. M. Kahn, “The gender wage gap: Extent, trends, and explanations,” *Journal of Economic Literature*, vol. 55, no. 3, pp. 789–865, 2017.
- [18] J. Mincer, *Schooling, Experience, and Earnings*. New York: Columbia University Press, 1974.
- [19] P. Kagundu and O. Pavlova, “Gender wage differentials in Uganda: Evidence from the Uganda national household survey,” Working Paper 07-26, Andrew Young School of Policy Studies, Georgia State University, Atlanta, GA, 2007.
- [20] R. Sebaggala, “Wage determination and gender discrimination in Uganda,” Research Series 50, Economic Policy Research Centre (EPRC), Kampala, Uganda, 2007.
- [21] T. Lemieux, “The “Mincer equation” thirty years after *Schooling, Experience, and Earnings*,” in *Jacob Mincer: A Pioneer of Modern Labor Economics* (S. Grossbard, ed.), pp. 127–145, Boston, MA: Springer, 2006.
- [22] G. Psacharopoulos and H. A. Patrinos, “Returns to investment in education: A decennial review of the global literature,” *Education Economics*, vol. 26, no. 5, pp. 445–458, 2018.
- [23] M. J. Lombardía, E. López-Vizcaíno, and C. Rueda, “A new approach to the gender pay gap decomposition by economic activity,” *Journal of the Royal Statistical Society: Series A*, vol. 185, no. 1, pp. 219–245, 2022.
- [24] R. Mora and J. Herrera, “Wage gap decomposition with repeated cross-sections,” *Journal of Applied Econometrics*, vol. 40, no. 1, pp. 1–20, 2025.
- [25] R. Oaxaca, “Male-female wage differentials in urban labor markets,” *International Economic Review*, vol. 14, no. 3, pp. 693–709, 1973.
- [26] A. S. Blinder, “Wage discrimination: Reduced form and structural estimates,” *Journal of Human Resources*, vol. 8, no. 4, pp. 436–455, 1973.
- [27] D. Neumark, “Employers’ discriminatory behavior and the estimation of wage discrimination,” *The Journal of Human Resources*, vol. 23, no. 3, pp. 279–295, 1988.
- [28] J. J. Heckman, L. J. Lochner, and P. E. Todd, “Earnings functions, rates of return and treatment effects: The Mincer equation and beyond,” in *Handbook of the Economics of Education* (E. A. Hanushek and F. Welch, eds.), vol. 1, pp. 307–458, Elsevier, 2006.
- [29] R. Oaxaca and E. Sierminska, “Revisiting the Kitagawa–Oaxaca–Blinder decomposition,” *IZA Discussion Paper*, no. 16145, 2023.
- [30] J. J. Heckman, “Sample selection bias as a specification error,” *Econometrica*, vol. 47, no. 1, pp. 153–161, 1979.
- [31] J. Cotton, “On the decomposition of wage differentials,” *The Review of Economics and Statistics*, vol. 70, no. 2, pp. 236–243, 1988.
- [32] J. DiNardo, N. M. Fortin, and T. Lemieux, “Labor market institutions and the distribution of wages, 1973–1992: A semiparametric approach,” *Econometrica*, vol. 64, no. 5, pp. 1001–1044, 1996.
- [33] J. A. F. Machado and J. Mata, “Counterfactual decomposition of changes in wage distributions using quantile regression,” *Journal of Applied Econometrics*, vol. 20, no. 4, pp. 445–465, 2005.

- [34] H. Ñopo, “Matching as a tool to decompose wage gaps,” *The Review of Economics and Statistics*, vol. 90, no. 2, pp. 290–299, 2008.
- [35] B. Jann, “The Blinder–Oaxaca decomposition for linear regression models,” *The Stata Journal*, vol. 8, no. 4, pp. 453–479, 2008.
- [36] N. M. Fortin, T. Lemieux, and S. Firpo, “Decomposition methods in economics,” in *Handbook of Labor Economics* (O. Ashenfelter and D. Card, eds.), vol. 4A, pp. 1–102, Elsevier, 2011.
- [37] S. Mullainathan and J. Spiess, “Machine learning: An applied econometric approach,” *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 87–106, 2017.
- [38] S. Athey and G. W. Imbens, “Machine learning methods that economists should know about,” *Annual Review of Economics*, vol. 11, pp. 685–725, 2019.
- [39] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, “Double/debiased machine learning for treatment and structural parameters,” *The Econometrics Journal*, vol. 21, no. 1, pp. C1–C68, 2018.
- [40] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, “From local explanations to global understanding with explainable AI for trees,” *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020.
- [41] S. Firpo, N. M. Fortin, and T. Lemieux, “Unconditional quantile regressions,” *Econometrica*, vol. 77, no. 3, pp. 953–973, 2009.
- [42] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [43] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.
- [44] D. Cengiz, A. Dube, A. Lindner, and D. Zentler-Munro, “Seeing beyond the trees: Using machine learning to estimate the impact of minimum wages on labor market outcomes,” *Journal of Labor Economics*, vol. 40, no. S1, pp. S203–S247, 2022.
- [45] M. T. Ribeiro, S. Singh, and C. Guestrin, ““why should I trust you?”: Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.
- [46] D. Weichselbaumer and R. Winter-Ebmer, “A meta-analysis of the international gender wage gap,” *Journal of Economic Surveys*, vol. 19, no. 3, pp. 479–511, 2005.
- [47] S. Appleton, J. Hoddinott, and P. Krishnan, “The gender wage gap in three african countries,” *Economic Development and Cultural Change*, vol. 47, no. 2, pp. 289–312, 1999.
- [48] C. J. Nordman, A.-S. Robilliard, and F. Roubaud, “Gender and ethnic earnings gaps in seven west african cities,” *Labour Economics*, vol. 18, pp. S132–S145, 2011.
- [49] Economic Policy Research Centre, “Labour market transitions of young women and men in Uganda,” tech. rep., Economic Policy Research Centre (EPRC), Kampala, Uganda, 2018. EPRC Research Series.
- [50] Uganda Bureau of Statistics, “Uganda national household survey 2012/13,” tech. rep., UBOS, Kampala, Uganda, 2014.
- [51] Uganda Bureau of Statistics, “Uganda national household survey 2016/17,” tech. rep., UBOS, Kampala, Uganda, 2021.

- [52] Uganda Bureau of Statistics, “Uganda national household survey 2019/20,” tech. rep., UBOS, Kampala, Uganda, 2021.
- [53] I. Fortier *et al.*, “Maelstrom research guidelines for rigorous retrospective data harmonization,” *International Journal of Epidemiology*, vol. 46, no. 1, pp. 103–105, 2017.
- [54] T. M. Smeeding and J. Romich, “The luxembourg income study (LIS): A cross-national data resource for policy-relevant research,” in *International Encyclopedia of Housing and Home*, pp. 246–252, Elsevier, 2012.
- [55] M. Sobek *et al.*, “IPUMS-international: A global census microdata resource,” *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, vol. 44, no. 2, pp. 70–78, 2011.
- [56] M. Kołczyńska, “Combining multiple survey sources: A reproducible workflow and toolbox,” *Political Analysis*, vol. 30, no. 2, pp. 283–299, 2022.
- [57] United Nations Statistics Division, “ISIC rev. 4: International standard industrial classification of all economic activities,” tech. rep., UNSD, New York, 2008.
- [58] World Bank LSMS Team, “A longitudinal cross-country dataset on agricultural productivity and welfare in sub-Saharan Africa,” *Scientific Data*, vol. 12, p. article 899, 2025.
- [59] S. Van Buuren, *Flexible Imputation of Missing Data*. CRC Press, 2nd ed., 2018.
- [60] W. Saeed and C. Omlin, “Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities,” *Knowledge-Based Systems*, vol. 263, p. 110273, 2023.
- [61] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [62] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A next-generation hyperparameter optimization framework,” *arXiv preprint arXiv:1907.10902*, 2019.
- [63] A. R. Hevner, S. T. March, J. Park, and S. Ram, “Design science in information systems research,” *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004.
- [64] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, “A design science research methodology for information systems research,” *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2007.
- [65] J. W. Creswell and J. D. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Thousand Oaks, CA: SAGE Publications, 5th ed., 2018.
- [66] M. Y. Feilzer, “Doing mixed methods research pragmatically: Implications for the rediscovery of pragmatism as a research paradigm,” *Journal of Mixed Methods Research*, vol. 4, no. 1, pp. 6–16, 2010.
- [67] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 1, pp. 267–288, 1996.
- [68] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2nd ed., 2009.
- [69] International Labour Organization, “Women and men in the informal economy: A statistical picture,” tech. rep., International Labour Office, Geneva, 2018.

Appendix A

Research Project Timelines

The research was managed across four phases over three months from May to end of July 2026. Each phase builds on the outputs of the preceding phase, creating a sequential pipeline from raw data to final dissertation.

Phase 1: Data Acquisition and ETL Harmonisation (May 2026)

#	Task	Deliverable	Duration
1.1	Acquire four UNHS microdata rounds from UBOS	Final datasets (4 rounds)	Week 1
1.2	Variable inventory and mapping across rounds	Harmonisation log	Week 1–2
1.3	Schema alignment, education/ISIC/ISCO cross-walking	Standardised round datasets	Week 2–3
1.4	Sample restriction, complete-case filtering, minimum-wage flooring	Pooled analysis-ready dataset	Week 3–4
1.5	Retention and coverage checks across rounds	Harmonisation quality log	Week 4

Dependencies: Tasks 1.2–1.5 depend on successful data acquisition (1.1).

Phase deliverable: Harmonised pooled dataset with documented harmonisation log (Objective 1).

Phase 2: Model Development and Explainability (June 2026)

#	Task	Deliverable	Duration
2.1	Feature engineering and encoding	Final feature matrix	Week 5
2.2	Train/validation/test split (70/15/15, stratified by round and sex)	Training, validation, and test sets	Week 5
2.3	OLS and LASSO baseline models	Baseline performance metrics	Week 5
2.4	Random Forest and XGBoost with Optuna tuning	Tuned ML models	Week 5–7
2.5	Model comparison across rounds	Performance comparison table	Week 7
2.6	SHAP global, subgroup, and temporal analysis	SHAP plots and rankings	Week 7–8

Dependencies: All Phase 2 tasks depend on the pooled dataset from Phase 1 (Task 1.4). Tasks 2.4–2.6 depend on baseline results (2.3) for comparison.

Phase deliverable: Trained models with comparative metrics and SHAP-based feature importance analysis (Objective 2).

Phase 3: Wage Gap Decomposition and Temporal Analysis (Late June – Early July 2026)

#	Task	Deliverable	Duration
3.1	Classical OLS Oaxaca–Blinder decomposition	OLS decomposition results	Week 8
3.2	ML-augmented decomposition (gender gap, 4 rounds)	Gender gap decomposition	Week 8–9
3.3	ML-augmented decomposition (urban–rural gap, 4 rounds)	Spatial gap decomposition	Week 9
3.4	Temporal trend analysis of explained/unexplained shares	Trend charts and tables	Week 9–10
3.5	Robustness checks (reference group, model, sample)	Sensitivity results	Week 10

Dependencies: All Phase 3 tasks depend on the best-performing model from Phase 2 (Task 2.5). Task 3.4 depends on completion of both gender (3.2) and spatial (3.3) decompositions.

Phase deliverable: Decomposed wage gaps across four waves with temporal tracking and robustness analysis (Objective 3).

Phase 4: Final Thesis Writing and Submission (July 2026)

#	Task	Deliverable	Duration
4.1	Draft Chapters 4 and 5 (Results and Discussion)	Draft chapters	Week 10–11
4.2	Draft Chapter 6 (Conclusions and Recommendations)	Draft chapter	Week 11
4.3	Compile full dissertation and internal review	Complete draft	Week 12
4.4	Supervisor review and revisions	Revised dissertation	Week 12–13
4.5	Final submission	Bound copies + soft copy	Week 13

Dependencies: Phase 4 depends on completion of all analytical outputs from Phases 1–3.

Phase deliverable: Final dissertation submitted for examination.

Gantt Chart

Table A.1: Project Gantt Chart: May – July 2026 (13 weeks)

	May				June				July				
	1	2	3	4	5	6	7	8	9	10	11	12	13
Phase 1: Data & ETL (Objective 1)													
1.1 Data acquisition	•												
1.2–1.3 Inventory + schema alignment	•	•	•										
1.4–1.5 Cleaning, QA			•	•									
Phase 2: Models & SHAP (Objective 2)													
2.1–2.3 Features + base-lines					•								
2.4 RF + XGBoost tuning					•	•	•						
2.5–2.6 Comparison + SHAP							•	•					
Phase 3: Decomposition (Objective 3)													
3.1–3.3 OB decomposition								•	•				
3.4–3.5 Trends + robustness									•	•			
Phase 4: Report Writing													
4.1–4.2 Draft chapters										•	•		
4.3–4.5 Review + submit												•	•

• = active task period. Columns represent weeks 1–13 (May–July 2026). Each phase depends on the deliverables of the preceding phase.

Appendix B

Declaration of AI Use

Declaration Statement:

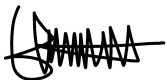
I, *BOB ROBERT TUMUSHIIME*, declare that I used the following AI tools during the preparation of this proposal:

- i. *Claude (claude.ai)* was used to assist with:
 - Formatting and structuring the L^AT_EX document, including table of contents layout, page alignment, and appendix organisation.
 - Drafting and refining academic prose in selected sections, which I then reviewed, revised, and rewritten in my own words to reflect my own understanding.
 - Checking consistency of variable names, dataset references, and citation keys across chapters.
 - Generating initial outlines for the methodology and expected results chapters, which I restructured and rewrote based on my own research design decisions.

Representative prompts included:

- (a) “help me fix the table of contents formatting in my LaTeX proposal”
- (b) “help me structure the methodology chapter for my wage gap study”
- (c) “check if my dataset references are consistent across chapters”
- (d) “help me organise the expected results around my three objectives”

In every case, the AI-generated output served as a starting point rather than a finished product. I reviewed each output against the primary sources (UNHS reports, journal articles, UBOS publications), corrected inaccuracies, rewrote passages in my own voice, and made all substantive decisions including the choice of research design, analytical methods, data sources, and variable selection independently. The AI tool was treated as a drafting and formatting assistant in line with UCU’s Ethical Generative AI Usage policy, not as a replacement for critical thinking or original analysis.

Signature:  _____

Date: 14th September 2026

Appendix C

List of 10 Major Journal Publications Reviewed

Table C.1: Major Publications Reviewed

#	Author(s) & Year	Journal/Source	Type	Relevance	IF
1	ILO (2022)	<i>Global Wage Report 2022-23</i>	Report	Inflation and real wages globally	–
2	Kagundu & Pavlova (2007)	Working Paper 07-26, Andrew Young School of Policy Studies, Georgia State University	Working paper	Gender wage gap in Uganda using OLS decomposition	–
3	Sebagala (2007)	<i>EPRC Research Series</i> No. 50	Report	Wage differentials in Uganda's labour market	–
4	Oaxaca (1973)	<i>Int. Econ. Review</i> , 14(3)	Journal	Original wage decomposition method	2.1
5	Mincer (1974)	<i>Schooling, Experience, Earnings</i>	Book	Human capital earnings function	–
6	Bonaccolto-Töpfer & Briel (2022)	<i>Labour Economics</i> , 78	Journal	ML-augmented Oaxaca–Blinder decomposition	2.8

#	Author(s) & Year	Journal/Source	Type	Relevance	IF
7	Lundberg & Lee (2017)	<i>NeurIPS</i> , 30	Conference	SHAP framework for model explainability	–
8	Chen & Guestrin (2016)	<i>KDD</i> , ACM	Conference	XGBoost gradient boosting algorithm	–
9	Cengiz et al. (2022)	<i>J. Labor Economics</i> , 40(S1)	Journal	ML methods applied to wage analysis	5.2
10	UBOS (2025)	<i>UNHS Report</i> 2023/24	Report	Primary data source for this study	–

Type key: Journal = peer-reviewed journal article; Conference = peer-reviewed conference proceedings; Report = institutional report; Working paper = unrefereed institutional working paper; Book = published monograph. IF = approximate 2024 Impact Factor where available; working papers, reports, books, and conference proceedings do not carry one.

Appendix D

Research Alignment with Programme, National, and Global Agenda

D.1 Alignment with MSc Data Science and Analytics Programme

This research draws on nearly every core area of the MSc Data Science and Analytics programme at Uganda Christian University. The data engineering work harmonising four rounds of household survey microdata that use different occupation codes, industry classifications, and geographic boundaries is the kind of messy, real-world data wrangling that the programme trains students to handle. The modelling component exercises supervised machine learning (Random Forest, XGBoost) and model evaluation skills, while the SHAP analysis puts into practice the programme’s emphasis on interpretable and explainable AI.

What makes this project a good fit for a data science thesis rather than a pure economics thesis is the methodological contribution. The Oaxaca–Blinder decomposition is a well-known econometric tool, but swapping in ML predictions for OLS fitted values is a relatively new idea that sits at the intersection of machine learning and labour economics. Building that bridge, by taking a classical method and extending it with modern ML techniques, is exactly the sort of cross-disciplinary work the programme is designed to encourage.

On a practical level, the study also involved working with data that arrives with gaps, inconsistent coding, and classification changes between rounds. Cleaning and harmonising that data was not a preliminary chore before the real work began. It is a substantial part of the data science contribution itself, and one that future researchers will benefit from through the documented harmonisation log.

D.2 Alignment with National Development Agenda

D.2.1 National Development Plan IV (NDP IV)

Uganda’s Fourth National Development Plan puts human capital development, inclusive economic growth, and income inequality reduction among its top priorities. This study speaks to the human capital pillar directly. It produces figures on what drives wage differences between men and women, and between urban and rural workers, two groups that NDP IV’s equity and social inclusion framework singles out for policy attention.

The decomposition results showed, in concrete terms, how much of the observed wage gap traced back to measurable characteristics such as education and sector of employment, and how much sat in the unexplained residual. That distinction has real policy implications. Where most of the gap is explained, investment in education and skills training is the indicated response. Where a large unexplained portion persists, the drivers lie outside what the survey measures, and identifying them would require research designed for causal inference rather than a further decomposition.

D.2.2 Uganda Vision 2040

Vision 2040 lays out a path toward a middle-income economy with a competitive labour market, less gender inequality, and better rural livelihoods. All three of those goals connect to what this study examined. By establishing which factors, among education, occupation, sector, and location, account for the largest shares of the gender and spatial wage gaps, the research can point policymakers toward the levers most likely to make a difference.

The temporal dimension adds something that most existing Ugandan wage studies lack. With four rounds of data stretching from 2012/13 to 2023/24, the study can examine whether Uganda has been making progress toward Vision 2040's targets for inclusive growth, or whether the gaps have stayed in place across twelve years of economic change.

D.3 Alignment with Global Development Agenda

D.3.1 Sustainable Development Goals (SDGs)

SDG 8 — Decent Work and Economic Growth. Target 8.5 calls for equal pay for work of equal value by 2030. This study measures the wage gaps that stand between Uganda and that target and breaks them down into their component parts. That kind of granular evidence is what policymakers need if they are going to design interventions that work on the specific drivers, rather than broad programmes aimed at inequality in general.

SDG 5 — Gender Equality. The gender wage gap decomposition addresses a specific question: how much of the gap follows from women holding different qualifications and working in different sectors, and how much does not. A large unexplained component shows that observed characteristics do not account for the gap. It does not identify the cause. Unequal treatment is one of several possibilities, alongside differences in hours worked, job tenure, firm size, and reporting accuracy, and separating them would require further research designed for causal inference. On that understanding, the decomposition contributes to the evidence base behind SDG 5's goal of ending wage discrimination against women, without itself measuring discrimination.

SDG 10 — Reduced Inequalities. The urban–rural wage gap analysis documents spatial income disparities within Uganda and identifies the observable factors behind them. Rural workers are not simply paid less; the question is which characteristics account for the difference, and whether those reasons changed over the twelve years covered. The answers feed into the evidence base that SDG 10 requires for designing policies to narrow the gap between rural and urban populations.

D.3.2 African Union Agenda 2063

Agenda 2063 envisions a prosperous continent built on inclusive growth, gender parity, and productive employment for young people. This study contributes to Aspiration 1 (a prosperous

Africa based on inclusive growth) and Aspiration 6 (people-driven development) by generating the kind of detailed, country-level wage inequality evidence that continent-wide policy frameworks often lack.

Although the data comes from Uganda, the method itself is not Uganda-specific. ML-augmented Oaxaca–Blinder decomposition can be applied wherever nationally representative household survey data exist, and thanks to the World Bank’s LSMS programme, that covers most of Sub-Saharan Africa. Since the approach held up here, it offers a template that researchers in neighbouring countries could pick up and adapt with relatively little modification.

Appendix E

Research Poster Layout



Appendix F

Research Ethics Committee (REC) Approval

The signed approval issued by the Uganda Christian University Research Ethics Committee is reproduced below.

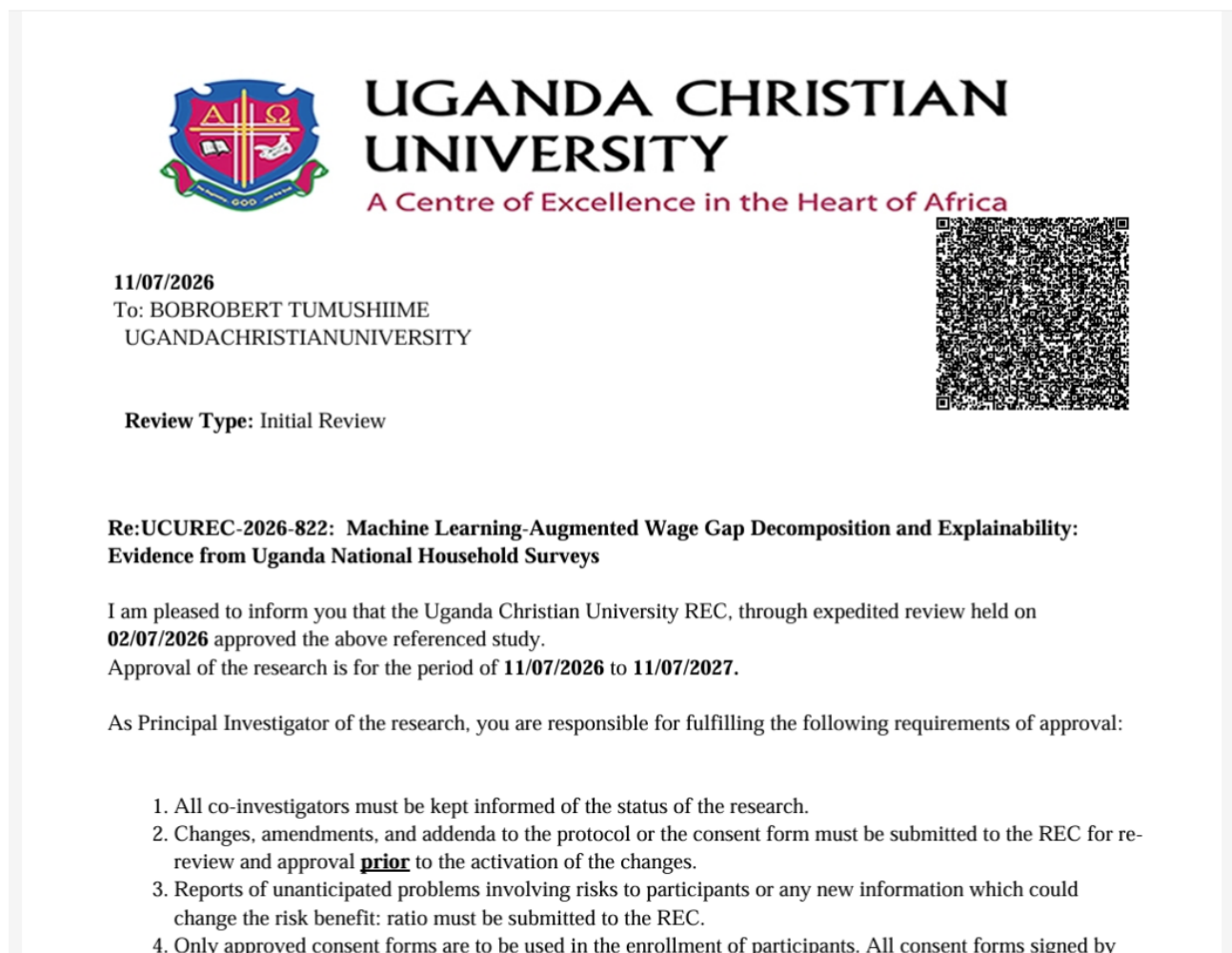


Figure F.1: Research Ethics Committee approval, page 1.

participants and/or witnesses should be retained on file. The REC may conduct audits of all study records, and consent documentation may be part of such audits.

5. Continuing review application must be submitted to the REC **eight weeks** prior to the expiration date of **11/07/2027** in order to continue the study beyond the approved period. Failure to submit a continuing review application in a timely fashion may result in suspension or termination of the study.
6. The REC application number assigned to the research should be cited in any correspondence with the REC of record.
7. You are required to register the research protocol with the Uganda National Council for Science and Technology (UNCST) for final clearance to undertake the study in Uganda.

The following is the list of all documents approved in this application by Uganda Christian University REC:

No.	Document Title	Language	Version Number	Version Date
1	Protocol	English	June 2026	2026-06-06
2	Data collection tools	English	2026	2026-06-01
3	Main Consent forms	English	2026	2026-06-01

Yours sincerely,



Edward Mukooza Kibikyo
For: Uganda Christian University REC

Figure F.2: Research Ethics Committee approval, page 2.

Appendix G

Data Collection Tools

This study relied entirely on secondary data sources. No primary data collection instruments (questionnaires, interview guides, observation checklists) were developed or used. All microdata were obtained from the Uganda Bureau of Statistics (UBOS) through a formal written request to the UBOS Data Archive.

Table G.1: Secondary Data Sources Obtained from UBOS

Dataset	Description	Source	Access Mode	
UNHS 2012/13	Labour force module, earnings, and demographic variables	UBOS	Formal request	written
UNHS 2016/17	Labour force module, earnings, and demographic variables	UBOS	Formal request	written
UNHS 2019/20	Labour force module, earnings, and demographic variables	UBOS	Formal request	written
UNHS 2023/24	Labour force module, earnings, and demographic variables	UBOS	Formal request	written

Data Access Procedure

All microdata was obtained from Uganda Bureau of Statistics through a formal written application stating the research purpose, intended use, variables of interest, and institutional affiliation (Uganda Christian University, MSc Data Science and Analytics programme). The data was used solely for academic research and was not shared with third parties. All datasets received from UBOS were already anonymized and did not contain any personally identifiable information.

Appendix H

Student–Supervisor Meeting/Supervision Plan

Thesis: *ML-Augmented Wage Gap Decomposition and Explainability: Evidence from Uganda National Household Surveys*

1. Particulars

Student name	Bob Robert Tumushiime
Programme	MSc in Data Science and Analytics
Registration No.	J25M19/009
Student No.	B31332
Institution	Uganda Christian University (UCU)
Supervisor name	Dr Kakuba Samuel
Department / Faculty	College of Engineering and Computing
Academic year	2025/2026

2. Purpose of the supervision plan

This plan set out the agreed arrangements for regular supervision between the student and supervisor throughout the thesis. It established a weekly meeting rhythm, defines the responsibilities of each party, and maps meetings to clear milestones and deliverables so that progress can be monitored and any difficulties addressed early. The plan was a working agreement and was revised by mutual consent as the research progressed.

3. Meeting arrangements

Frequency	Once per week (weekly), with additional ad-hoc sessions by arrangement when needed
Proposed day & time	Every Friday, 8:00–9:00 PM
Duration	45–60 minutes
Mode	virtual via Zoom and Google Meet
Location / Platform	meeting link
Review of work	Drafts and deliverables submitted at least 2 working days before each meeting

4. Roles and responsibilities

Student responsibilities

- Prepared and circulated a brief agenda 24–48 hours before each meeting.
- Submitted drafts and deliverables at least 2 working days ahead for review.
- Took and shared meeting minutes / action points within 24 hours.
- Maintained the supervision log and progress tracker.
- Came prepared with specific questions, results, and blockers.
- Completed agreed action items before the next meeting.

Supervisor responsibilities

- Provided timely, constructive feedback on submitted work.
- Guided on methodology, scope, and academic standards.
- Advised on data access, tools, and relevant literature.
- Monitored progress against milestones and flag risks early.
- Approved key deliverables and readiness for submission / defense.
- Supported with departmental and administrative requirements.

5. Ensured effective Communication between meetings

Primary channel	Email — tumbobs@gmail.com / sirkaks2@gmail.com
Expected response time	Within 2 working days
Document sharing	via Email
Urgent matters	Phone / WhatsApp by prior agreement
