Predicting Educational Attainment in Kenya: A Machine Learning Approach Using Socioeconomic and Geographic Data

by

Stella Kemboi
169123

Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Data Science and Analytics at Strathmore University

Strathmore Institute of Mathematical Sciences
Strathmore University
Nairobi, Kenya

June, 2025

Declaration and Approval

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the dissertation contains no material previously published or written by another person except where due reference is made in the dissertation itself.

Student's Name: Stella Kemboi

Sign: SK **Date:** 26/05/2025

The dissertation of Stella Kemboi was reviewed and approved by the following:

Dr. Kennedy Senagi
Postdoctoral Research Fellow - Data Management
International Centre of Insect Physiology and Ecology

Dr. Godfrey Madigu
Dean, Institute of Mathematical Sciences
Strathmore University

Prof. Bernard Shibwabo
Director of Graduate Studies
Strathmore University

Abstract

Educational disparities in Kenya remained a critical challenge, particularly between urban and rural regions and across socio-economic groups. Despite the implementation of FPE and FSE, inequalities persisted, impacting students' ability to complete various educational levels. This study aimed to address this issue by leveraging ML to predict educational attainment using key variables such as household wealth, access to services, and geographic data from the Kenya DHS. A ML model was developed to analyze these socio-economic and geographic factors, providing policymakers with data-driven insights to design targeted interventions aimed at closing educational gaps. By deploying the ML model through a web-based tool, stakeholders were able to identify at-risk regions and populations, leading to more effective resource allocation. The anticipated outcome was a more equitable education system, contributing to Vision 2030 by improving long-term educational outcomes and reducing socio-economic barriers to learning.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Acronyms

AUC-ROC Area Under the Receiver Operating Characteristic Curve.

CRISP-DM Cross Industry Standard Process for Data Mining.

CSV Comma-Separated Values.

DBMS Database Management System.

DHS Demographic and Health Survey.

DT Decision Tree.

EDA Exploratory Data Analysis.

EDM Educational Data Mining.

ERD Entity Relationship Diagram.

FDSE Free Day Secondary Education.

FPE Free Primary Education.

FSE Free Day Secondary Education.

GB Gradient Boosting.

GPS Global Positioning System.

IQR Interquartile Range.

KNBS Kenya National Bureau of Statistics.

LR Logistic Regression.

ML Machine Learning.

NN Neural Network.

RF Random Forest.

SDG Sustainable Development Goals.

SHAP SHapley Additive exPlanations.

SMOTE Synthetic Minority Over-sampling Technique.

SVM Support Vector Machine.

UML Unified Modeling Language.

UNCTAD United Nations Conference on Trade and Development.

UNESCO United Nations Educational, Scientific and Cultural Organization.

Vision 2030 Kenya's National Development Blueprint.

XGBoost eXtreme Gradient Boosting.

Acknowledgements

I am deeply grateful to God for the strength, wisdom, and grace to complete this journey. His unending favor, love, and guidance have been the foundation upon which this work has been built. Without His presence and provision, none of this would have been possible.

I also extend my heartfelt appreciation to my loving husband, Dr. Collins Tanui. Your unwavering belief in me, constant encouragement, and countless sacrifices have carried me through the most challenging moments of this academic pursuit. Thank you for your love, support, and understanding. To our beautiful children — your laughter, hugs, and presence have been a constant source of joy and motivation. You are my greatest inspiration, and I dedicate this work to you and the bright future I hope to help create for you.

Finally, I wish to express my sincere gratitude to my supervisor, Dr. Kennedy Senagi, for his exceptional guidance, patience, and unwavering support throughout this journey. His expertise, insightful feedback, and encouragement have been instrumental in shaping the direction and quality of this research. I am truly privileged to have worked under his mentorship.

Chapter 1: Introduction

1.1 Background

Education is universally recognized as one of the most powerful drivers of economic growth, poverty reduction, and social development. Globally, access to education has significantly improved in the last few decades, contributing to better livelihoods and reduced inequality in many regions. The United Nations Sustainable Development Goals (SDG) (United Nations, 2021), particularly Goal 4: Quality Education, emphasize the need for inclusive and equitable quality education and promote lifelong learning opportunities for all. Despite global commitments, disparities in educational attainment remain prevalent, particularly in low- and middle-income countries, where socioeconomic and geographic barriers hinder access to education.

Across Africa, significant strides have been made to increase school enrollment rates. Countries like South Africa, Nigeria, and Ghana have made efforts to improve access to education through government policies aimed at providing free or subsidized schooling. However, the continent still faces considerable challenges in achieving universal education. According to the United Nations Educational, Scientific and Cultural Organization (UNESCO), Sub-Saharan Africa has the highest rate of education exclusion, with one in five children aged 6-11 and one in three adolescents aged 12-14 not attending school (UNESCO Institute for Statistics, 2020). These numbers highlight the enduring barriers to education, such as poverty, long distances to schools, and gender inequality, which disproportionately affect vulnerable populations.

In East Africa, the situation is similar, with countries like Tanzania, Uganda, and Rwanda implementing policies aimed at expanding access to primary and secondary education. However, despite improvements in enrollment, these countries still face challenges in terms of retention rates and educational quality. For instance, rural areas often lag behind urban centers in terms of school infrastructure, teacher availability, and access to learning mate-

rials. This disparity contributes to lower educational attainment, particularly in rural and underprivileged communities, where children struggle to progress through the various levels of education.

Focusing on Kenya, the government has introduced several initiatives to improve access to education, including the Free Primary Education (FPE) policy in 2003 and the Free Day Secondary Education (FDSE) policy in 2008 (Kenya National Bureau of Statistics, 2021). These efforts have led to a significant increase in school enrollment, with millions of children entering the education system. However, despite these gains, educational attainment—defined as the highest level of education an individual completes—remains highly unequal. The Kenya National Bureau of Statistics (KNBS) reports that only 45% of children from urban areas complete secondary education, compared to a lower 25% in rural areas (Kenya National Bureau of Statistics, 2021). This disparity is driven by broader socioeconomic inequalities, where children from wealthier households are significantly more likely to progress through various levels of education, from primary to tertiary than those from poorer backgrounds.

Kenya's FPE, which seeks to transform the country into a middle-income nation, recognizes education as a key pillar of national development. However, persistent disparities in access to education, particularly in rural areas, continue to threaten the realization of these goals. Rural communities face unique challenges, including long distances to schools, inadequate infrastructure, and a shortage of trained teachers. These obstacles are further compounded by limited access to basic services like electricity, clean water, and sanitation, which are critical to creating a conducive learning environment. The World Bank estimates that 65% of rural households in Kenya still lack access to electricity, and 40% lack adequate sanitation facilities—factors that significantly hinder children's ability to succeed in school (World Bank, 2021).

By applying advanced Machine Learning (ML) techniques, such as Random Forest (RF) and eXtreme Gradient Boosting (XGBoost), this study uncovered the complex relationships between household and geographic factors and educational attainment. The findings from this research have made significant contributions to Kenya's national efforts to close the

educational gap, in alignment with Kenya's National Development Blueprint (Vision 2030). Through data-driven insights, policymakers are able to focus resources more effectively on areas and populations in need, ensuring that all Kenyan students—regardless of their socio-economic background—have the opportunity to achieve their full educational potential.

1.2 Problem Statement

Educational disparities in Kenya, particularly between urban and rural areas and across socio-economic groups, remain a critical challenge. Traditional methods for analyzing educational outcomes rely on manual data collection and descriptive analysis, which are often time-consuming, inefficient, and prone to inaccuracies. These approaches typically focus on short-term academic performance metrics like test scores and graduation rates, offering only a limited view of students' success (Mwangi & Kimenyi, 2019). There is a limited focus on long-term educational progression beyond immediate academic performance which lacks a holistic view of educational trajectories. By concentrating on immediate academic outcomes, these methods fail to account for broader socio-economic and geographic factors that influence long-term educational attainment, particularly in low-resource settings where educational gaps are most pronounced.

ML offers a transformative solution by automating the analysis of complex datasets, allowing for more efficient and holistic predictions of educational attainment. Unlike traditional methods, machine learning can integrate a wide range of socio-economic and geographic variables, providing data-driven insights into the long-term trajectories of students. This approach enabled policymakers to identify students at risk of not progressing through the education system and to design targeted interventions that address both academic and structural barriers to educational success (L. Hasan & Kapoor, 2019),(A. Bhutto & Zafar, 2020). By addressing the inefficiencies and limitations of manual analysis, machine learning presents a more accurate and actionable method for closing educational gaps in Kenya.

1.3 Objectives of the Study

1.3.1 Main Objective

To develop a ML model that automates the prediction of educational attainment in Kenya by analyzing socio-economic and geographic factors, providing accurate, data-driven insights to inform policy interventions and reduce educational disparities

1.3.2 Specific Objectives

(i) Comprehensive review of existing methods for automating educational attainment.

(ii) Comprehend key predictors of educational attainment across socio-economic and geographic groups.

(iii) Apply ML algorithms to automate predictions and overcome manual limitations.

(iv) Generate data-driven insights through interactive visualizations to inform policies aimed at enhancing educational equity.

1.3.3 Research Questions

(i) What methods have been used in prior studies to automate educational attainment analysis using data science??

(ii) What are the most significant socio-economic and geographic predictors of educational attainment in Kenya?

(iii) How accurately can ML models predict educational attainment outcomes?

(iv) How do socio-economic factors contribute to educational disparities?

(v) What data-driven insights can guide policies to improve educational equity?

1.4 Scope and Limitation

This research aligns with Kenya's Vision 2030, which identifies education as a key driver for transforming the country into a middle-income nation by fostering economic growth, reducing poverty, and improving quality of life. The study aimed to contribute to this vision by developing a machine learning model to predict educational attainment across all levels, leveraging socio-economic and geographic data from the Kenya Demographic and Health Survey (DHS). By providing insights into how factors such as household wealth, access to services, and regional disparities influence educational progression, the ML model is empowering policymakers to design targeted interventions that enhance educational equity, particularly in underserved regions. This aligns with UN SDG 4, which calls for inclusive and equitable quality education and the promotion of lifelong learning opportunities for all (United Nations, 2021).

While the study was limited by the DHS dataset, which lacked variables like school quality and teacher effectiveness, it addressed significant gaps in the understanding of educational disparities across Kenya. The inclusion of geographic data, although restricted to GPS coordinates, allowed for a spatial analysis of educational outcomes, which is critical for addressing regional inequalities. In addition, this study responded to the need for automated data-driven solutions that has reduced the reliance on inefficient manual methods, providing evidence-based insights to guide educational policy and resource allocation. By integrating machine learning into this process, the research has contributed to more efficient, scalable solutions for addressing educational challenges in Kenya and beyond, supporting broader global development agendas like the United Nations Conference on Trade and Development (UNCTAD) focus on leveraging technology to enhance educational outcomes in developing regions (United Nations Conference on Trade and Development, 2021).

1.5 Justification

Achieving Kenya's Vision 2030, which seeks to transform the country into a middle-income nation, is fundamentally dependent on improving educational outcomes, particularly in marginalized regions. While education is a powerful tool for reducing poverty, persistent educational disparities—rooted in socioeconomic and geographic factors continue to impede progress. Leveraging machine learning to predict educational attainment is vital for addressing these challenges and developing data-driven policy interventions that can close these gaps.

Most studies in Kenya focus on academic performance, leaving a gap in understanding the long-term factors influencing educational attainment across multiple levels. By integrating socioeconomic and geographic data, this study has offered a comprehensive view of the barriers to educational attainment, especially in rural areas where access to education is more limited. Machine learning provides a data-driven approach to uncovering key predictors of educational success, enabling policymakers to allocate resources more effectively and design interventions that meet the specific needs of underserved populations.

The ultimate beneficiaries of this research are education policymakers, school administrators, and development organizations. By deploying the prediction system, these stakeholders can identify vulnerable regions with poor educational attainment and direct resources or interventions accordingly. This approach promotes more equitable access to education, particularly in underrepresented and marginalized areas. The use of DHS data, based on nationally representative samples, ensured that the findings were widely applicable across Kenya's diverse educational contexts.

Chapter 2: Literature Review

2.1 Introduction

Educational Data Mining (EDM) is a field that focuses on applying data mining and machine learning techniques to analyze and predict educational outcomes. EDM leverages large datasets from academic environments, such as student performance records, socio-economic data, and learning management systems, to uncover patterns that can improve the learning process and address educational disparities (Romero & Ventura, 2020). The increase in digital education platforms has made EDM more relevant, as these platforms generate massive amounts of data that can be analyzed to provide insights into how students learn and how external factors such as socioeconomic background influence their success (Baker & Yacef, 2009).

Machine learning models in EDM have been used to predict a wide variety of educational outcomes, including academic performance, retention rates, and dropouts, making EDM an essential tool for educators and policymakers. By analyzing past educational data, predictive models can identify students at risk of poor academic performance or early dropout, helping institutions implement timely interventions. This study aims to leverage EDM to predict educational attainment in Kenya, focusing on socio-economic and geographic factors that influence the likelihood of completing different levels of education.

2.2 Applications of Machine Learning in Educational Data Mining

ML techniques have been widely applied to predict student outcomes, but many of these studies focus on specific outcomes such as academic performance or dropout rates. For example, (J. Musso & Cascallar, 2020) used ML models to predict academic performance in Vietnam based on socio-demographic factors like family background and school trajectory. While their findings demonstrated that socio-economic variables are strong predictors of

academic success, the study's limitation lies in its narrow focus on performance metrics. It doesn't account for broader educational trajectories or educational attainment across multiple levels. Additionally, the reliance on cognitive and socio-demographic factors without incorporating behavioral or environmental data restricts the depth of insights that could be derived.

(A. Bhutto & Zafar, 2020) applied Support Vector Machine (SVM) and Logistic Regression (LR) to predict student performance based on household income and parental education. While the study achieved an accuracy of 78%, it primarily emphasized accuracy without exploring other important metrics like precision, recall, or F1-score, which could provide a more nuanced understanding of the model's performance. Furthermore, although the study showed that socio-economic background significantly influences educational outcomes, it did not delve into the long-term impact of these socio-economic factors on educational attainment. Its focus on immediate performance outcomes may overlook critical aspects of how students progress through different stages of education.

(J. Xu & Huang, 2019) used Decision Tree (DT) and Neural Network (NN) to analyze student behavior and predict academic performance. Their study incorporated data on internet usage and study habits, reflecting a broader attempt to include non-academic factors in predicting educational outcomes. However, the study's limitation lies in its focus on a single dimension of student behavior—online activity—while ignoring other environmental or social influences that might be equally important. By not incorporating variables like geographic location or access to resources, the study presents an incomplete picture of the factors influencing educational success, particularly in diverse contexts such as Kenya, where geographic and infrastructural factors play a significant role in educational outcomes.

Several studies have applied ML in the Kenyan educational context, although they often focus on academic performance rather than educational attainment. For example, (Mwangi & Kimenyi, 2019) applied RF and LR to predict secondary school performance in several Kenyan counties. While the study provided valuable insights into the predictors of academic success, including parental education and household income, it primarily targeted

secondary school students and did not explore the full educational trajectory. This limits the study's applicability to predicting educational attainment across various levels of education, a broader goal that would require additional factors such as regional disparities or infrastructure availability.

Similarly, (Omondi & Obura, 2020) used NN to predict high school academic performance, focusing on factors like parental involvement and school infrastructure. The study demonstrated that these variables have a significant impact on academic success. However, it did not address how these factors influence long-term educational attainment, particularly in underprivileged regions where resources and infrastructure are lacking. While their neural network approach provided high accuracy, the study would benefit from incorporating rural vs. urban comparisons to highlight the disparities in educational opportunities across different geographic regions.

(E. Ndung'u & Mwangi, 2021) adopted ensemble methods like boosting and bagging to predict school dropout rates in Kenya. The use of ensemble methods was particularly effective in improving the accuracy of predictions. However, while ensemble techniques are useful in minimizing error rates, they also introduce complexity, which may lead to issues with model interpretability—a key concern when the goal is to provide policymakers with actionable insights. Although their model identified significant predictors of dropout rates, such as household income and parental employment, the study did not address other potentially influential factors like access to educational resources or quality of school infrastructure. This limitation reduces the model's ability to provide comprehensive solutions to dropout prevention in contexts where multiple factors are at play.

2.3 Machine Learning Models in Educational Data Mining EDM

In the field of Educational Data Mining EDM, several ML models have been widely applied to predict educational outcomes. Each of these models offers distinct advantages depending on the nature of the data and the complexity of the relationships between variables.

One commonly used model is the decision tree, known for its simplicity and interpretability. Decision trees work by recursively splitting a dataset into smaller subsets based on the most informative features, allowing for straightforward decision-making (L. Hasan & Kapoor, 2019). In the context of predicting educational attainment, decision trees could identify key determinants such as parental education level, household income, or access to school facilities—factors that often play a significant role in determining whether a student completes various stages of education.

Building on decision trees, RF enhance predictive power by averaging the results of multiple decision trees, thereby reducing the likelihood of overfitting. This ensemble method is particularly well-suited to datasets with complex interactions between variables (Breiman, 2001). Given that educational attainment is influenced by a combination of socioeconomic and geographic factors, RF can effectively capture these intricate relationships, especially when handling large datasets with numerous variables, such as those in this study.

SVM is another powerful ML model frequently employed in educational predictions. SVM is particularly effective for high-dimensional datasets and is used to classify data points into distinct categories (A. Bhutto & Zafar, 2020). For example, SVM could be applied to distinguish between students who complete secondary education and those who do not, based on a combination of socio-economic and geographic factors. SVM's ability to handle non-linear decision boundaries makes it highly suitable for educational data, where such factors often interact in complex, non-linear ways.

Additionally, NN have proven to be valuable tools when modeling non-linear relationships between variables. NN can capture complex patterns of interaction between various predictors of educational attainment, such as the interplay between household wealth, parental education, and geographic accessibility (I. Goodfellow, Y. Bengio, and A. Courville, 2016). These models are particularly useful for large, intricate datasets, where relationships between predictors are not easily discernible using linear models.

Furthermore, ensemble learning methods, such as Gradient Boosting (GB) and Stacking, are increasingly being used to improve the accuracy of educational predictions (A. Jayaprakash

& Mandal, 2020). By combining multiple models, ensemble techniques leverage the strengths of each individual model to produce more robust predictions. In the context of predicting educational attainment, ensemble methods are especially beneficial for handling imbalanced data and capturing complex feature interactions, which are often present in socio-economic and geographic datasets.

In sum, these ML models—whether used individually or in combination through ensemble learning—offer a powerful, data-driven approach to predicting educational attainment. By applying these models to socio-economic and geographic data, this study aims to provide a comprehensive analysis of the factors influencing educational outcomes, with the goal of informing targeted interventions to reduce educational disparities.

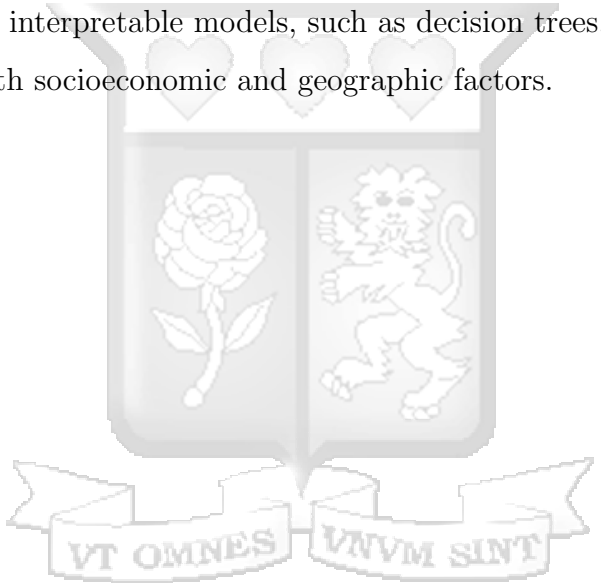
2.4 Research Gaps and Opportunities

A major gap in the existing literature is the focus on academic performance rather than educational attainment. Many studies prioritize short-term outcomes like test scores or graduation rates without considering how students progress through various stages of education. For example, the work by (J. Musso & Cascallar, 2020) and (A. Bhutto & Zafar, 2020) provides insights into immediate academic success but does not address the long-term educational trajectories that are crucial for understanding overall educational attainment. Additionally, the limited use of geographic data in most of these studies overlooks the critical role that regional disparities play in shaping educational outcomes, particularly in developing countries like Kenya.

Interpretability is also key consideration in policy-oriented machine learning applications, where decision-makers require transparent reasoning for predictions. Decision Trees offer high interpretability due to their clear rule-based structure, making them suitable for stakeholder communication. Random Forests, while more accurate, aggregate multiple trees and therefore lose some transparency. XGBoost, though highly performant, presents significant interpretability challenges due to its ensemble structure and complex interactions. While

tools like SHAP enhance post hoc explainability for models like XGBoost, they still require technical expertise to interpret correctly. This trade-off between model performance and interpretability can hinder policy adoption, especially in environments with limited technical capacity.

Moreover, many studies focus on single machine learning models rather than exploring ensemble methods or more complex techniques that could yield better predictive accuracy. While (E. Ndung'u & Mwangi, 2021) did employ ensemble techniques, their study's complexity may hinder its practical applicability for policymakers who require more interpretable models for decision-making. This project aims to address these gaps by applying ensemble methods alongside more interpretable models, such as decision trees, to predict educational attainment based on both socioeconomic and geographic factors.



Chapter 3: Methodology

This study followed the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology to develop a ML model for predicting educational attainment in Kenya. The CRISP-DM framework provides a structured approach to data mining and is particularly well-suited to projects involving large, complex datasets such as the Kenya Demographic and Health Survey (DHS). The methodology was applied through six key phases, each tailored to the specific objectives and context of this study, as outlined below.

1. **Business Understanding**: The study aimed to support education policy in Kenya by predicting educational attainment using socio-economic and geographic data. The objectives were aligned with Vision 2030 and SDG 4 to ensure relevance and impact.

2. **Data Understanding**: The Kenya DHS dataset was explored to understand variable distributions and relationships among key predictors such as household wealth, access to services, and geographic location.

3. **Data Preparation**: Data cleaning, feature engineering, and handling of missing and imbalanced data were carried out. New variables like distance to the nearest school and a household service index were created to improve model accuracy.

4. **Modeling**: Several machine learning algorithms, including Decision Trees, Random Forests, SVM, Logistic Regression, Neural Networks, and XGBoost, were applied to classify individuals into educational levels. Stratified k-fold cross-validation ensured model robustness.

5. **Evaluation**: Models were assessed using accuracy, precision, recall, F1-score, and AUC-ROC to determine their ability to predict educational attainment effectively across different groups.

6. **Deployment**: The final model was deployed via a web-based tool with interactive visualizations, allowing users to explore predicted educational outcomes and regional disparities to inform targeted policy interventions.

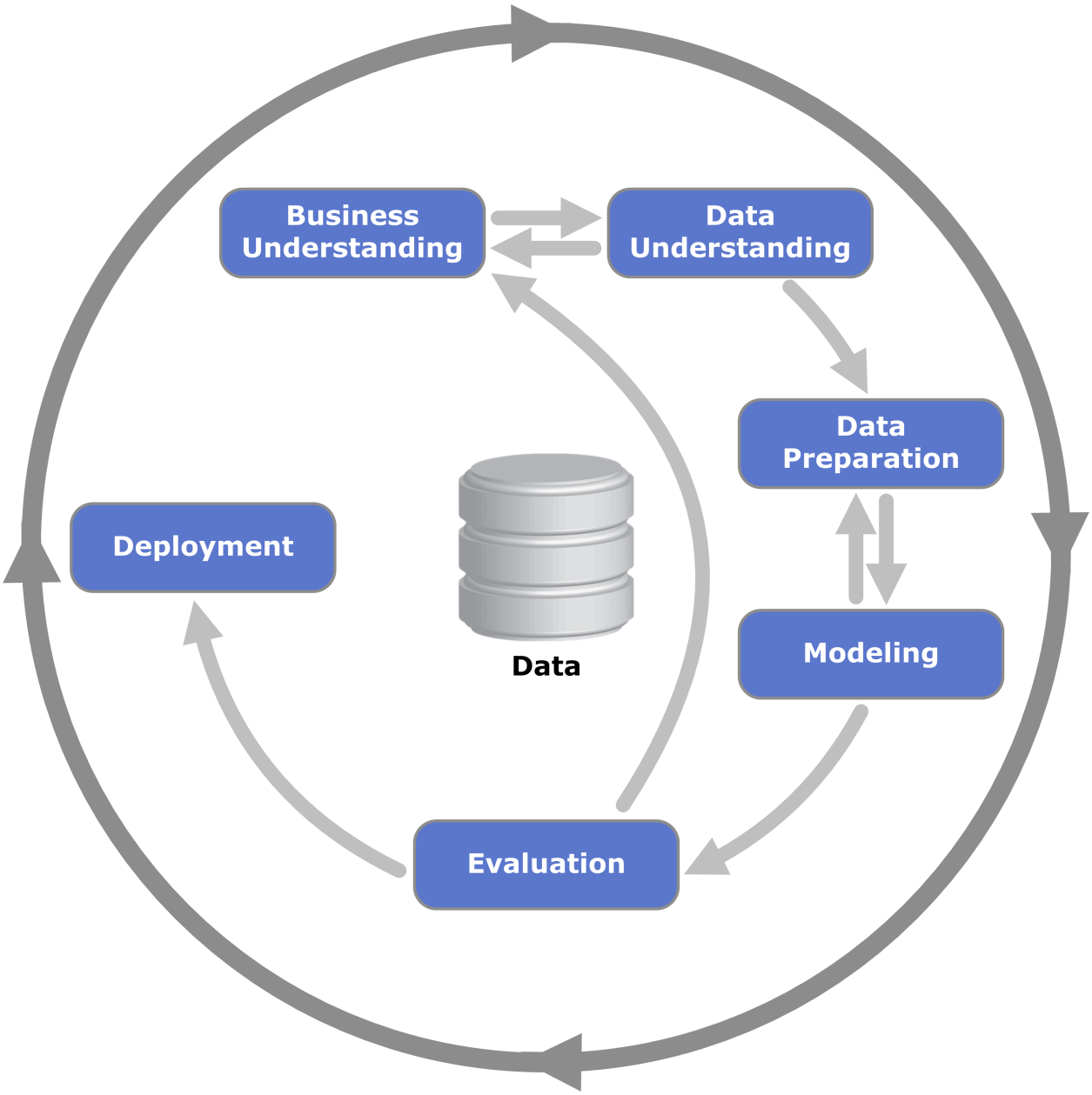


Figure 3.1: CRISP-DM Diagram by Kenneth Jensen, CC BY-SA 3.0, Wikimedia Commons

3.1 Business Understanding

The business understanding stage of this dissertation focused on addressing the persistent educational disparities in Kenya, particularly those influenced by socio-economic and geographic factors. The primary objective was to develop a data-driven solution that supports

policy formulation and resource allocation aimed at improving educational outcomes. This stage involved a detailed examination of the problem context, key stakeholders, and project requirements. The key components of the business understanding included:

1. **Identifying Stakeholders**: It was important to understand the target users and beneficiaries of this work. The stakeholders include education policymakers, government institutions such as the Ministry of Education, non-governmental organizations, and data analysts involved in the education sector. By identifying these users, the project aimed to create a solution that aligns with their decision-making needs and policy goals.

2. **Defining Objectives**: The primary and specific objectives of this study were clearly defined to guide the research process. These objectives, as outlined in Chapter 1, Section 1.3, focused on developing a ML model to predict educational attainment based on socio-economic and geographic factors, and to deliver actionable insights through an interactive visualization platform.

3. **Determining Requirements**: Based on the defined objectives, specific system and data requirements were established. These included access to a nationally representative dataset, specifically the Kenya DHS, selection of relevant predictive variables, integration of Global Positioning System (GPS)-based analysis, and development of a web-based tool. Initial design elements such as data flow diagrams and dashboard mockups were considered to visualize how end-users would interact with the system, ensuring usability and policy relevance.

Overall, this phase provided a foundational understanding of the problem domain, clarified the needs of stakeholders, and established the direction for developing a practical, data-informed tool to support educational planning and equity in Kenya.

3.2 Data Understanding

The data understanding phase of this study involved a systematic exploration of multiple datasets to support the development of a ML model for predicting educational attainment in Kenya. The primary source of data was the 2022 Kenya DHS, which provides nationally representative household-level information across socio-economic, demographic, and health domains. To enhance this dataset with geographic context, additional spatial data were incorporated from the DHS GPS dataset and other external geospatial sources.

3.2.1 Kenya DHS 2022 Dataset:

The individual and household recode files from the 2022 Kenya DHS were explored to extract key socio-economic indicators relevant to educational attainment. These included variables related to household wealth, size, water and sanitation access, media exposure, and demographic characteristics such as the age and sex of the household head. Variable codes from the DHS were mapped to descriptive names using the official DHS recode manuals and documentation to ensure clarity and consistency.

3.2.2 DHS GPS Dataset

The DHS GPS dataset provides geospatial coordinates for each sampled cluster, with slight displacement to preserve participant anonymity. These coordinates were used to spatially link DHS data with geographic features such as towns, schools, and healthcare facilities. The integration of spatial data allowed for the creation of new features—such as distance to the nearest school, town, or health facility—that capture access to essential services and infrastructure.

3.2.3 External Geospatial Data

To derive the distance-based variables, geospatial layers containing the coordinates of schools, towns, and healthcare facilities were obtained from publicly available sources, including the Ministry of Education, OpenStreetMap, and other government datasets. These coordinates were mapped onto the DHS clusters using spatial joins and distance computations performed in GIS software and Python. This enriched the dataset with geographic proximity indicators critical to educational access.

3.2.4 Feature Mapping and Preprocessing

After integrating the spatial and survey datasets, variable names were standardized, and categorical codes were translated into interpretable labels. Initial data profiling and Exploratory Data Analysis (EDA) were conducted to understand the distribution, completeness, and variability of each feature. This helped in identifying potential outliers, missing values, and relationships between predictors and the education level outcome.

Tables 3.1 and 3.2 provide an overview of the predictor variables used in the machine learning model and the classification labels for the target variable, respectively.

Table 3.1: Predictor Variables Used in the Education Level Model

Variable	**Description**	**Data Type**
Wealth Index Score	Composite score representing household economic status	Numerical
Household Size	Total number of individuals in the household	Numerical
Years in Current Place	Duration (in years) the household has lived in the same location	Numerical
Household Head Age	Age of the head of household in years	Numerical
Listens to Radio	Frequency of listening to the radio	Categorical
Internet Access	Availability of internet in the household	Categorical
Internet Usage Frequency	How often household members use the internet	Categorical
Drinking Water Source	Primary source of drinking water	Categorical
Toilet Facility	Type of sanitation facility used by the household	Categorical
Time to Fetch Water (min)	Time (in minutes) taken to fetch water from the main source	Numerical
Distance to School (km)	Approximate distance to the nearest school	Numerical
Distance to Healthcare (km)	Distance to the nearest healthcare facility	Numerical
Distance to Town (km)	Distance to the nearest town or trading center	Numerical

Table 3.2: Target Variable: Education Level

Class Label	Description
0	No Education
1	Primary Education
2	Secondary Education
3	Higher Education

3.3 Data Preparation

Data preparation was a critical phase in this dissertation, as it directly impacted the performance and interpretability of the machine learning models developed to predict educational attainment. This phase involved cleaning, transforming, and integrating both survey and geospatial data into a structured format suitable for analysis and modeling. Given the heterogeneous nature of the data, preparation was conducted in two parallel streams: the Kenya DHS survey data and the spatially derived distance features. Each stream required a unique set of preprocessing techniques to ensure data consistency, completeness, and suitability for training machine learning algorithms.

3.3.1 Kenya DHS Survey Data

(a) **Feature Selection:**

Relevant socio-economic and household-level features were selected from the DHS dataset based on domain knowledge and prior literature. These included variables such as household wealth index, household size, years in current place of residence, household head age, and access to services such as internet, toilet facilities, and safe drinking water.

(b) **Variable Mapping:**

Original DHS variable codes were mapped to meaningful variable names using the

official DHS recode manual. This ensured clarity in both feature interpretation and downstream documentation.

(c) **Handling Missing Values:** Missing values were assessed for each feature. For categorical features, imputation was performed using the most frequent category, while numerical features were imputed using mean and multiple imputation by chained equations (MICE), depending on their distribution and correlation with other variables.

(d) **Encoding Categorical Variables:**

Categorical features such as toilet facility, water source, and media access were transformed into numeric format using ordinal or one-hot encoding based on their nature. This step enabled the models to process categorical inputs effectively.

(e) **Outlier Detection and Treatment:** Outliers in continuous variables (e.g., household size, age of household head) were examined using Interquartile Range (IQR) and z-score methods. Extreme values were either capped or removed based on domain-valid thresholds to maintain data integrity.

(f) **Class Imbalance in Education Level:** The distribution of the target variable—education level—was examined. The dataset exhibited significant class imbalance, with the majority of observations concentrated in the "Primary Education" category, and relatively fewer instances in "No Education" and "Higher Education." To mitigate this imbalance, Synthetic Minority Over-sampling Technique (SMOTE) was applied during model training to synthetically generate samples of the minority classes. Performance was compared with baseline (non-resampled) models to assess the impact of balancing.

3.3.2 Geospatial Distance Features

(a) **Geospatial Integration:**

Using the DHS GPS cluster coordinates, spatial joins were performed to calculate distances from each cluster to the nearest school, healthcare facility, and town. Coor-

dinates for external features were obtained from OpenStreetMap and the Ministry of Education.

(b) **Distance Calculation:**

Haversine distance was used to compute straight-line distances between DHS clusters and nearby infrastructure points. These values were transformed into continuous numerical features: distance to school (km), distance to healthcare (km), and distance to town (km).

(c) **Missing Values and Anomalies:**

Clusters with missing or erroneous coordinates were excluded from the geospatial join process. Additionally, extremely high distances were flagged and removed if they exceeded reasonable rural coverage thresholds, ensuring data relevance.

(d) **Feature Scaling:**

All distance-based variables were normalized using min-max scaling to bring them into a common scale. This improved convergence and interpretability during model training.

(e) **Class Imbalance Considerations:**

Although these geospatial features were not targets, their distribution was assessed in relation to the class imbalance in the education level variable. Exploratory analysis EDA revealed that remote households (with longer distances to schools or towns) were disproportionately represented in lower education categories. This informed the stratified sampling strategy applied during cross-validation to maintain class representation across geographic clusters.

3.4 Machine Learning Modeling

To build a robust predictive model for educational attainment, the cleaned and engineered dataset was split into two subsets: 80% for model training and 20% for testing. The predic-

tion target was the respondent's education level, categorized into four groups: no education, primary, secondary, and higher education. Predictor variables included both household-level socio-economic features and geospatial indicators derived from DHS clusters.

Stratified 10-fold cross-validation was applied to the training set to ensure that each fold maintained the same distribution of education classes, particularly important given the imbalance between the categories. This approach enhanced the reliability of model performance estimates and minimized the risk of overfitting.

Three supervised ML were selected for experimentation, each chosen for its strengths in handling structured data and classification tasks:

a. **Decision Tree (DT)** – A rule-based model that provides insights into feature splits and decision paths.

b. **Random Forest (RF)** – An ensemble method that aggregates multiple decision trees to improve predictive accuracy and reduce variance.

c. **XGBoost (XGB)** – An advanced gradient boosting algorithm known for its speed, accuracy, and regularization capabilities.

To handle class imbalance during training, the SMOTE technique was applied within each cross-validation fold, generating synthetic examples for underrepresented education categories. This ensured that the models were exposed to a more balanced training distribution and improved their ability to predict minority classes.

The models were evaluated on the test set using a range of performance metrics—accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC)—to provide a comprehensive understanding of their strengths and weaknesses in classifying educational levels.

3.4.1 Decision Trees

Decision Trees are transparent and interpretable classification models that recursively partition the dataset into subsets based on input feature values (Loh, 2011; Quinlan, 1986). At each internal node, the algorithm selects a feature and corresponding threshold that best separates the data, typically by minimizing impurity (J. Han & Pei, 2011). In this study, Gini Impurity was used as the splitting criterion, calculated as:

$$Gini = 1 - \sum_{i=1}^{n} p_i^2$$

where p_i denotes the proportion of class i instances at a given node, and n is the number of target classes—in this case, four education levels: No Education, Primary, Secondary, and Higher Education.

The DT algorithm was trained on the preprocessed dataset, which included both socio-economic variables (e.g., wealth index, household size, water and sanitation access) and geospatial indicators (e.g., distance to school, healthcare, and town). The model recursively split the dataset based on these features to classify individuals into the correct education level.

Each terminal node (leaf) of the tree corresponds to a final prediction, representing one of the four education categories. These leaf nodes also contain probability distributions across classes, providing insight into model confidence. For example, a leaf node might assign 70% probability to "Primary Education" and 30% to "Secondary Education", depending on the feature values in that path.

The trained DT revealed that certain features, such as wealth index score, household head age, and distance to school, appeared frequently near the root nodes, indicating their strong influence on predicting education levels. Due to its simplicity and interpretability, the DT model also served as a baseline for explaining individual predictions in later stages of the project.

While DT are prone to overfitting, especially in noisy datasets, early stopping criteria such as maximum depth and minimum samples per leaf were applied to reduce over-complexity (Murthy, 1998). The model's performance was assessed using cross-validation, and its results compared with other classifiers in terms of accuracy, precision, recall, F1-score, and AUC-ROC.

3.4.2 Random Forests

Random Forests are ensemble learning models that construct a collection of decision trees and combine their outputs to improve classification accuracy and robustness (Z.-H. Zhou, 2021). Each tree is trained on a random subset of the original dataset through bootstrap sampling (sampling with replacement). Additionally, at each split in a tree, only a random subset of features is considered, which introduces further diversity among the individual trees and helps to reduce correlation between them.

The final prediction of the ensemble is made by aggregating the outputs of all trees, either through majority voting (for classification) or averaging (for regression):

$$\hat{y} = \text{Mode}(T_1(x), T_2(x), \ldots, T_n(x))$$

where $T_i(x)$ is the prediction of the i-th tree in the forest.

In this study, the RF model was applied to predict educational attainment across four levels: No Education, Primary, Secondary, and Higher Education. The model was trained using the full set of socio-economic and geographic features, including wealth index, household demographics, access to services, and proximity to infrastructure. Due to its ability to handle nonlinear relationships and feature interactions, RF were particularly well-suited to the heterogeneous nature of the DHS dataset (Wang & Sun, 2020).

To address the issue of class imbalance in the target variable, oversampling using the SMOTE technique was incorporated during cross-validation, ensuring that minority classes were adequately represented during training (Haixiang et al., 2019). The model's hyperparameters,

such as the number of trees, maximum depth, and minimum samples per split, were optimized to balance performance and overfitting.

Feature importance scores were extracted from the trained model, providing insight into the most influential predictors of educational attainment. Variables such as wealth index score, distance to school, and household head age consistently ranked among the top contributors, reaffirming findings from the Decision Tree analysis.

The RF model achieved strong performance across evaluation metrics, offering a balance of accuracy, generalization, and interpretability, which made it a key candidate for deployment within the decision-support tool.

3.4.3 XGBoost

Extreme Gradient Boosting (XGBoost) is a high-performance, ensemble learning algorithm based on gradient boosting decision trees (Chen & Guestrin, 2020). Unlike bagging methods like RF, XGBoost builds trees sequentially, where each new tree attempts to correct the errors made by the ensemble of previously constructed trees. The model is trained by minimizing a regularized objective function that combines a differentiable loss function (e.g., softmax for multiclass classification) with a penalty for model complexity.

The predicted class $\hat{y}$ is obtained by summing the contributions of all trees in the ensemble:

$$\hat{y} = \sum_{m=1}^{M} f_m(x), \quad f_m \in \mathcal{F}$$

where each f_m is a decision tree from the space of regression trees $\mathcal{F}$.

In this project, XGBoost was applied to predict educational attainment levels across four categories: No Education, Primary, Secondary, and Higher Education. The model was trained on a feature-rich dataset that included household-level socio-economic indicators and geospatial variables such as distance to the nearest school, healthcare facility, and town (Albelbisi

& Yusop, 2021).

XGBoost was particularly effective for this task due to its ability to handle heterogeneous feature types, manage missing data internally, and resist overfitting through built-in regularization (Brownlee, 2019). During training, stratified 10-fold cross-validation was used alongside SMOTE-based oversampling to mitigate the effects of class imbalance and ensure balanced learning across all education categories (Z. Zhou et al., 2020).

Hyperparameter tuning was performed to optimize tree depth, learning rate, and the number of boosting rounds (Zhang & Zhang, 2021). The model achieved high predictive performance and demonstrated robustness across multiple evaluation metrics, including accuracy, F1-score, and AUC-ROC.

Feature importance analysis revealed that XGBoost prioritized variables such as wealth index, household head age, and access to basic services as strong predictors of educational attainment. These insights not only contributed to model explainability but also supported policy-level interpretations related to barriers in educational access.

3.5 Machine Learning Model Evaluation and Optimization

To evaluate the effectiveness of the machine learning models in predicting educational attainment, a range of evaluation metrics were employed. These metrics—AUC-ROC, precision, recall, F1-score, and accuracy—were selected for their relevance to multi-class classification and their ability to offer a comprehensive performance overview. Each metric highlights a different aspect of model behavior, particularly in the presence of class imbalance, which was addressed during data preparation (see Section 3.3).

3.5.1 Accuracy

Accuracy represents the proportion of correctly classified instances out of the total number of predictions (Hossin & Sulaiman, 2019). It is often used as a basic performance metric

and is particularly effective when the classes are well-balanced. Although class imbalance existed in the original dataset, resampling techniques such as SMOTE were applied during training to reduce this skew (Fernández et al., 2019), allowing accuracy to be interpreted more reliably.

The confusion matrix for each model was used to compute accuracy, defined by the formula:

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

where:

a. TP = True Positives

b. TN = True Negatives

c. FP = False Positives

d. FN = False Negatives

A comparison of model accuracies is presented in Chapter 6, Section 6.4.1.

3.5.2 Area Under the Curve (AUC-ROC

AUC-ROC was used to evaluate the discriminatory power of the models across all classification thresholds (Bradley, 2019). It quantifies the likelihood that a randomly selected sample from a higher education class will be ranked above a randomly selected sample from a lower class (Hand, 2021). While AUC-ROC is often used in binary classification, a one-vs-rest strategy was applied to adapt it to this multi-class problem (Ferrari et al., 2020).

Higher AUC-ROC values indicate stronger classification performance. Details of model AUC-ROC results are provided in Chapter 6, Section 6.4.2.

3.5.3 Recall

Recall, also known as sensitivity or the true positive rate, measures how well a model identifies all relevant instances of a class (Goutte & Gaussier, 2019). For this study, it represents the proportion of individuals correctly predicted to belong to a specific education level among all individuals who actually fall within that category. It is particularly useful when failing to identify underrepresented education levels (false negatives) is costly (Mujtaba et al., 2021).

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

This metric is explored further in Chapter 6, Section 6.4.3.

3.5.4 Precision

Precision measures the accuracy of positive predictions (Saito & Rehmsmeier, 2019). It assesses the proportion of correct predictions among all instances that the model labeled as belonging to a given education level. High precision is important in situations where false positives may lead to misinformed interventions (Luque et al., 2019).

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

Model precision scores are discussed in Chapter 6, Section 6.4.4.

3.5.5 F1-Score

The F1-score is the harmonic mean of precision and recall, providing a single metric that balances the trade-off between them (Opitz & Maclin, 2019). It is especially informative when dealing with class imbalance, as it reflects both the cost of missed detections and incorrect assignments (Tharwat, 2020).

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

This metric was essential in determining how well the models handled both over- and underrepresented education classes. F1-score results are analyzed in Chapter 6, Section 6.4.5.

3.5.6 Summary

Together, these evaluation metrics provided a multidimensional view of model performance.AUC-ROC offered a general perspective on class discrimination, while recall and precision illuminated the models' tendencies toward false negatives and false positives, respectively. The F1-score helped assess overall balance, and accuracy gave a high-level view of correctness. This combination of metrics ensured a thorough and reliable evaluation framework for selecting the most effective model.

3.5.7 Model Optimization

Model optimization involved tuning hyperparameters to enhance predictive performance and reduce overfitting. Techniques such as grid search and randomized search were applied within cross-validation loops to identify optimal values for parameters like learning rate, tree depth, number of estimators, and regularization strength. Each model was fine-tuned independently, ensuring it achieved the best balance between complexity and generalization.

3.6 Deployment

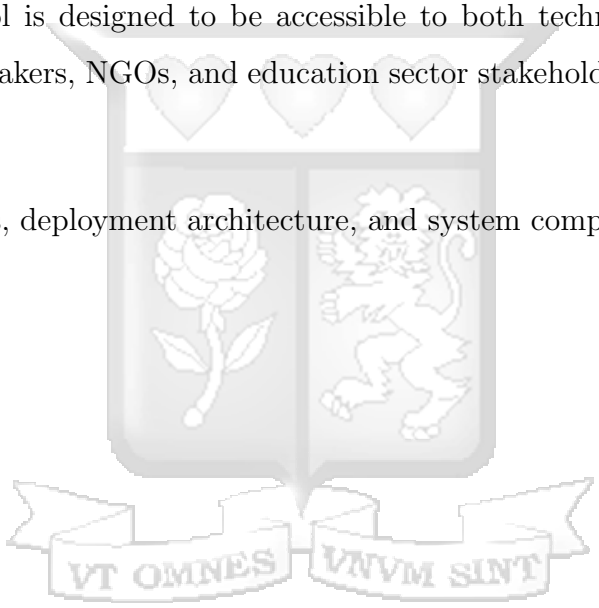
Following model evaluation, the XGBoost classifier was selected for deployment due to its consistent performance across evaluation metrics, including accuracy, F1-score, and AUC-ROC, as detailed in Chapter 6, Section 6.4.5. The objective of this phase was to transform the model from an offline analytical tool into a user-facing solution that supports interactive

exploration and interpretation of educational outcomes.

The deployed application enables users to input household-level information—such as wealth score, internet access, distance to school, and water source—and receive a real-time prediction of the most likely education level for individuals in similar circumstances. This prediction is supplemented with interpretable explanations using SHAP values, allowing users to understand how each input feature contributed to the model's output.

Built using the Streamlit framework, the platform provides a clean user interface with dedicated pages for exploratory data analysis, education level prediction, and SHAP-based model interpretation. The tool is designed to be accessible to both technical and non-technical users, including policymakers, NGOs, and education sector stakeholders seeking data-driven insights.

Further technical details, deployment architecture, and system components are described in Chapter 5.



Chapter 4: System Design and Architecture

4.1 System Modeling

To structure and visualize the system's functionality, Unified Modeling Language (UML) was employed. UML provides a widely accepted modeling language in system development, allowing clear communication among developers, designers, and stakeholders throughout the design phase.

In this study, a Use Case Diagram was used to depict how different user roles interact with the Education Attainment Level Predictor. This behavioral diagram offered a high-level representation of the platform's core functions—such as data input, prediction generation, and result interpretation—and the actors involved, including data analysts, policy decision-makers, and general users.

By mapping these interactions, the use case diagram ensured a shared understanding of user expectations and helped guide interface and backend development decisions.

Figure 4.1 presents the system's use case diagram, highlighting key components and their relationships.

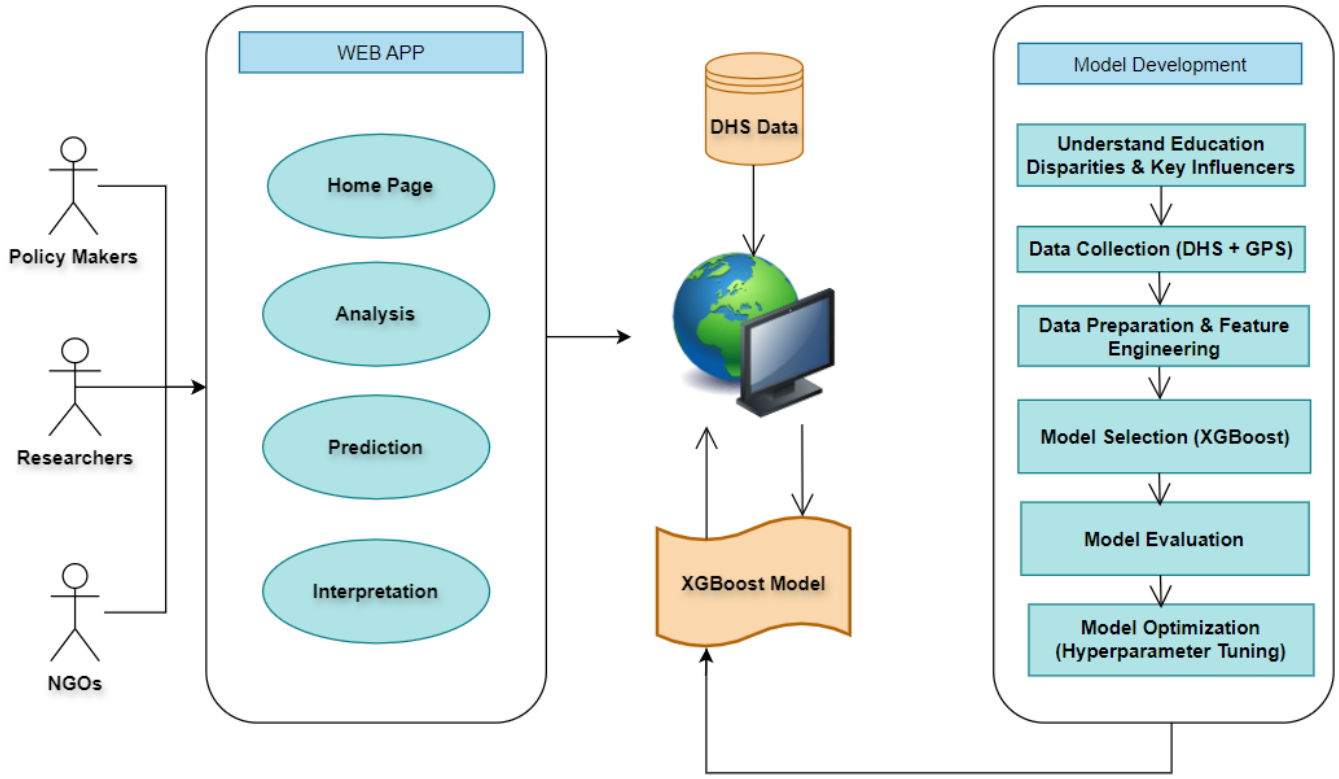


Figure 4.1: UML Use Case Diagram for the Education Attainment Prediction System

4.2 System Components

The system developed in this study consisted of three main components: a structured dataset derived from DHS files, a trained XGBoost model, and a user-friendly web-based application. Together, these components enabled an integrated workflow—from data exploration and prediction to interpretation—designed to support evidence-based decision-making in the education sector.

Data Layer: The raw data was sourced from the Kenya Demographic and Health Survey (DHS) and accompanying GPS dataset. After preprocessing and feature engineering,

the data was organized into two main structured tables—Household and Cluster—linked by `Cluster_ID`. Although no formal database system was used, the data was modeled as relational tables to streamline merging, querying, and analysis.

Model Layer: A ML model was trained using the structured data to predict educational attainment levels. XGBoost was selected for its high performance, interpretability, and ability to handle missing values and class imbalance. The final model was serialized and integrated into the app using Python.

Web Interface: Built with Streamlit, the front-end interface allowed users to interact with the model. Users could input household features, run predictions, and view interpretable SHAP-based visual explanations. The interface also included pages for exploratory data analysis and model insights, making it accessible to both technical and non-technical users.

These components worked cohesively to offer a complete and interactive platform for understanding and predicting educational outcomes based on socio-economic and geographic data.

4.2.1 Data Structure and Modeling

Although the data used in this study was not retrieved from a traditional relational database, it was organized and modeled in a structured format following the download from the Kenya Demographic and Health Survey (DHS) repository. The original datasets were obtained as Comma-Separated Values (CSV) files, including household-level survey data and geospatial cluster information with GPS coordinates.

To facilitate analysis and prediction, the data was restructured into two core entities: **Household** and **Cluster**. The Household table captured socio-economic and demographic variables for each respondent, including features such as household size, education level, internet access, and wealth index. The Cluster table contained location-specific attributes, including GPS coordinates and computed distances to the nearest school, healthcare facility, and town center.

The two tables were linked using the `Cluster_ID` variable, which served as a common key across both datasets. This relationship enabled spatial attributes to be integrated into household-level records, supporting feature engineering and geospatial analysis.

Figure 4.2 presents the Entity Relationship Diagram (ERD), which outlines the structure and linkage of the two main tables used during model training and deployment.

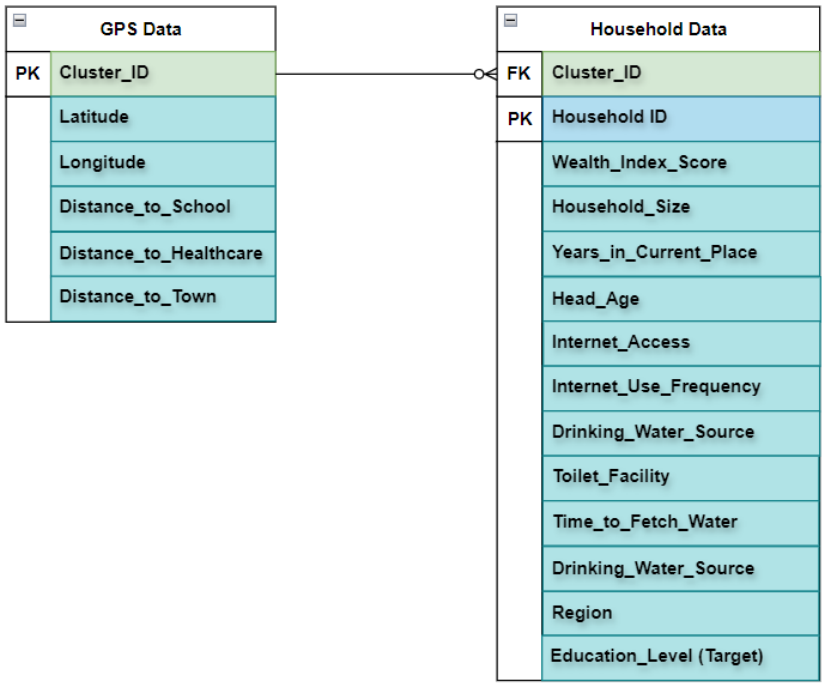


Figure 4.2: Entity Relationship Diagram (ERD) for Household and Cluster Data Structure

Table 4.1 offers a comprehensive breakdown of each data table that was constructed for the prediction system.

Table 4.1: Structured Data Tables for Education Prediction

Table	**Field Name**	**Data Type**	**Description**
GPS Data	Cluster_ID	int (PK)	Unique identifier for geographic cluster
	Latitude	float	Latitude coordinate of the cluster
	Longitude	float	Longitude coordinate of the cluster
	Distance_to_School	float	Distance from the cluster to the nearest school (in km)
	Distance_to_Healthcare	float	Distance to the nearest healthcare facility (in km)
	Distance_to_Town	float	Distance to the nearest town (in km)
Household Data	Household_ID	int (PK)	Unique identifier for household
	Cluster_ID	int (FK)	Foreign key referencing the GPS Data table
	Wealth_Index_Score	float	Household's calculated wealth index
	Household_Size	int	Number of individuals in the household
	Years_in_Current_Place	int	Duration (in years) the household has stayed in current location
	Head_Age	int	Age of the household head
	Internet_Access	string	Indicates whether the household has internet access
	Internet_Use_Frequency	string	Frequency of internet usage in the household
	Drinking_Water_Source	string	Main source of drinking water
	Toilet_Facility	string	Type of toilet facility used
	Time_to_Fetch_Water	int	Time in minutes to fetch water
	Region	string	Administrative region or location of the household
	Education_Level	int (Target)	Categorical target variable representing level of education attained

4.2.2 Web Portal

The web portal developed for this study was structured into four primary sections, each designed to address different aspects of educational data exploration and prediction. These sections included a Home Page that introduces the platform's objectives, an Analysis section for examining socio-economic and geographic disparities, a Prediction interface for generating educational attainment forecasts, and an Interpretation page for visualizing model explanations. To support intuitive navigation and illustrate the layout, both a sitemap and wireframes were created and are presented below.

(i) Sitemap

A sitemap provides a structural overview of the web portal, illustrating the hierarchy and navigation flow between its core sections. It outlines how users move between pages such as data analysis, prediction input, and interpretation, offering a streamlined representation of the user journey. This visual guide enhances usability by ensuring that key components are clearly laid out and logically connected. The sitemap is shown in Figure 4.3.

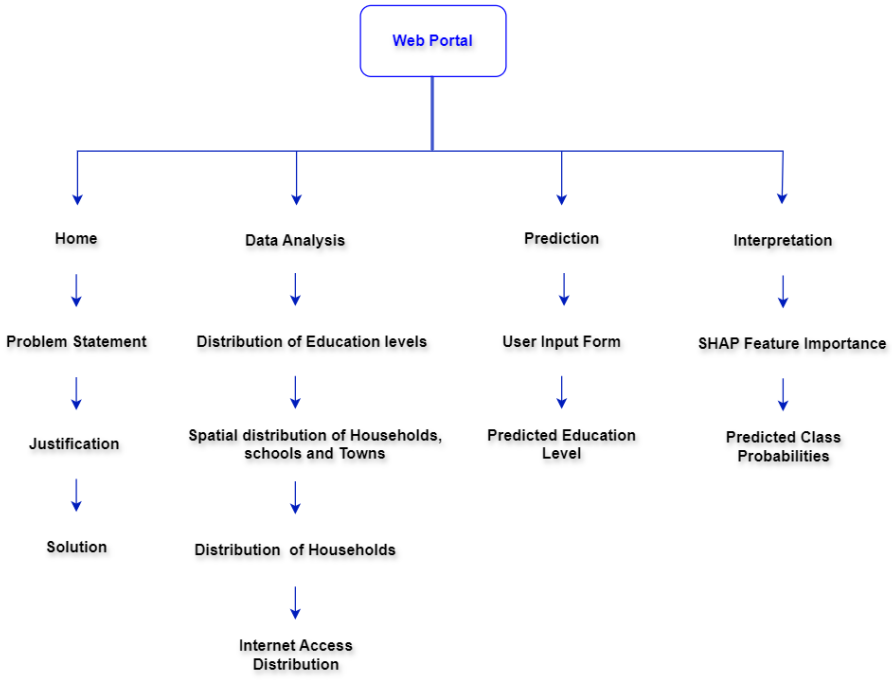


Figure 4.3: Sitemap of the Education Attainment Prediction Web Portal

(ii) Wireframes

Wireframes are schematic illustrations that depict the basic structure of a web application's layout. They represent the arrangement of interface elements—such as menus, input fields, and content blocks—without emphasizing visual design or aesthetics. These blueprints guide the development of intuitive and user-friendly interfaces by allowing early evaluation of functionality and navigation flow. In this study, wireframes were created using Draw.io, an open-source diagramming tool, to visualize the structure of the education prediction portal prior to implementation.

(a) Home Page Wireframe

The home page provides an introductory overview of the education prediction platform, outlining its objectives, functionalities, and potential impact. It also serves as a central navigation point, offering users easy access to other sections of the portal such as data analysis, prediction, and interpretation.

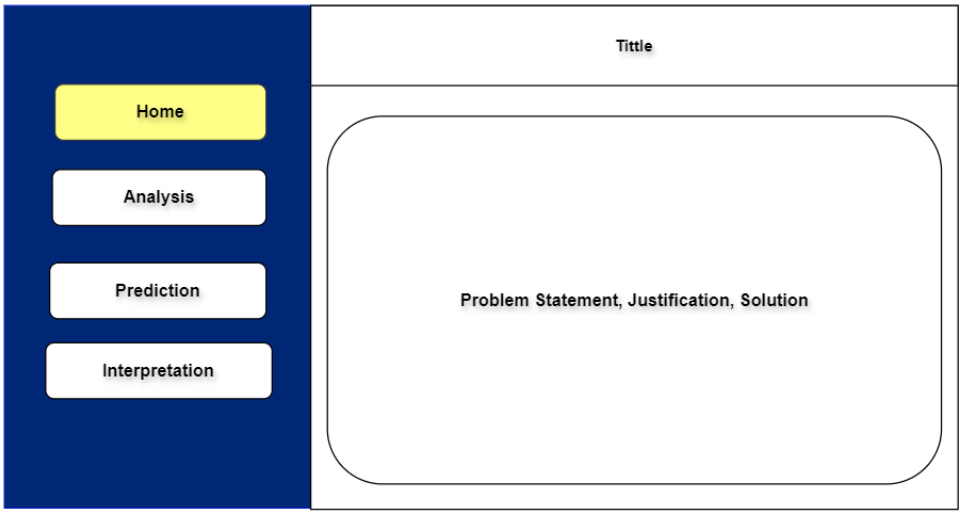


Figure 4.4: Wireframe of the Home Page

(b) Data Analysis Wireframe

The Data Analysis page presents interactive insights derived from the DHS dataset, focusing on patterns in educational attainment across different regions. It includes visualizations such as spatial distribution maps, bar charts, and socio-economic breakdowns, allowing users to explore correlations between household characteristics, geographic location, and education levels.

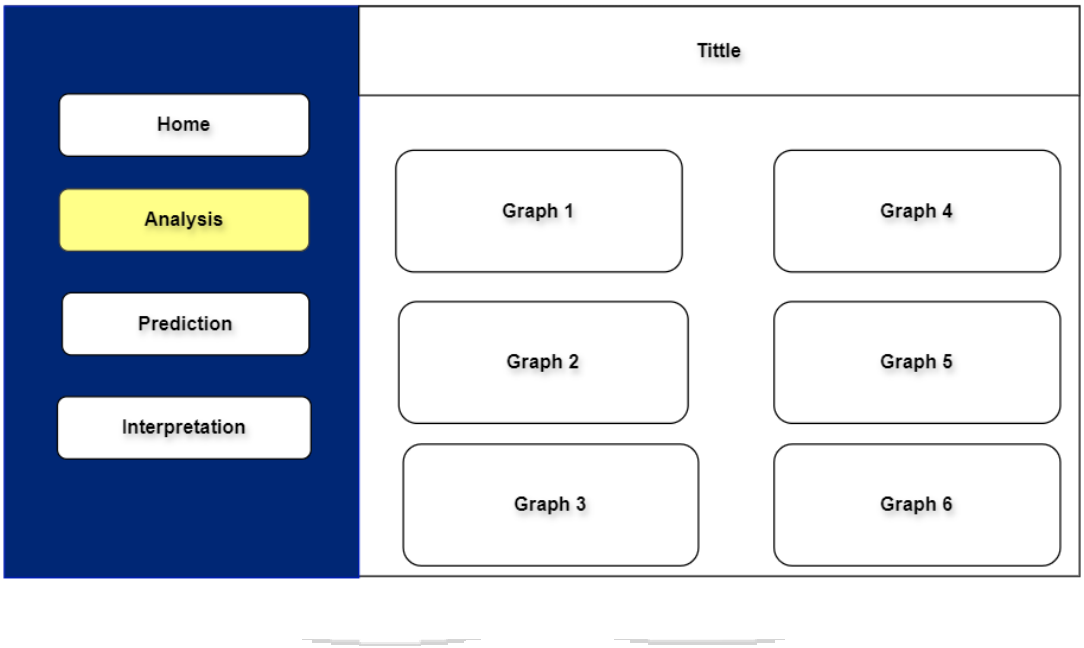


Figure 4.5: Wireframe of the Data Analysis Page

(c) Prediction Page Wireframe

The Prediction page allows users to input household and geographic features—such as wealth index, household size, and distance to services—in order to receive an estimated education level. The interface was designed to be interactive and intuitive, enabling real-time prediction based on the trained XGBoost model.

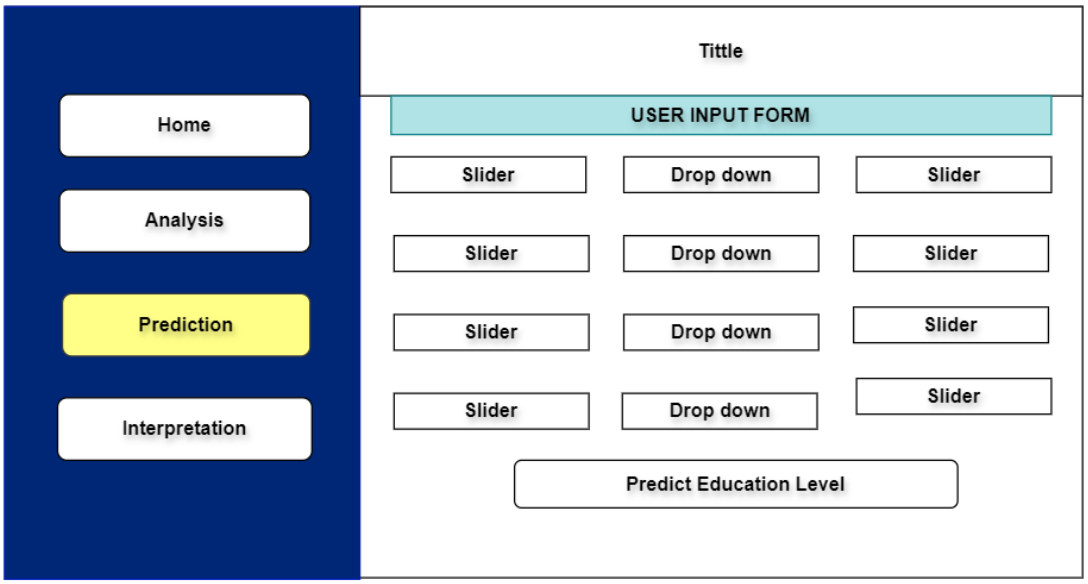


Figure 4.6: Wireframe of the Prediction Page

(d) Interpretation Page Wireframe

The Interpretation page presents SHapley Additive exPlanations (SHAP)-based visual explanations of the model's predictions. It allows users to understand how individual input features—such as wealth index, internet access, or distance to school—contributed to a specific predicted education level. This enhances transparency and trust in the model's decision-making process.

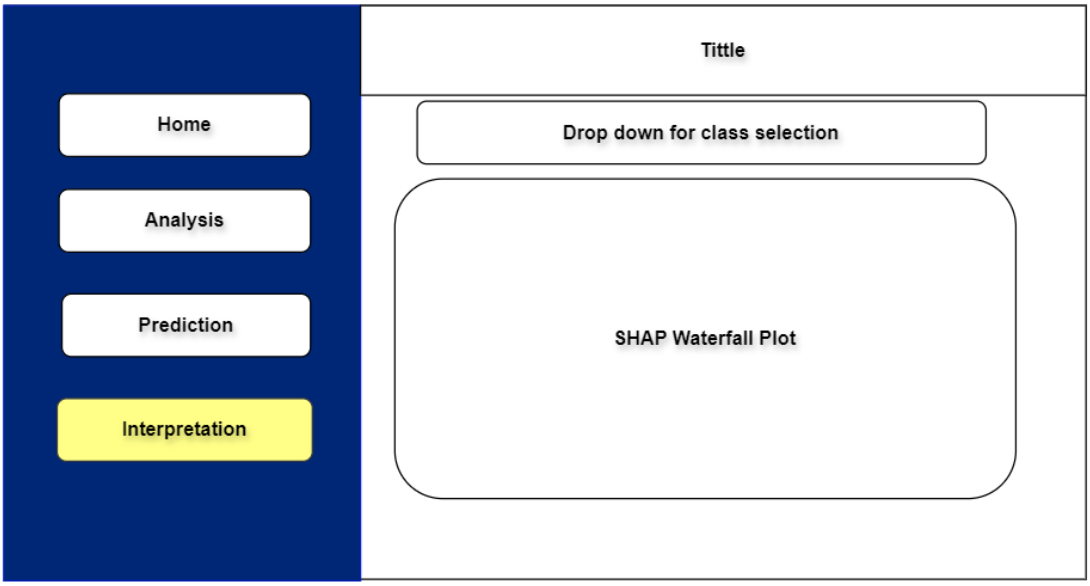


Figure 4.7: Wireframe of the Interpretation Page

Chapter 5: System Implementation and Testing

This chapter presents a detailed account of the technical implementation and development workflow behind the education attainment prediction system. It outlines the design considerations, chosen technologies, and the practical steps followed in building the integrated solution—comprising a machine learning model, backend data infrastructure, and a user-facing web application.

The chapter elaborates on how data was prepared and deployed, how the model was integrated with the web portal, and how each component was engineered to support real-time interaction and interpretability. Emphasis is placed on ensuring usability, scalability, and accessibility for key stakeholders such as policymakers, NGOs, and researchers.

In addition, testing approaches are described to evaluate the system's functionality and assess whether it meets the goals outlined in the problem statement. These evaluations help determine the platform's reliability in supporting data-driven education planning and decision-making in Kenya.

5.1 System Implementation

5.1.1 Database

Although this study did not rely on a traditional relational database for backend storage, the data used in the system was structured and managed using well-defined data models. The Kenya Demographic and Health Survey (DHS) data and the accompanying GPS dataset were initially downloaded in CSV format and preprocessed using Python.

To simulate a relational schema and support structured analysis, the data was modeled into two main entities: Household and Cluster. These were organized as flat files, but the relationship between them mirrored a normalized database structure, with Cluster_ID

serving as the linking key. This organization made it possible to efficiently merge geographic attributes with household-level survey responses.

The data modeling and merging process was handled using Pandas within the Streamlit framework. This enabled seamless ingestion and transformation of the data for use in model training and real-time prediction. Feature selection, transformation, and engineered variables (such as distances to schools and healthcare facilities) were incorporated prior to deployment.

While no Database Management System (DBMS) was deployed in this implementation, the data design adhered to normalization principles, ensuring minimal redundancy, clarity of relationships, and readiness for scaling in future database-backed versions of the platform.

5.1.2 Web Portal

The web portal for this project was implemented using Streamlit, a lightweight and open-source Python framework designed for building data applications with minimal effort. Streamlit was selected for its simplicity, seamless integration with machine learning workflows, and suitability for rapid prototyping and deployment of interactive user interfaces.

The platform runs entirely in Python, allowing the integration of preprocessing logic, model prediction, and interpretability features without requiring separate frontend or backend development. The app structure is organized into multiple pages, each serving a distinct purpose—ranging from exploratory data analysis (EDA) to model-based prediction and result interpretation.

Interactive widgets such as sliders, dropdowns, and input fields allow users to provide household and geographic data, which is then passed to the trained model for real-time prediction. The predicted education level is displayed along with SHAP-based visual explanations that highlight how each feature influenced the outcome.

The portal is designed to be intuitive and responsive, enabling accessibility across devices without requiring advanced technical knowledge. Its deployment as a web-based interface

makes it a practical tool for policymakers, researchers, and organizations interested in identifying and addressing educational disparities in Kenya.

(a) Homepage

The homepage, as shown in Figure 5.1, was designed to offer users a concise introduction to the platform. It summarizes the problem of educational disparities in Kenya and highlights how the tool leverages machine learning to support data-driven policy decisions. Clear navigational links are provided to guide users to other core sections of the application, including data analysis, prediction, and model interpretation. The homepage also includes acknowledgments to supporting initiatives and resources used in the development of the platform.

Figure 5.1: Homepage of the Education Prediction Web Application

(b) Data Analysis Page

The Data Analysis section provided visual insights into the spatial and socio-economic patterns present in the dataset. Through interactive charts and maps, users could explore the distribution of households, schools, and healthcare facilities, as well as understand how variables such as wealth and internet access relate to educational outcomes. Figure 5.2 illustrates a snapshot of the interface.

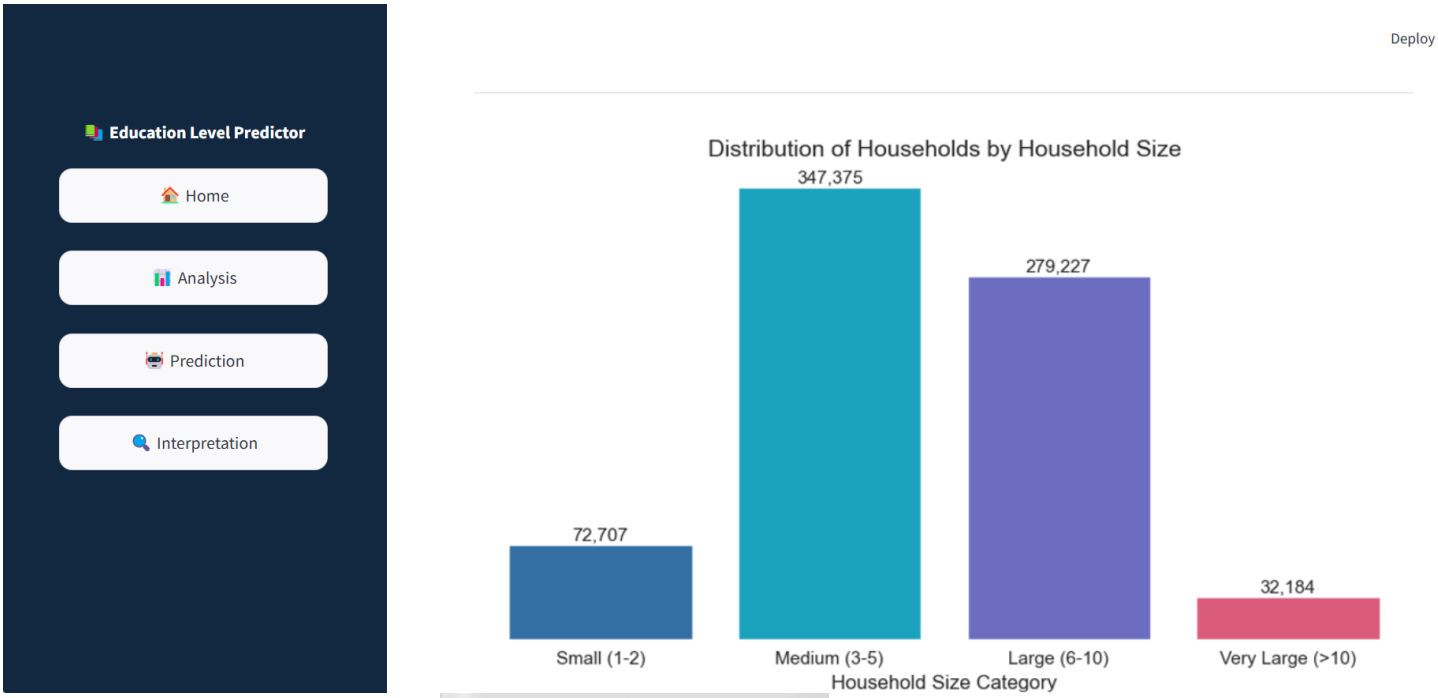


Figure 5.2: Data Analysis Page Showing Insights from Exploratory Data Analysis

(c) Prediction Page

The Prediction section allowed users to input household-specific information such as wealth index, internet access, water source, and distance to services. Based on these inputs, the trained XGBoost model generated real-time predictions for the likely level of educational attainment. This interactive feature enabled stakeholders to simulate different household conditions and assess their impact on education outcomes. Figure 5.3 shows the layout of this functionality.

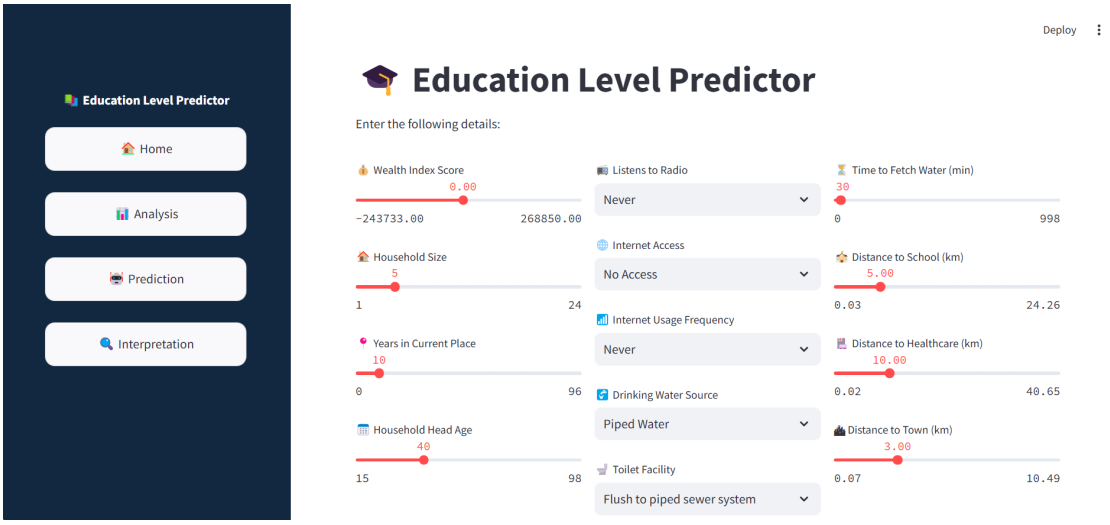


Figure 5.3: Prediction Page for Real-Time Education Level Forecasting

(d) Interpretation Page

The Interpretation section offered model explainability through SHAP (SHapley Additive exPlanations) visualizations. These plots illustrated how each input feature influenced the model's prediction for a specific household. By breaking down the contribution of each variable, users could gain a deeper understanding of the decision process behind the predicted education level. As shown in Figure 5.4, the SHAP waterfall plot highlighted both positive and negative influences on the prediction, enhancing the transparency of the machine learning model.

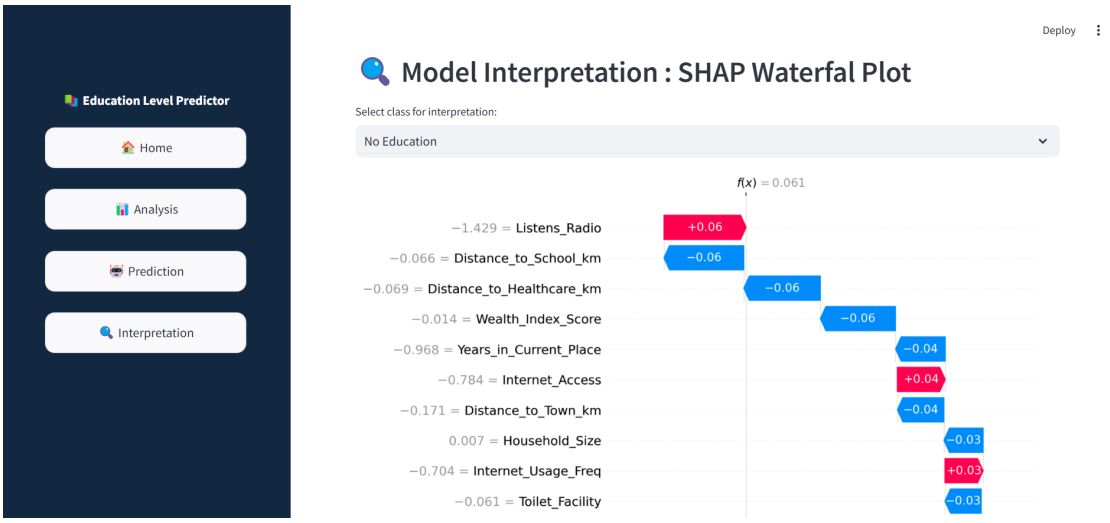


Figure 5.4: Interpretation Page Displaying SHAP-Based Explanations

5.2 Testing

Testing was conducted to ensure the Education Attainment Prediction System met its functional, usability, and performance requirements. This phase verified that the platform operated as intended across all components—data input, prediction, and interpretation—while also ensuring compatibility across devices and browsers. Security and validation checks were carried out to confirm safe data handling and consistent model accuracy. The following subsections provide a breakdown of the key testing categories implemented in the study.

5.2.1 Functionality Testing

Functionality testing was conducted to ensure that all core components of the education prediction system performed as expected. This included verifying the prediction pipeline—starting from user input through to classification of education level—and checking the interpretability module powered by SHAP. Each page of the web portal (Home, Analysis, Prediction, and Interpretation) was tested to confirm that interactive elements responded correctly and rendered appropriate outputs. This process ensured that the system reliably delivered pre-

dictions and visual explanations based on user-supplied household and geographic data.

5.2.2 Usability Testing

Usability testing was conducted to evaluate how easily users could interact with the education prediction web portal and navigate through its key functionalities. The testing involved simulating end-user scenarios where individuals explored the Home, Analysis, Prediction, and Interpretation sections of the platform. Observations focused on the clarity of labels, ease of navigation, and responsiveness of interface elements.

Participants were asked to input household data, interpret predictions, and switch between pages without prior instructions. Their experiences helped identify areas where the interface could be made more intuitive or streamlined. Special attention was given to the prediction input form—ensuring that sliders and dropdowns were easy to use and clearly labeled.

Overall, the feedback collected during usability testing contributed to minor UI adjustments, resulting in a more user-friendly design that accommodates both technical and non-technical users.

5.2.3 Compatibility Testing

Compatibility testing was conducted to verify that the education prediction web portal functioned consistently across various platforms and devices. The system was accessed using multiple web browsers, including Google Chrome, Mozilla Firefox, Microsoft Edge, and Safari, to test cross-browser support. Each section of the portal—from homepage to prediction and interpretation—was examined for responsiveness, rendering accuracy, and feature availability.

Additionally, mobile responsiveness testing was performed to assess how well the platform adapted to different screen sizes and resolutions. The portal was accessed using smartphones and tablets, and users were asked to complete key tasks such as submitting household data

and viewing prediction results. This helped ensure that layout elements, input forms, and visualizations displayed properly on smaller screens. Any minor inconsistencies observed were documented and resolved to improve accessibility and ensure a smooth user experience across devices.

5.2.4 Security Testing

Security testing was conducted to evaluate how well the system safeguarded user inputs and internal data processing. Although the system did not involve the collection of sensitive personal data, basic security protocols were implemented to prevent misuse. Input validation checks were tested to ensure that user-provided values did not lead to crashes or unintended behavior. The integrity of the deployed model and interface was monitored to prevent unauthorized modifications. Additionally, the deployment environment was reviewed to ensure safe handling of form data and to mitigate common security risks such as code injection or URL tampering.

5.2.5 Validation Testing

Validation testing was carried out to determine whether the system met the objectives defined in the problem statement. This included assessing the accuracy and consistency of the education level predictions under varying household and geographic inputs. Several test cases were developed to reflect real-world conditions, and the model's outputs were reviewed for plausibility and alignment with expected outcomes. Furthermore, feedback was gathered from users with experience in education policy and data analytics to validate the system's usefulness in supporting data-driven decisions. Their insights were instrumental in confirming the platform's effectiveness and relevance in addressing educational disparities.

Chapter 6: Discussion of Results

This section discusses the key outcomes of the education prediction system, drawing from the results obtained during data exploration, model evaluation, and deployment. It highlights the main patterns observed, the performance of the machine learning models, and the contribution of different input features to the prediction task. These results are interpreted in relation to the study's objectives and their potential relevance to educational policy and planning.

6.1 Data Understanding

6.1.1 Education Distribution Across Counties

To understand regional disparities in education attainment, counties were grouped and visualized based on household education levels. Group 1 counties, predominantly from arid and semi-arid regions, showed a higher proportion of households with no formal education, such as Mandera (72.4%) and Garissa (65.7%). In contrast, Group 4 counties, mainly from central and highland regions, had significantly lower rates of no education and higher levels of secondary and post-secondary attainment. For instance, Nyamira and Kiambu showed over 85% of households having at least primary education. This contrast underscores the strong geographic influence on education outcomes in Kenya.

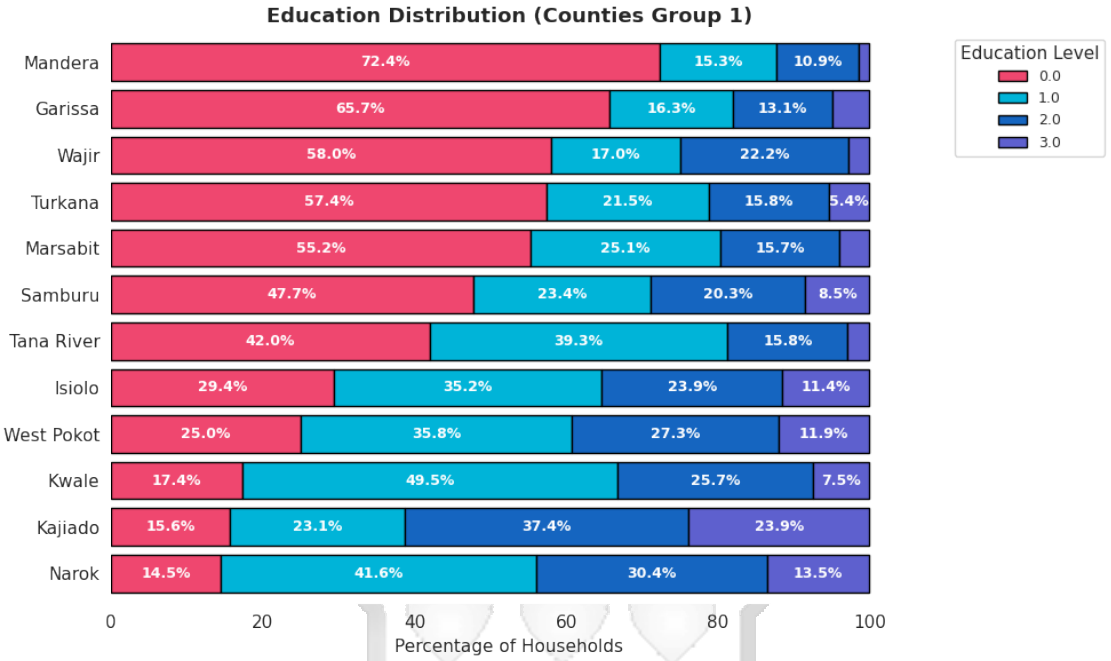


Figure 6.1: Counties in the northern and arid regions show the highest concentration of households with no formal education.

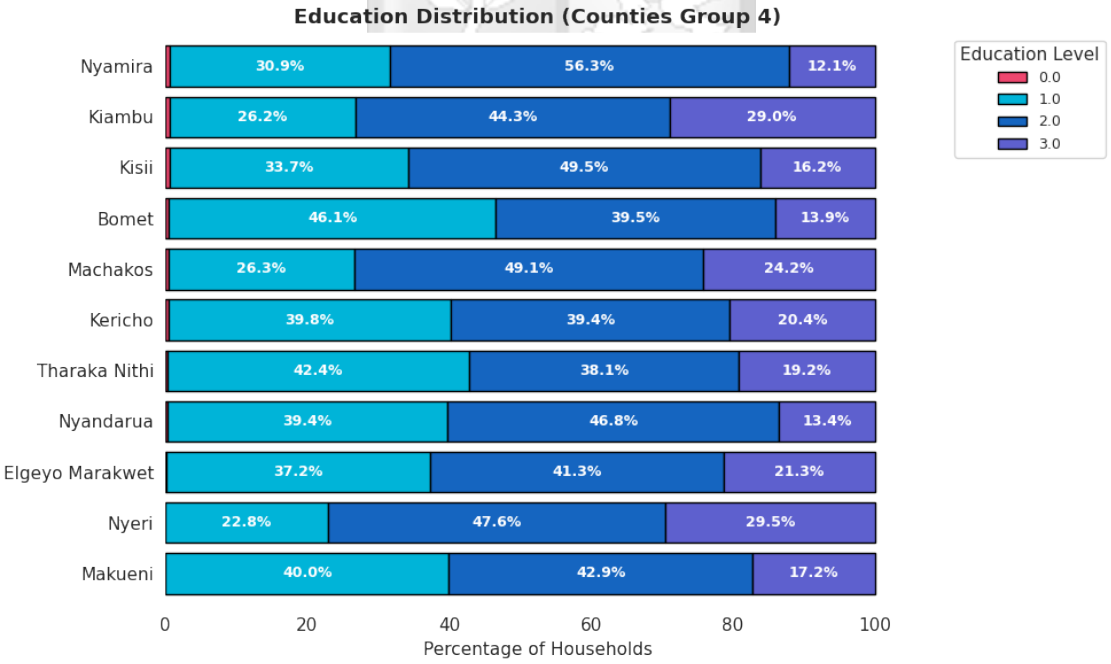


Figure 6.2: Counties in the central and highland regions exhibit higher proportions of households attaining secondary and post-secondary education.

6.1.2 Education Levels by Wealth Category

Wealth-related disparities in educational attainment are evident in the chart. Households in the *low wealth category* show the highest percentage at lower education levels, particularly Level 1. Conversely, households in the *high wealth category* dominate the higher levels of education, with Level 3 showing a sharp increase. This pattern underscores the strong influence of economic status on access to and progression through formal education. It highlights the need for policies that bridge the education gap for financially disadvantaged populations.

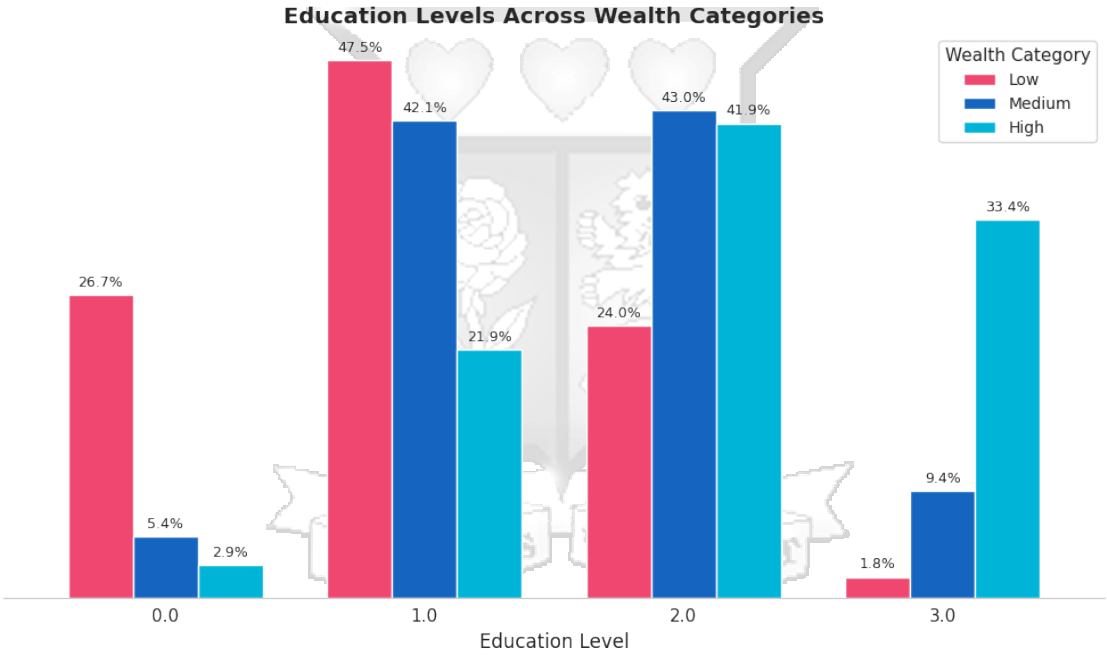


Figure 6.3: Educational attainment by household wealth category, illustrating disparities in access to higher education levels among low-income versus high-income households.

6.1.3 Education Levels by Household Size Category

The visualization reveals interesting patterns in educational attainment across different household size categories. Households with a medium size (3–5 members) and large size (6–10 members) tend to have higher representation at the middle education levels (Levels

1 and 2). Very small households (1–2 members) have a relatively lower percentage across all education levels, while very large households (more than 10 members) show a higher concentration at the lower levels of education, particularly Level 0 and Level 1. These trends suggest that extremely large households may face resource constraints that hinder progression through the education system, highlighting the potential impact of family size on educational outcomes.

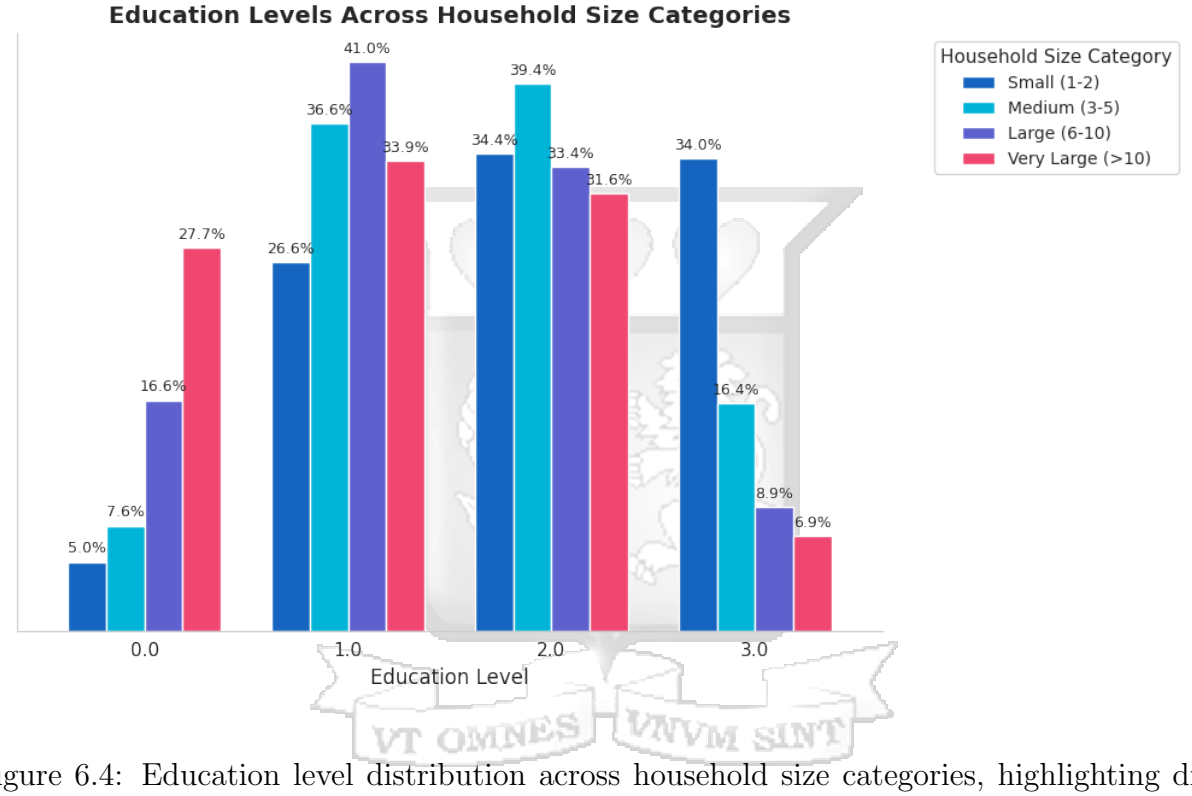


Figure 6.4: Education level distribution across household size categories, highlighting disparities in attainment based on family size.

6.1.4 Internet Access Distribution by Wealth Category

The chart illustrates a clear digital divide based on wealth categories. Households in the low wealth category predominantly lack internet access, with a significantly higher percentage reporting no access compared to medium and high wealth households. In contrast, mobile and fixed internet access increase with wealth, with the high wealth group showing the highest levels of connectivity. This disparity emphasizes the correlation between economic

status and digital inclusion, which has direct implications for access to educational resources and opportunities in the digital age.

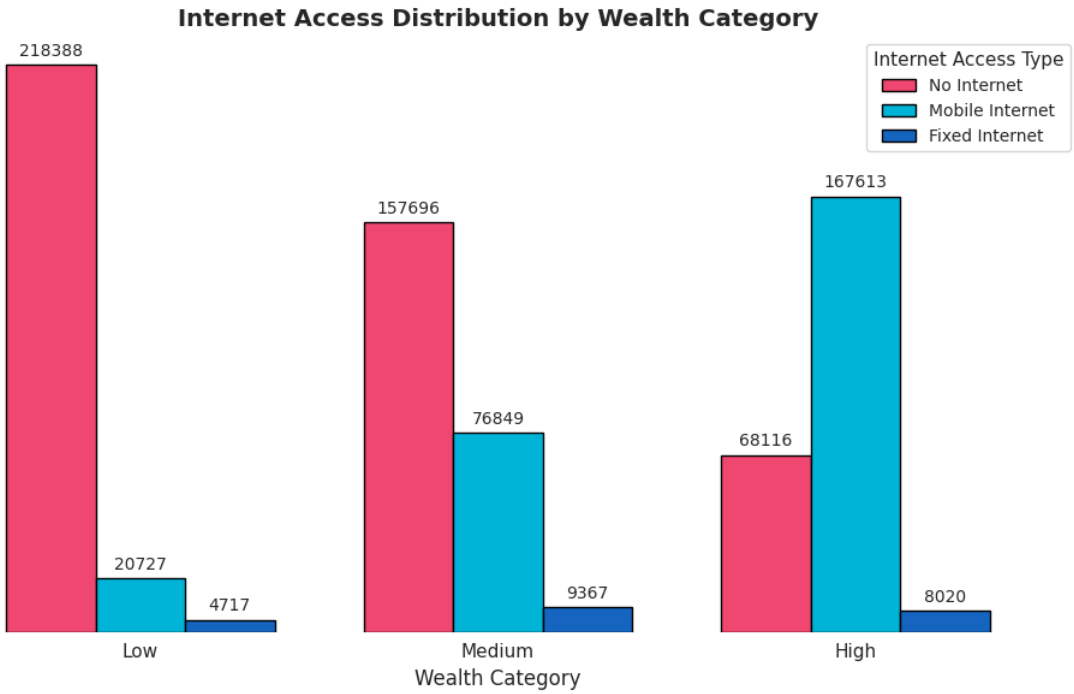


Figure 6.5: Internet access distribution across wealth categories, illustrating the digital divide that exists between low and high-income households.

6.1.5 Distribution of Households by Distance to School

The distribution of household distances to the nearest school highlights a critical barrier to educational access. A significant proportion of households are located more than 2 kilometers from the nearest school, with some residing over 5 kilometers away. This physical separation can contribute to lower school attendance and completion rates, particularly in rural and underserved areas. These findings support the inclusion of distance-related variables in the model, reinforcing the importance of geographic access in understanding educational attainment disparities.

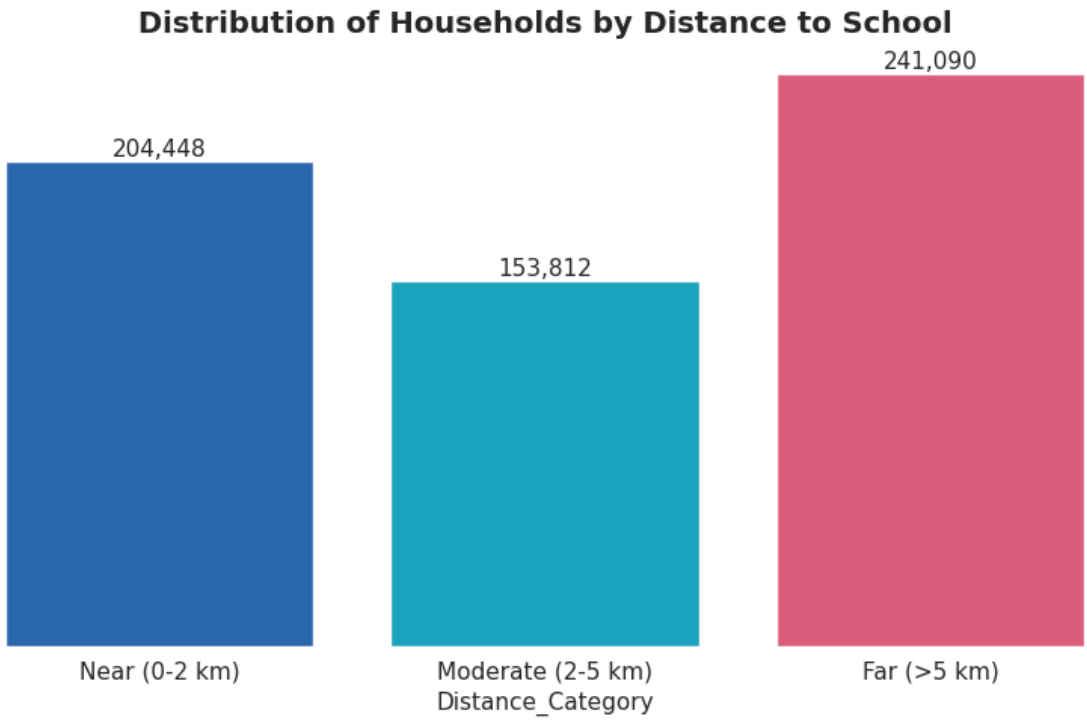


Figure 6.6: Distribution of households by distance to the nearest school. Longer travel distances may hinder school attendance, especially in remote areas.

6.1.6 Spatial Distribution of Households and Key Services

Figure 6.7 shows how households, schools, healthcare facilities, and towns are distributed across Kenya. Services tend to cluster in urban areas, while rural regions remain underserved. This spatial imbalance may limit educational opportunities for remote households. These patterns support the inclusion of both distance-based variables and the `region` variable in the model to account for accessibility and broader geographic disparities in educational outcomes.

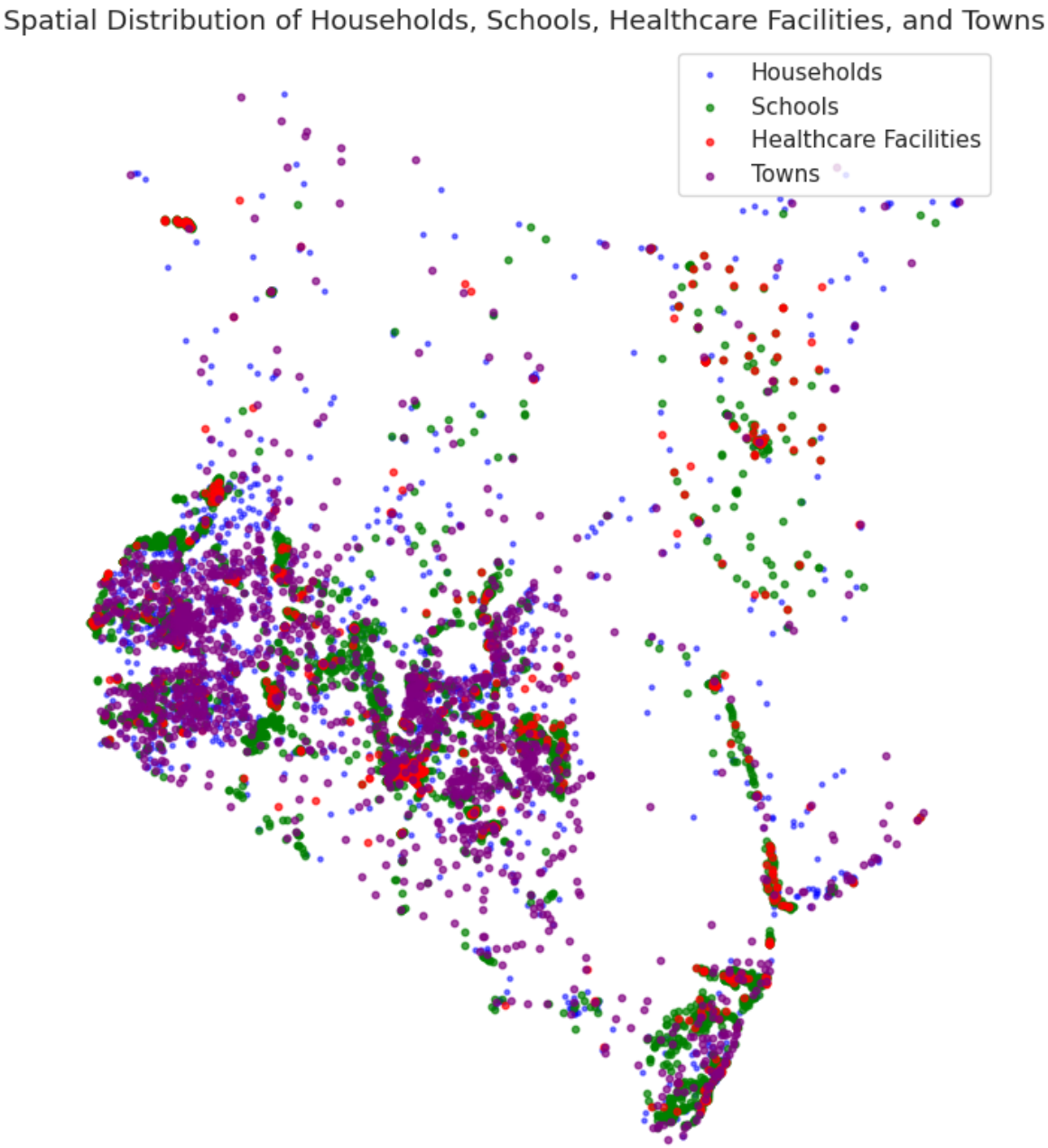


Figure 6.7: Spatial distribution of households and key public services, revealing disparities in access across rural and urban regions.

6.1.7 Distribution of Households by Drinking Water Source

Figure 6.8 displays the distribution of households by their primary drinking water sources. The data reveals that piped water is the most common source, followed by rainwater and public taps. A notable number of households still rely on unimproved or unsafe sources such as unprotected springs, wells, and surface water. These disparities in access to safe water are important socioeconomic indicators, as inadequate water access may correlate with lower school attendance and poorer educational outcomes, particularly in rural settings.

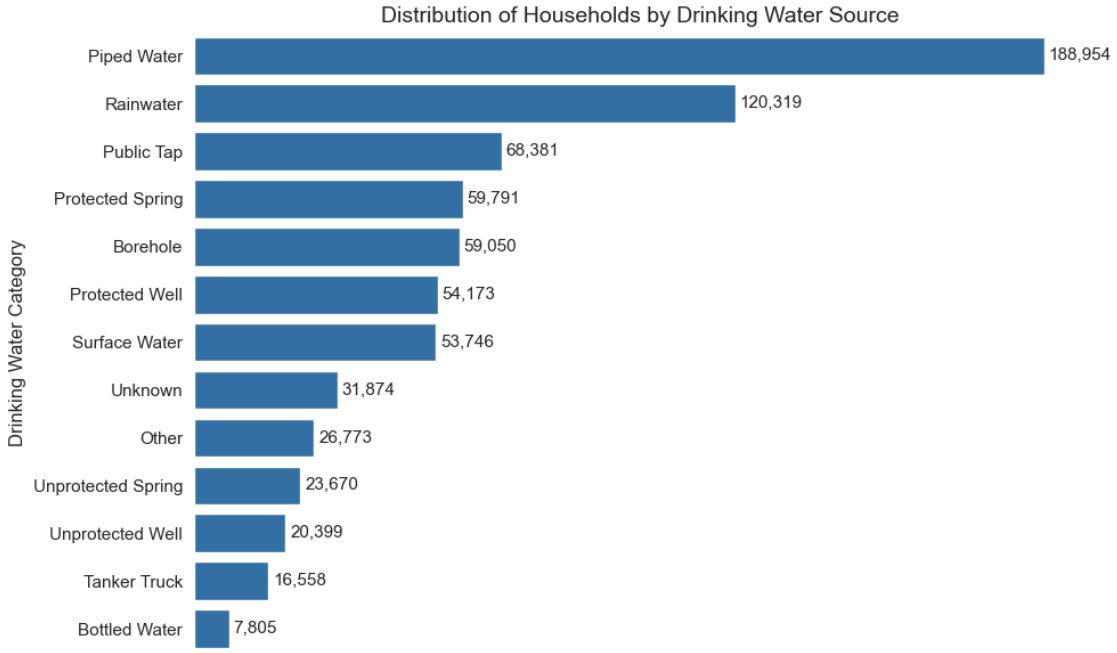


Figure 6.8: Distribution of households by drinking water source

6.1.8 Education Level Distribution by Distance to School

Figure 6.9 illustrates how household proximity to schools correlates with education levels. Households located nearer to schools (0–2 km) show a higher proportion of members attaining post-primary education levels (Levels 2 and 3). In contrast, households situated further away (over 5 km) exhibit a noticeable decline in the proportion of individuals reaching the highest education level. This supports the hypothesis that increased physical distance to educational

facilities may serve as a barrier to continued schooling, particularly in underserved regions.

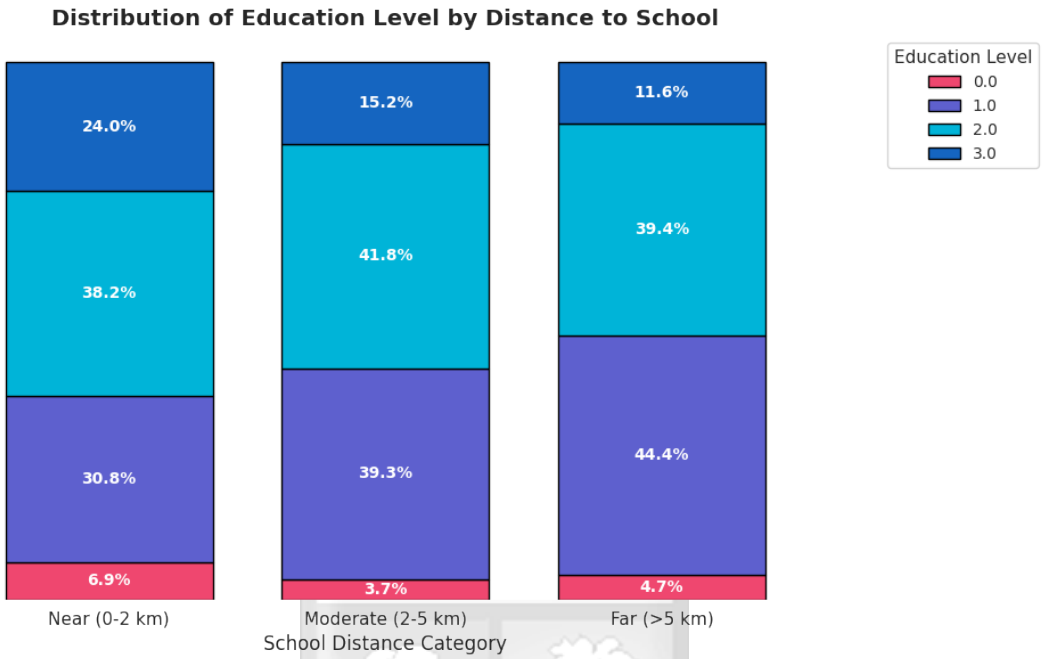


Figure 6.9: Distribution of education level by school distance category

6.2 Data Preparation

6.2.1 Class Imbalance

An analysis of the target variable—education level—revealed a significant class imbalance, with Level 0 (no education) and Level 3 (higher education) being notably underrepresented compared to Levels 1 and 2. This imbalance posed a risk of biased model learning, particularly in predicting less frequent education levels.

To address this, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. As illustrated in Figures 6.10 and 6.11, SMOTE balanced the dataset by synthetically generating instances for the minority classes. This preprocessing step was essential to improve model generalization and ensure fair prediction across all education levels.

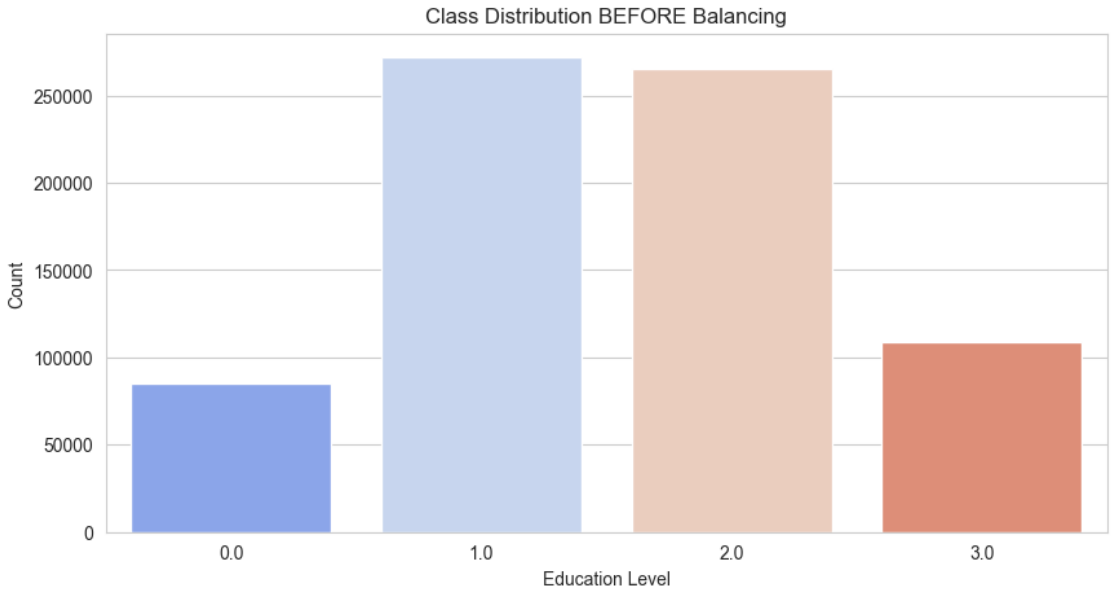


Figure 6.10: Class distribution before applying SMOTE

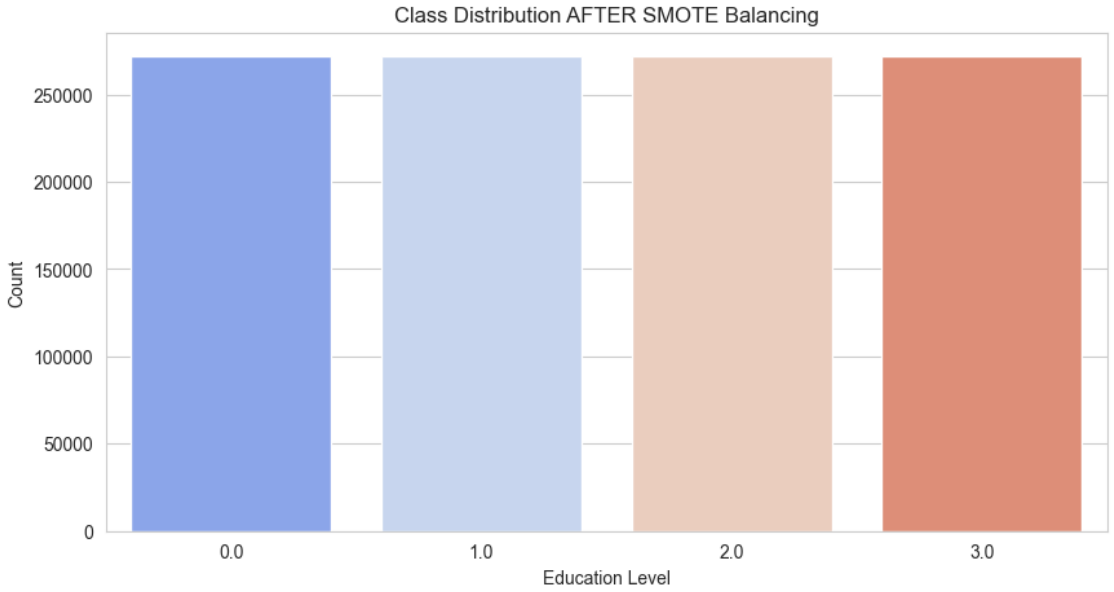


Figure 6.11: Class distribution after applying SMOTE

6.3 Machine Learning Modeling

In this study, the Decision Tree classifier was used as a baseline model due to its simplicity, interpretability, and ease of visualization. Given that the prediction task involved categorical educational attainment levels that do not follow a strict ordinal sequence, ordinal regression was deemed unsuitable. Furthermore, while Support Vector Machines (SVMs) were considered during the literature review, they were not implemented in the experimental phase due to their computational demands and limited model transparency. In contrast, tree-based ensemble methods such as Random Forest and XGBoost offer better performance while also supporting interpretability via feature importance and SHAP values, which are essential for policy-oriented applications.

6.3.1 Decision Tree Classifier

The Decision Tree classifier was implemented as part of the modeling workflow due to its simplicity, interpretability, and ability to handle both numerical and categorical variables. While it achieved a moderate overall accuracy of 72.7%, the model exhibited strong performance in predicting classes 0 and 3, with relatively lower precision and recall for classes 1 and 2. These variations highlight the model's sensitivity to class distribution and its tendency to overfit, especially when applied to imbalanced datasets. Nonetheless, its transparent decision paths make it a useful baseline for comparison with more complex models.

Table 6.1: Classification Report for Decision Tree Model

Class	Precision	Recall	F1-Score	Support
0.0	0.85	0.92	0.89	54371
1.0	0.63	0.64	0.64	54370
2.0	0.62	0.47	0.54	54371
3.0	0.76	0.87	0.81	54370
Accuracy			**0.73**	217482
Macro Avg	0.72	0.73	0.72	217482
Weighted Avg	0.72	0.73	0.72	217482

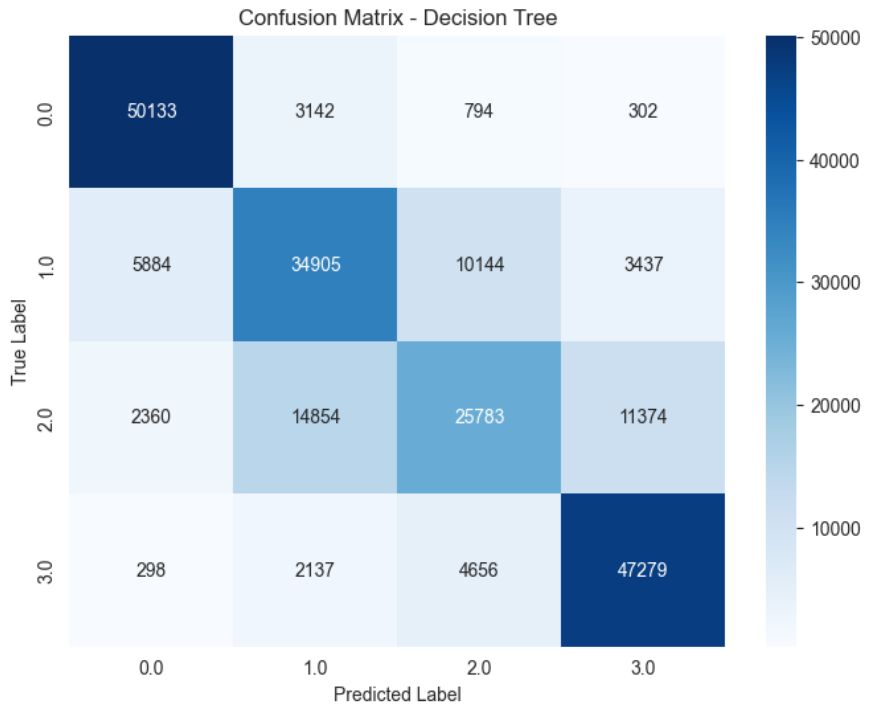


Figure 6.12: Confusion Matrix for Decision Tree Classifier

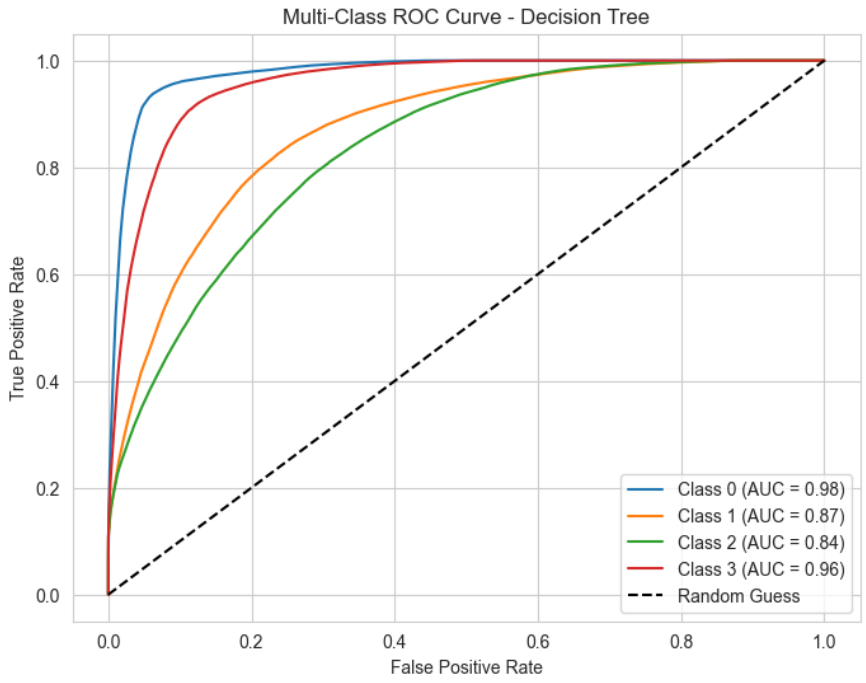


Figure 6.13: AUC-ROC Curve for Decision Tree Classifier

6.3.2 Random Forest Classifier

The Random Forest classifier demonstrated improved performance compared to the Decision Tree, achieving an accuracy of 78.6%. The ensemble nature of the model allowed it to generalize better across classes, with notable improvements in the prediction of class 2 compared to the Decision Tree. It showed high precision and recall for classes 0 and 3, affirming its robustness in distinguishing both low and high education levels. These results suggest that the Random Forest model offers a balanced trade-off between performance and interpretability, making it suitable for complex classification tasks in this study.

Table 6.2: Classification Report for Random Forest Model

Class	Precision	Recall	F1-score	Support
0.0	0.88	0.93	0.90	54371
1.0	0.69	0.75	0.72	54370
2.0	0.79	0.54	0.64	54371
3.0	0.79	0.92	0.85	54370
Accuracy		0.79		217482
Macro Avg	0.79	0.79	0.78	217482
Weighted Avg	0.79	0.79	0.78	217482

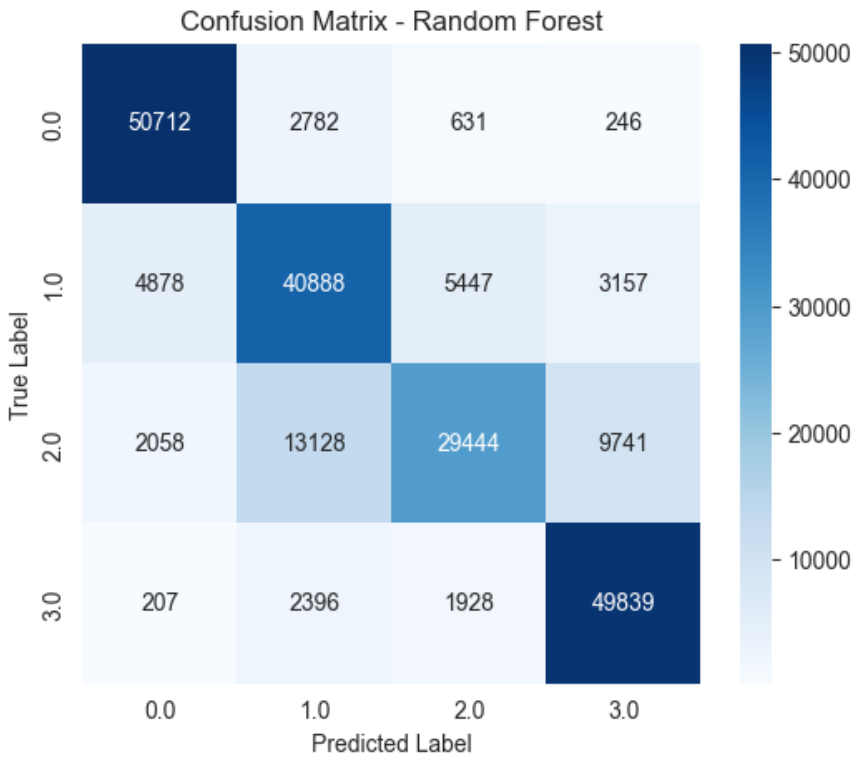


Figure 6.14: Confusion Matrix for Radom Forest Classifier

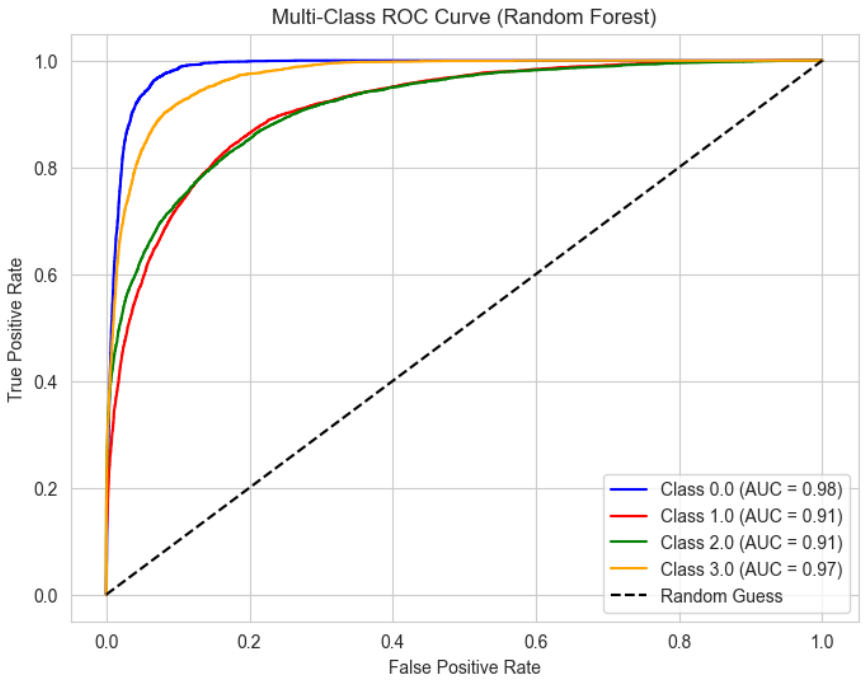


Figure 6.15: AUC-ROC Curve for Random Forest Classifier

6.3.3 XGBoost Classifier

The XGBoost classifier outperformed all other models in this study, achieving the highest accuracy of 83.3%. Its ability to handle complex interactions and minimize overfitting through regularization contributed to superior predictive performance across all education levels. Class 0 and Class 3, in particular, showed outstanding recall and precision scores, indicating that the model excelled at identifying both households with no education and those with higher education. These results highlight the effectiveness of XGBoost as a robust and scalable solution for multi-class classification tasks in education data modeling.

Table 6.3: Classification Report for XGBoost Model

Class	Precision	Recall	F1-score	Support
0.0	0.89	0.98	0.93	54371
1.0	0.78	0.76	0.77	54370
2.0	0.81	0.65	0.72	54371
3.0	0.84	0.94	0.89	54370
Accuracy		0.83		217482
Macro Avg	0.83	0.83	0.83	217482
Weighted Avg	0.83	0.83	0.83	217482

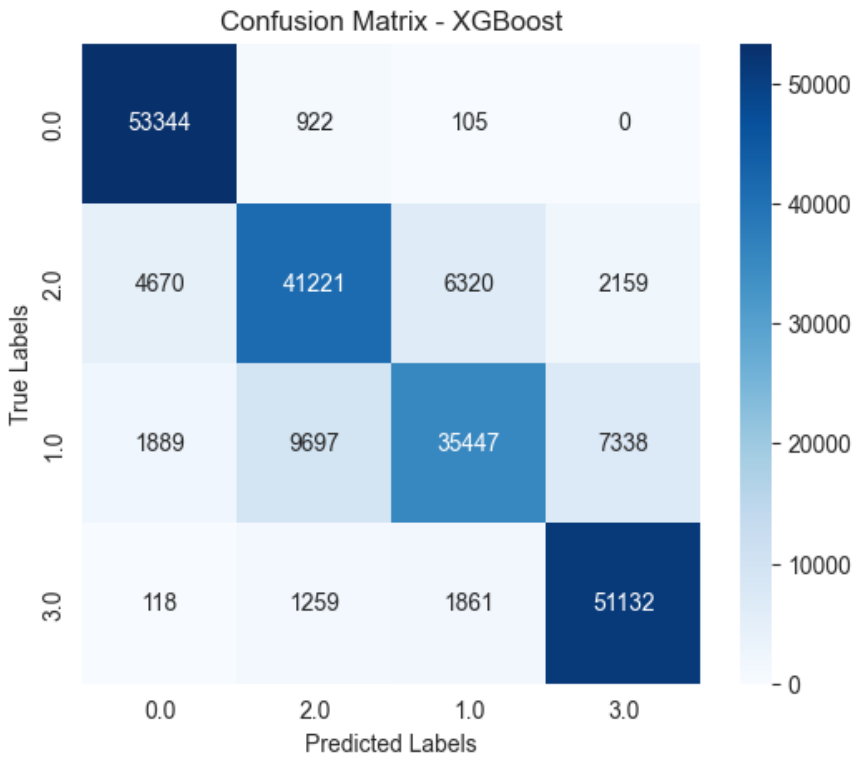


Figure 6.16: Confusion Matrix for XGBoost Classifier

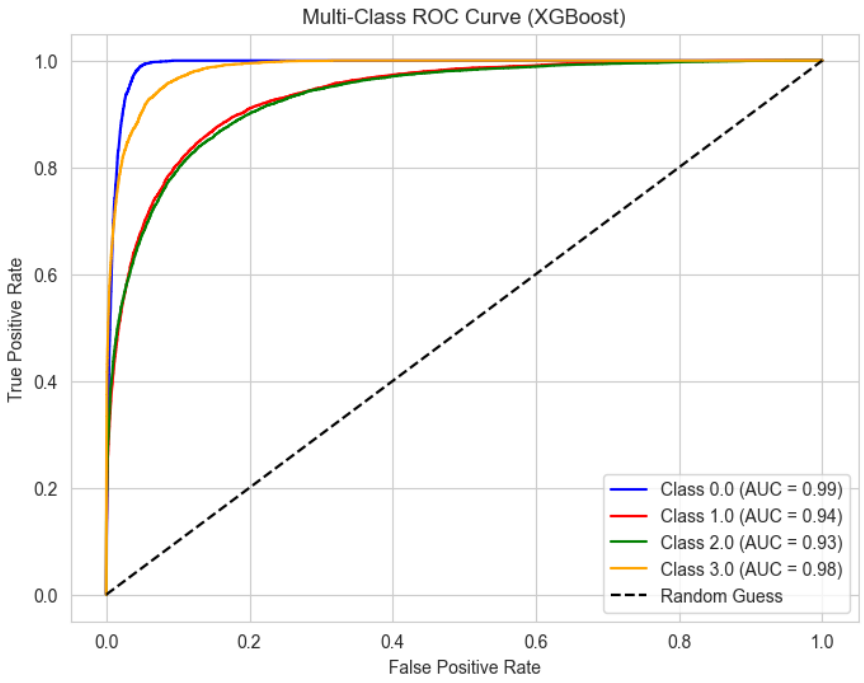


Figure 6.17: AUC-ROC Curve for XGBoost Classifier

6.4 Model Evaluation and Optimization

6.4.1 Accuracy

Accuracy serves as a fundamental metric for evaluating the overall performance of the classification models in predicting household education levels. Among the tested models, XGBoost demonstrated the highest accuracy of 83.3%, highlighting its superior ability to correctly classify instances across all four education categories. Random Forest followed with an accuracy of 78.6%, showing strong capability in handling the multi-class nature of the task. The Decision Tree model yielded an accuracy of 72.7%, suggesting comparatively lower effectiveness in capturing the complex relationships within the dataset.

These accuracy scores reflect each model's ability to learn patterns from socio-economic and spatial features and make reliable predictions. The exceptional performance of XGBoost underscores its strength in managing both linear and non-linear patterns while mitigating

overfitting through regularization and boosting.

Table 6.4: Accuracy Scores of the Developed Models

Model	**Accuracy Score**
Decision Tree	0.73
Random Forest	0.79
XGBoost	0.83

6.4.2 Area Under the Curve (AUC-ROC)

The Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) offers a robust measure of the model's ability to distinguish between the education level classes. AUC values range from 0 to 1, with higher values indicating better performance.

XGBoost recorded the highest AUC scores across all classes, achieving near-perfect separability with Class 0 (AUC = 0.99), Class 1 (AUC = 0.94), Class 2 (AUC = 0.93), and Class 3 (AUC = 0.98). These values underscore the model's superior capability in distinguishing between the education levels.

Random Forest also demonstrated strong discriminatory power, with Class 0 (AUC = 0.98), Class 1 (AUC = 0.91), Class 2 (AUC = 0.91), and Class 3 (AUC = 0.97). The model showed consistent and reliable performance across all target classes.

Decision Tree, while performing slightly lower, still yielded high AUCs: Class 0 (AUC = 0.98), Class 1 (AUC = 0.87), Class 2 (AUC = 0.84), and Class 3 (AUC = 0.96). Although effective, its performance was comparatively less robust in predicting middle classes (1 and 2).

Overall, XGBoost emerged as the most effective model in terms of AUC, followed closely by Random Forest, while Decision Tree showed slightly lower separation capacity—especially for intermediate education levels.

6.4.3 F1-Score

The F1-score provides a balanced measure of model performance by combining precision and recall into a single metric. It is particularly useful for evaluating classification tasks with class imbalance or varying error costs.

Among the evaluated models, **XGBoost** demonstrated the highest average F1-score (0.83), indicating strong overall performance in correctly classifying all four education levels. It maintained consistently high F1-scores across individual classes, particularly for Class 0 (0.93) and Class 3 (0.89), which shows effectiveness in identifying both low and high education outcomes.

Random Forest followed with an average F1-score of 0.78, reflecting solid predictive power. While it performed well for Class 0 (0.90) and Class 3 (0.85), it had a notably lower F1-score for Class 2 (0.64), suggesting challenges in capturing the mid-range education category.

Decision Tree yielded the lowest average F1-score (0.72), mainly due to a low score of 0.54 for Class 2. This suggests the model struggled with misclassifying middle education levels, though it maintained strong performance on Class 0 (0.89) and Class 3 (0.81).

Overall, the F1-score analysis highlights XGBoost as the most balanced model in terms of both precision and recall, especially for critical decision-making classes.

6.4.4 Recall

Recall measures the model's ability to correctly identify all relevant instances of each class. High recall values indicate fewer false negatives, which is crucial for identifying underrepresented or high-impact categories.

XGBoost achieved the highest overall recall (macro avg: 0.83), with outstanding recall for Class 0 (0.98) and Class 3 (0.94). These results imply that XGBoost is especially effective at detecting both ends of the education spectrum.

Random Forest also performed strongly in terms of recall, with an average of 0.79. It maintained excellent recall for Class 0 (0.93) and Class 3 (0.92), though performance dropped for Class 2 (0.54), indicating some difficulty in recalling mid-level education outcomes.

Decision Tree had the lowest average recall (0.73), with a sharp decline in Class 2 (0.47). While still competent in recognizing Class 0 (0.92) and Class 3 (0.87), its limited ability to recall middle-range categories limits its overall reliability.

These findings reinforce XGBoost's superiority in minimizing false negatives, especially for crucial categories, making it highly suitable for policy applications where inclusion of at-risk households is a priority.

6.4.5 Statistical Significance Testing of Model Performance

To evaluate whether the observed performance differences between models were statistically significant, a paired t-test was conducted on the cross-validation accuracies of the XGBoost and Random Forest classifiers across the same data folds. The test yielded a t-statistic of 16.00 and a p-value of 0.000089, indicating that the performance advantage of XGBoost is not due to random chance but reflects a genuine improvement in classification capability. This result statistically reinforces the superiority of XGBoost in this context, complementing its higher accuracy, better handling of feature interactions, and enhanced interpretability through SHAP values. These findings justify the selection of XGBoost as the final model for deployment in the education-level prediction system. Future studies may extend this analysis to include additional metrics and alternative statistical tests, such as McNemar's test or Wilcoxon signed-rank test, to validate consistency across different evaluation criteria.

6.4.6 Model Optimization

Given XGBoost's superior baseline performance across all metrics, further hyperparameter tuning was conducted to enhance its predictive accuracy. Using a grid search approach,

key parameters such as `max_depth`, `learning_rate`, `n_estimators`, and `subsample` were systematically adjusted to identify the optimal configuration.

As illustrated in Figure 6.16, the model's accuracy improved notably after tuning, reaching **86%**—a significant gain from the baseline of 83.3%. This improvement highlights the sensitivity of XGBoost to parameter tuning and its ability to adapt to complex, non-linear relationships within the dataset.

The optimization process confirms that fine-tuning hyperparameters not only boosts accuracy but also enhances the model's generalization capacity, making it more reliable for predicting educational attainment across diverse household profiles.

6.5 Deployment

6.5.1 Feature Importance

Figure 6.18 displays the feature importance plot generated by the XGBoost model. The most influential variables in predicting household education levels include **Wealth Index Score**, **Head Age**, **Region**, and **Distance to School**. These features contribute the most to the model's decision-making, indicating their strong relationship with educational attainment.

Socioeconomic indicators such as household size, toilet facility type, and drinking water source also played notable roles, suggesting that basic living conditions and access to amenities are crucial factors in educational outcomes. Interestingly, digital access variables like internet usage frequency and access type appeared further down the list, yet still contributed to the model's performance.

This analysis provides valuable insight into the underlying drivers of educational attainment, aligning well with the study's objective to identify key socio-spatial inequalities. These insights can guide policy decisions by emphasizing the factors most associated with poor education outcomes.

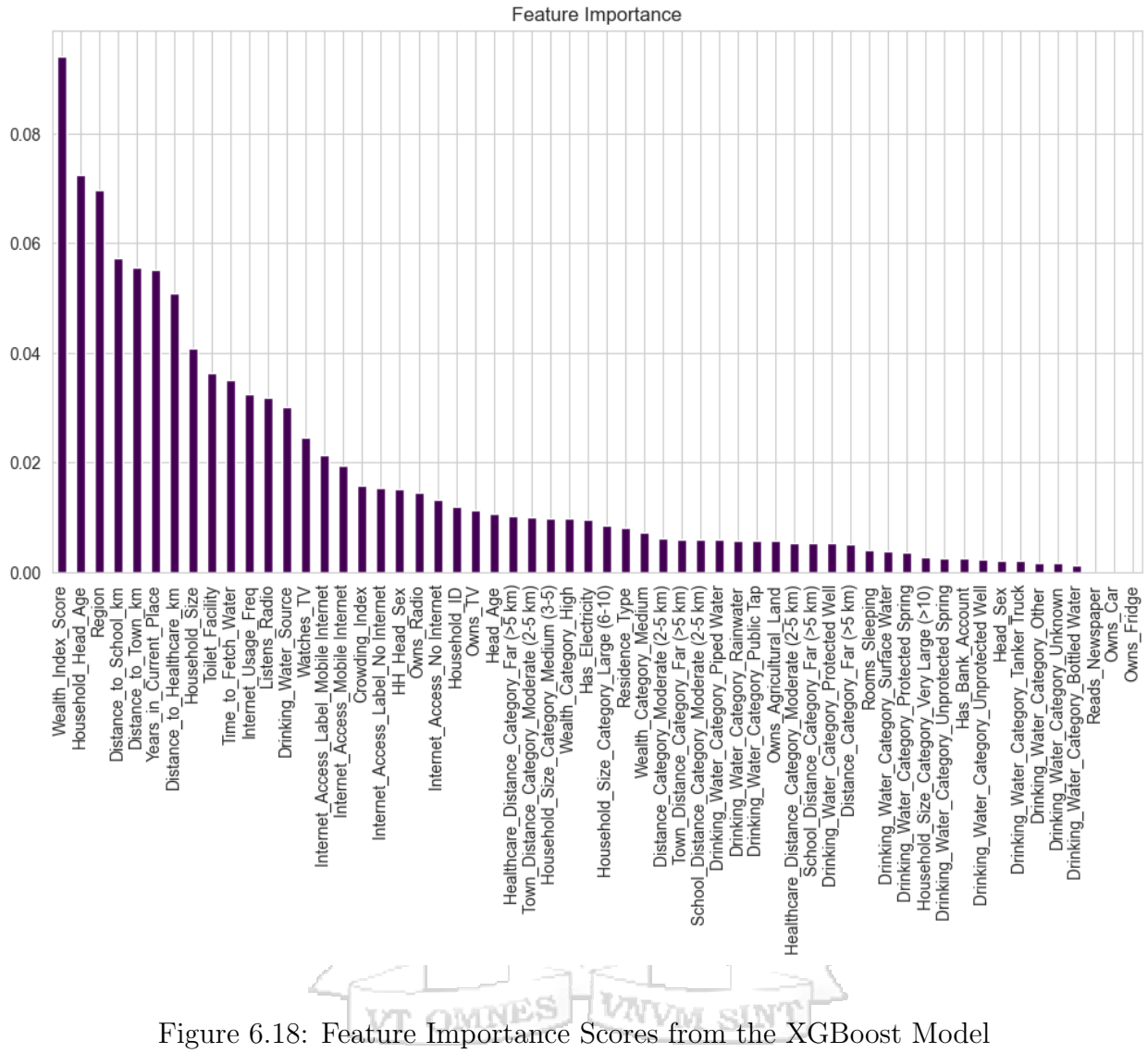


Figure 6.18: Feature Importance Scores from the XGBoost Model

6.5.2 Model Interpretation

To enhance the explainability of the predictions made by the XGBoost model, SHAP (SHapley Additive exPlanations) values were employed. Figure 6.19 presents the SHAP summary plot, which offers insight into the contribution and direction of influence each feature has on the model's output.

The most influential features include `Region`, `Internet Usage Frequency`, `Listens to Radio`, and `Wealth Index Score`. These variables demonstrate high SHAP values, indi-

cating that variations in these features lead to significant changes in the predicted education level. For instance, higher wealth index scores generally increase the likelihood of higher education levels, while regional disparities significantly impact model outcomes, validating the inclusion of region as a categorical predictor.

Color gradients further illustrate the value ranges of each feature, where red indicates higher feature values and blue represents lower ones. This visual distinction reveals important interactions—such as households with high internet usage and frequent media exposure being positively associated with higher educational attainment. By interpreting the SHAP plot, stakeholders can understand not only which features are most predictive but also how they influence the model's decisions.

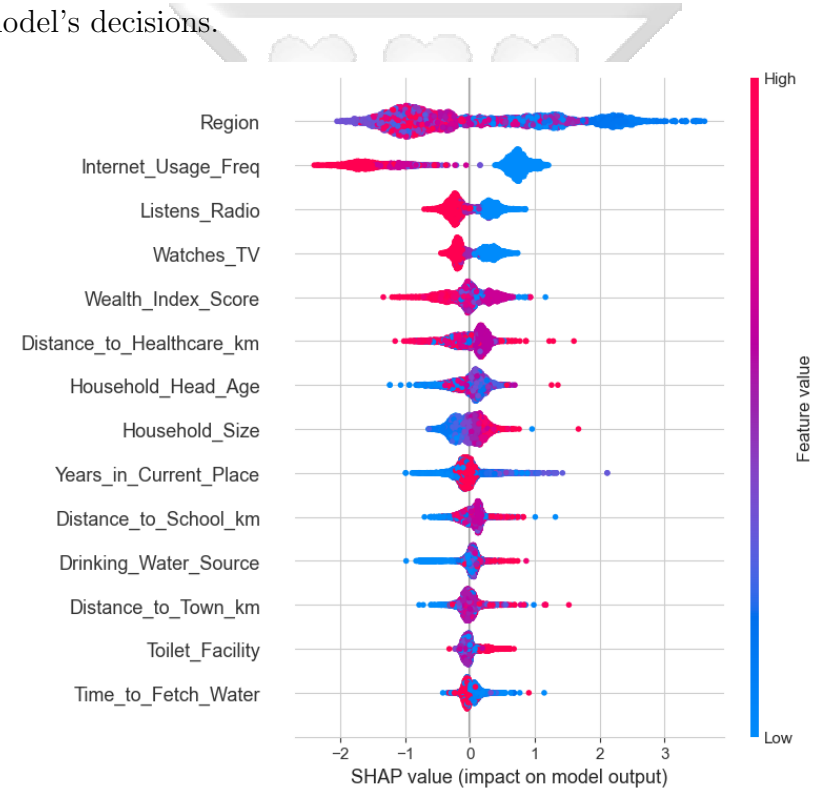


Figure 6.19: SHAP Summary Plot Showing Feature Impact on Model Output

While SHAP (SHapley Additive exPlanations) was employed to interpret the XGBoost model's outputs, it is important to acknowledge that such post-hoc interpretability methods have inherent limitations. XGBoost, as a powerful ensemble method, captures complex non-linear interactions, but these interactions are often difficult to explain in simple, actionable

terms for non-technical stakeholders. SHAP values offer local approximations but can be computationally expensive and sensitive to correlated features. As such, while SHAP improves transparency, the broader challenge of explaining model behavior in policy-sensitive environments remains, especially when results are used to make equity-driven decisions

6.6 Summary

This chapter presented the results of the machine learning modeling process, providing insights into data understanding, model performance, optimization, and interpretation. Among the three evaluated models—Decision Tree, Random Forest, and XGBoost—the XGBoost model consistently achieved the highest accuracy (83.3%) and F1 scores, making it the most suitable for predicting education levels in Kenyan households.

Feature importance analysis identified *Wealth Index Score*, *Region*, *Household Head Age*, and *Distance to School* as top predictors, reinforcing the influence of socio-economic and spatial factors on educational attainment. Model interpretation using SHAP further highlighted how these variables impact predictions at the individual level, enhancing transparency.

Hyperparameter tuning led to an improved XGBoost accuracy of 86%, demonstrating the value of model optimization. Overall, the results support the effectiveness of machine learning in uncovering key patterns in education data, offering a reliable foundation for data-driven education policy formulation.

While this study leverages advanced techniques such as SMOTE to address class imbalance, the ethical implications of oversampling must be acknowledged. Synthetic instances may inadvertently reinforce biases or misrepresent minority populations if not carefully validated. Additionally, the use of GPS-based proximity as a proxy for access to services carries limitations; it does not account for real-world barriers such as infrastructure quality, transportation availability, or service usability. Finally, although XGBoost was selected for its superior performance, its complexity poses interpretability challenges. Even with the integration of SHAP values to enhance transparency, the interactions within the model may still be diffi-

cult for non-technical stakeholders to interpret and act upon. These limitations highlight the importance of cautious deployment and stakeholder collaboration during implementation.

Moreover, the system's deployment context raises concerns around the validity of predictions in real-world conditions. The current model has not been empirically validated with end-users or tested in an operational environment. This limits the generalizability of results and raises the risk of misclassifying vulnerable households, which could lead to misguided interventions. Real-world validation through pilot testing, user feedback, and continuous monitoring would be essential to ensure the model's reliability, fairness, and contextual accuracy in deployment scenarios.

Chapter 7: Conclusions, Recommendations and Future Work

7.1 Conclusion

This study has demonstrated the potential of machine learning in advancing data-driven insights for improving education outcomes in Kenya. By applying models such as Decision Trees, Random Forest, and XGBoost to household survey data, we were able to accurately predict the highest level of education attained by household members based on socio-economic, geographic, and service accessibility indicators.

Among the models evaluated, XGBoost emerged as the top performer, achieving an accuracy of 83.3% before optimization and 86% after hyperparameter tuning. The model's predictive strength, coupled with interpretability through SHAP values and feature importance analysis, provided actionable insights into the most influential factors affecting educational attainment—including wealth, region, distance to school, and household head characteristics.

The deployment of the model via a user-friendly web portal further underscored the practical relevance of this research, allowing stakeholders to input household-level data and receive real-time predictions along with visual interpretations. These findings contribute to the growing body of work that positions machine learning as a transformative tool for evidence-based policy formulation in the education sector.

Despite the promising results of the developed model, several limitations must be acknowledged. One key concern is covariate drift — the possibility that the relationship between predictors and educational attainment may change over time due to evolving socioeconomic or policy environments. Additionally, the DHS sampling strategy, while nationally representative, may still introduce geographic sampling bias, especially in under-surveyed or remote regions. These limitations affect the generalizability of the model when deployed across different subpopulations or timeframes. Future iterations should consider periodic model

retraining and validation with new data to maintain predictive reliability across contexts.

Overall, this study affirms that integrating machine learning into national planning can support more targeted interventions and equitable resource allocation—helping to bridge education gaps, especially in underserved regions.

7.2 Recommendations

Based on the study's findings, several key recommendations are proposed to the Ministry of Education and relevant stakeholders to support equitable access to education across Kenya:

(i) **Targeted Regional Interventions:** Significant regional disparities in educational attainment suggest the need for region-specific policies. The Ministry should prioritize counties with persistently low education levels for enhanced support, including school infrastructure development and community-based learning programs.

(ii) **Expand Internet Connectivity:** Internet access was found to be a strong predictor of higher educational outcomes. Expanding affordable internet coverage—particularly in underserved rural and remote areas—can support digital learning, teacher training, and access to educational resources.

(iii) **Support for Economically Disadvantaged Households:** The wealth index score strongly influenced education attainment. The Ministry should strengthen financial aid schemes such as bursaries and school feeding programs to reduce economic barriers and keep children from low-income households in school.

(iv) **Leverage Media for Education Outreach:** Access to media—especially radio and television—was positively associated with higher education levels. The government should expand educational programming via public broadcast channels and support initiatives that provide solar-powered radios and TVs in low-access communities.

(v) **Improve Physical Access to Schools and Services:** Distance to schools, healthcare, and towns significantly affected educational attainment. Efforts should be made

to construct more schools within walking distance in remote areas and improve transportation options to reduce dropout rates and encourage consistent school attendance.

By acting on these recommendations, the Ministry of Education can address structural barriers and advance inclusive education for all segments of the population, aligning with the goals of equitable development and Vision 2030.

7.3 Future Work

While this study demonstrates the potential of machine learning in predicting educational attainment, several opportunities exist for further research and system enhancement:

(i) **Conduct usability validation:** To enhance real-world applicability, future work should prioritize engaging stakeholders such as the Ministry of Education, county education officers, and local NGOs through workshops and user feedback sessions. A proposed pilot use case would involve deploying the model as a decision-support tool in select counties with historically low educational attainment. Stakeholders can use the tool to identify high-risk clusters and allocate interventions accordingly. Feedback from these deployments would guide improvements to both model accuracy and interpretability, ensuring the solution aligns with policy-making needs.

(ii) **Incorporation of Longitudinal Data:** Future models can benefit from longitudinal education data to track household-level changes over time and improve predictive accuracy for policy planning.

(iii) **Integration with Real-Time Systems:** Developing a fully integrated dashboard for real-time predictions and education monitoring—linked with Ministry databases—can support dynamic policy responses.

(iv) **Expansion to Other Education Outcomes:** The model can be extended to predict related outcomes such as school dropout, performance scores, or transition rates between education levels.

(v) **Community-Level Feature Inclusion:** Incorporating more localized variables such as school quality, teacher availability, and community literacy rates could yield deeper insights into education drivers.

(vi) **User-Centered Design for Deployment:** Future versions of the web portal should incorporate user feedback from educators, parents, and policymakers to enhance usability and impact in real-world decision-making.

Continued research and development in these directions will further strengthen the role of data-driven tools in supporting inclusive and informed education policy across Kenya.

References

A. Bhutto, F. A., & Zafar, S. (2020). Educational data mining for student performance prediction using machine learning: A case study. *IEEE Access*, *8*, 76436–76451.

A. Jayaprakash, P. S., & Mandal, K. (2020). A review on educational data mining: Models and methods. *Journal of Advanced Computing*, *12*(3), 48–63.

Albelbisi, N., & Yusop, F. D. (2021). Predicting student learning outcomes using xgboost: A case study in higher education. *International Journal of Emerging Technologies in Learning*, *16*(2), 94–106.

Baker, S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), 3–17.

Bradley, A. P. (2019). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *44*(7), 1404–1419.

Breiman, T. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brownlee, J. (2019). *Xgboost with python: Gradient boosted trees for machine learning*. Machine Learning Mastery.

Chen, T., & Guestrin, C. (2020). Xgboost: A scalable tree boosting system. *Communications of the ACM*, *63*(10), 113–120.

E. Ndung'u, L. G., & Mwangi, M. (2021). Improving school dropout predictions using ensemble methods in kenya. *International Journal of Data Science and Analytics*, *10*(1), 45–58.

Fernández, A., García, S., & Herrera, F. (2019). Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, *61*, 863–905.

Ferrari, G., Pinetti, E., & Salini, S. (2020). Multi-class classification through roc surfaces and performance visualization. *Statistical Methods & Applications*, *29*(3), 565–582.

Goutte, C., & Gaussier, E. (2019). A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. *Information Processing & Management*, *56*(5), 1238–1249.

Haixiang, G., Yun, L., & et al. (2019). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220–239.

Hand, D. J. (2021). Measuring classifier performance: A coherent alternative to the area under the roc curve. *Machine Learning*, *110*(1), 131–145.

Hossin, M., & Sulaiman, M. (2019). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, *9*(2), 1–11.

I. Goodfellow, Y. Bengio, and A. Courville. (2016). *Deep learning*. MIT Press.

J. Han, M. K., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd). Elsevier.

J. Musso, P. K., & Cascallar, S. (2020). Predicting students' academic performance and socio-demographic variables: A machine learning approach. *Frontiers in Psychology*, *11*, 1–14.

J. Xu, W. H., & Huang, C. (2019). Using machine learning to predict student performance: A case study from china. *Journal of Educational Technology & Society*, *22*(4), 173–186.

Kenya National Bureau of Statistics. (2021). *Economic survey 2021: Socio-economic indicators*. Kenya National Bureau of Statistics. Retrieved October 2, 2024, from https://www.knbs.or.ke

L. Hasan, A. D. S., & Kapoor, M. (2019). An approach for student performance prediction using machine learning. *International Journal of Innovative Research in Computer and Communication Engineering*, *7*(3), 1255–1262.

Loh, W. Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 14–23.

Luque, A., Carrasco, A., Martín, A., & Herrero, E. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, *91*, 216–231.

Mujtaba, G., Anwar, S., & Shah, S. M. A. (2021). Evaluation metrics for classification problems: A survey. *Neurocomputing*, *450*, 335–347.

Murthy, S. K. (1998). Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery*, *2*, 345–389.

Mwangi & Kimenyi. (2019). Household wealth and secondary school performance: Evidence from kenyan counties. *Kenya Economic Research Institute*.

Omondi, A., & Obura, M. (2020). Predicting academic performance in kenya using neural networks: A comparative study of secondary school students. *Journal of Educational Research*, *20*(2), 35–45.

Opitz, D., & Maclin, R. (2019). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, *52*, 939–983.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106.

Romero, R., & Ventura, S. (2020). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics*, *40*(6), 601–618.

Saito, T., & Rehmsmeier, M. (2019). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, *14*(3), e0214102.

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168–192.

UNESCO Institute for Statistics. (2020). *Global education monitoring report 2020: Inclusion and education – all means all*. UNESCO. Retrieved October 2, 2024, from http://uis.unesco.org

United Nations. (2021). Sustainable development goals: Goal 4 - quality education [Accessed: 2021-09-15]. https://sdgs.un.org/goals/goal4

United Nations Conference on Trade and Development. (2021). *Leveraging technology and innovation for educational development in developing countries* (Accessed: 2024-10-02). United Nations Conference on Trade and Development. https://unctad.org/publication/leveraging-technology-and-innovation

Wang, S., & Sun, J. (2020). Random forest-based prediction of learning outcomes in higher education. *IEEE Access*, *8*, 65230–65238.

World Bank. (2021). *Access to electricity and basic services in kenya* (World Bank Development Report). World Bank. Retrieved October 2, 2024, from https://www.worldbank.org/en/country/kenya

Zhang, Y., & Zhang, X. (2021). Hyperparameter tuning of xgboost for imbalanced classification. *Expert Systems with Applications*, *168*, 114171.

Zhou, Z.-H. (2021). Ensemble learning: A survey. *Artificial Intelligence Review*, *55*, 601–645.

Zhou, Z., Li, W., & Liu, Y. (2020). A modified smote algorithm based on gaussian distribution. *IEEE Access*, *8*, 174151–174160.

Appendices

Appendix A: Similarity Report

turnitin Page 2 of 105 - Integrity Overview Submission ID trn:oid:::2945:275128786

19% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

- 231 Not Cited or Quoted 17%
  Matches with neither in-text citation nor quotation marks
- 34 Missing Quotations 2%
  Matches that are still very similar to source material
- 0 Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

12% Internet sources
9% Publications
15% Submitted works (Student Papers)

Integrity Flags

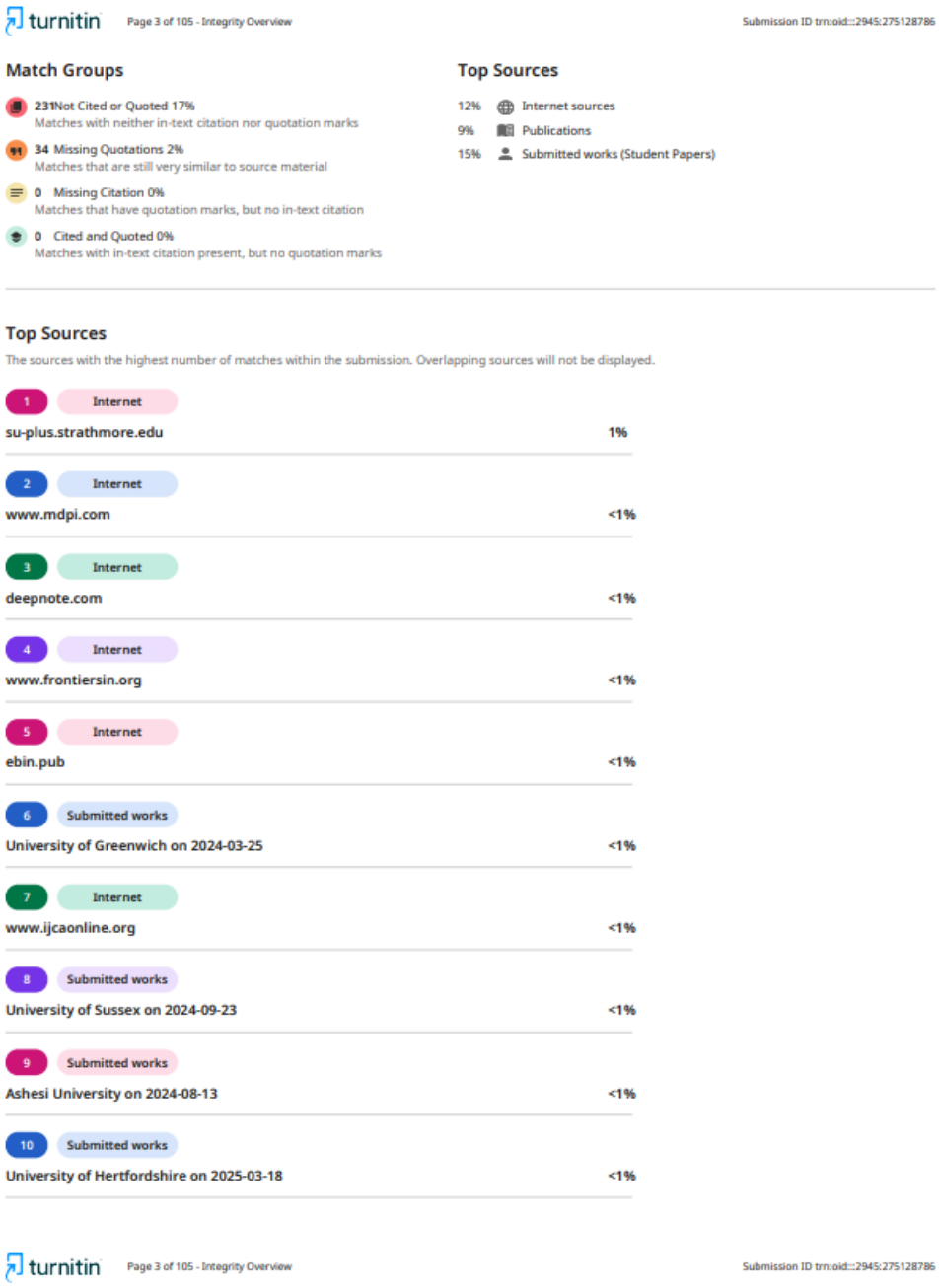
0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

turnitin Page 2 of 105 - Integrity Overview Submission ID trn:oid:::2945:275128786

Appendix B: Similarity Report



turnitin Page 3 of 105 - Integrity Overview Submission ID trn:oid:::2945:275128786

Match Groups

- 231 Not Cited or Quoted 17%
  Matches with neither in-text citation nor quotation marks
- 34 Missing Quotations 2%
  Matches that are still very similar to source material
- 0 Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

12% Internet sources
9% Publications
15% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	su-plus.strathmore.edu 1%
2	Internet	www.mdpi.com <1%
3	Internet	deepnote.com <1%
4	Internet	www.frontiersin.org <1%
5	Internet	ebin.pub <1%
6	Submitted works	University of Greenwich on 2024-03-25 <1%
7	Internet	www.ijcaonline.org <1%
8	Submitted works	University of Sussex on 2024-09-23 <1%
9	Submitted works	Ashesi University on 2024-08-13 <1%
10	Submitted works	University of Hertfordshire on 2025-03-18 <1%

turnitin Page 3 of 105 - Integrity Overview Submission ID trn:oid:::2945:275128786

Appendix C: Ethical Clearance Confirmation

Strathmore UNIVERSITY

19th December 2024

Ms Kemboi Stella,
stella.kemboi@strathmore.edu

Dear Ms Kemboi,

RE: Predicting Educational Attainment in Kenya: A Machine Learning Approach using Socioeconomic and Geographic Data

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2489/24.** The approval period is from **19th December 2024 to 18th December 2025.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu