Multilingual ISIC Code Classification System for Labour Force Survey Data Collection

An Automated Machine Learning Solution for Statistical Operations in Rwanda

---

Report Prepared By:

NGIRINSHUTI Fidele, Lead Data Scientist, model development, feature engineering, algorithm selection, and performance optimization

UMUHOZA Marie Ange, Data preprocessing, exploratory analysis, model evaluation, and validation

HATEGEKIMANA Emelyne, Streamlit application development, UI/UX design, and deployment

Institution:
National Institute of Statistics of Rwanda (NISR)

Date: January 2026

Contents

Executive Overview

This report summarizes the development and deployment of an automated system for assigning **International Standard Industrial Classification (ISIC Rev.4)** codes to business activity descriptions collected during Rwanda's **Labour Force Survey (LFS)**.

The system uses **machine learning and multilingual text processing** to classify free-text responses written in **Kinyarwanda, English, and French**. It was developed to address one of NISR's most expensive and time-consuming operations: manual ISIC coding of over **126,000 business descriptions**.

The model achieves **76.51% accuracy** across **396 ISIC codes** and produces **confidence scores** that allow high-quality automation of most cases. About **68% of records** can be coded automatically with **94% accuracy**, reducing manual workload by more than two-thirds while maintaining official statistics standards

Institutional Context

The National Institute of Statistics of Rwanda (NISR) relies on the Labour Force Survey to measure:

- Employment and unemployment
- Sectoral structure of the economy
- Informal sector size
- Gender and youth participation
- Skills and productivity trends

A key step in LFS processing is converting **text descriptions of economic activity** into **standard ISIC codes**, which are required for:

- International comparability (ILO, UN, World Bank)
- Policy design (industrial policy, agriculture, services)
- Monitoring Rwanda's Vision 2050

Without ISIC coding, the LFS cannot produce sector-based indicators.

The Operational Problem

Manual ISIC coding is a major bottleneck.

Additional problems:

- **Inconsistency**: Human coders agree only ~78%
- **Multilingual complexity**: Kinyarwanda, English, French
- **Training burden**: 2–3 weeks per new coder
- **Slow release**: Coding delays push publication beyond 6 months

This directly affects data quality, credibility, and policy relevance.

Project Objectives

The project aimed to modernize this process through automation.

Technical objectives

- Predict **4-digit ISIC codes** from free text
- Work across **three languages** without language detection
- Provide **confidence scores**
- Reach **≥70% accuracy**

Operational objectives

- Reduce manual coding by **at least 50%**
- Improve consistency
- Speed up data release

Institutional objectives

- Build NISR's **AI capacity**
- Create a **reusable framework**
- Enable future use for occupations and other classifications

Data and Preparation

The model was trained on **Rwanda LFS 2017-2023** data.

Item	Value
Records after cleaning	121,609
ISIC codes	396
Languages	Mixed
Avg. description	42 characters

Major challenges:

- Missing values
- Encoding errors
- Very **imbalanced ISIC distribution**
- Inconsistent human coding
- Many vague descriptions

After cleaning and filtering, a **70/30 stratified split** was used for training and testing.

Modeling Approach

The system converts text into numbers using two complementary techniques:

Word-level TF-IDF

Captures meaning:

- “construction”
- “retail trade”
- “food service”

Character-level TF-IDF

Captures morphology and spelling:

- “ubuhinzi” (farming)
- “construction”
- mixed-language words

This allows one model to work for **Kinyarwanda, English, and French**.

Three models were tested:

- Logistic Regression
- Random Forest
- **K-Nearest Neighbors (KNN)** → Best

Performance Results

Model	**Test Accuracy**
Logistic Regression	68.30%
Random Forest	65.10%
KNN	**76.51%**

KNN performed best because:

- It compares new descriptions with **similar past examples**
- It works very well with **text similarity**

The model performs extremely well on common activities such as:

- Agriculture
- Construction
- Retail
- Transport
- Education

Errors mostly occur between **closely related codes**, not unrelated sectors.

Confidence-Based Automation

Each prediction has a **confidence score**.

Confidence	**% of cases**	**Accuracy**
90–100%	47.90%	94.2%
70–90%	20.00%	81.7%
50–70%	18.00%	62.3%
Below 50%	14.10%	12–38%

This enables a **hybrid workflow**:

- **>70% confidence** → auto-code
- **50–70%** → supervisor review
- **<50%** → manual coding

This preserves quality while maximizing efficiency.

Operational Impact

Using the hybrid approach:

Measure	Manual	With AI
Records manually coded	126,817	~40,000
Person-days	840	269
Time saved	–	**571 days**
% Automated	0%	**68%**
Cost per 1,000	$31	$0.02

The model is:

- **300× faster**
- **Perfectly consistent**
- **Near-human accuracy**

Limitations and Future Direction

Limitations

- Rare ISIC codes lack training data
- Vague descriptions remain ambiguous
- No business context (location, size)
- KNN is hard to explain in simple words
- Depends on quality of human-coded data

Next Steps

- Integrate into NISR's survey systems
- Collect corrections and retrain annually
- Improve questionnaire design
- Apply model to occupations and other coding tasks

Conclusion

The Multilingual ISIC Classification System represents a **major modernization of Rwanda's Labour Force Survey operations**. It delivers:

- **Near-human accuracy**
- **Massive efficiency gains**
- **Better consistency**
- **Faster data release**