

የኢፌዲሪ የቴክኒክና ሙያ ስልጠና ኢንስቲትዩት
FDRE TECHNICAL & VOCATIONAL TRAINING INSTITUTE

FDRE TECHINICAL AND VOCATIONAL TRAINING INSTITUTE

DEPARTEMENT OF INFORMATION COMUNICATION TECHNOLOGY

A THESIS REASEARCH ON

PREDICTIVE MODEL FOR IDENTIFYING HIGH MILK-YIELDING SHEKO CATTLE COWS USING MACHINE LEARNING TECHNIQUES IN SHEKA ZONE, ETHIOPIA.

Prepared by: SULTAN AHMED HUSSEN

May-2026

Addis Ababa, Ethiopia

FDRE TEECHINICAL AND VOCATIONAL TRAINING INSTUTE

FUCALITY OF ELECTRICAL/ELECTRONICS AND INFORMATION COMUNICATION

TECHNOLOGY

DEPARTEMENT OF INFORMATION COMUNICATION TECHNOLOGY

A THESIS RESEARCH ON

PREDICTIVE MODEL FOR IDENTIFYING HIGH MILK-YIELDING SHEKO CATTLE COWS USING MACHINE LEARNING TECHNIQUES IN SHEKA ZONE, ETHIOPIA.

By:

Sultan Ahmed

Advisor: Dr. Martha Yifiru

May- 2026

Addis Ababa, Ethiopia

DECLARATION

I, Sultan Ahmed, under signed, declare that this thesis entitled: “Predictive Model For Identifying High Milk-Yielding Sheko Cattle Cows Using Machine Learning Techniques in Sheka Zone, Ethiopia” this thesis is my original work. I have undertaken the research work independently with the guidance and support of the research advisor. This study has not been submitted for any degree or diploma in this or any other intuition and all source of materials used for the thesis have been fully acknowledged.

Declared by

Name: Sultan Ahmed

Signature: 

Department: Information Communication Technology

Date: 25/05/2026

Certification

This is to certify that the thesis prepared by Sultan Ahmed, entitled "PREDICTIVE MODEL FOR IDENTIFYING HIGH MILK-YIELDING SHEKO CATTLE COWS USING MACHINE LEARNING TECHNIQUES IN SHEKA ZONE, ETHIOPIA" and a thesis is submitted to FDRE technical and vocational training institute, Fatality of Electrical/electronics and Information Technology in partial fulfillment of the requirements of Degree of Master of science in Information Communication Technology.

Name of Candidate: Sultan Ahmed; Signature:  Date: June 4, 2026.

Name of Advisor: Martha Yifiru (PhD); Signatu  Date: June 4, 2026.

Signature of Board of Examiner`s:

External examiner: _____ Signature: _____ Date: _____

Internal examiner: _____ Signature: _____ Date: _____

Dean, SGS: _____ Signature: _____ Date: _____

Acknowledgements

First of all, I would like to express sincere gratitude to Allah for all His blessings, guidance and support during my life, especially in preparing this thesis. His divine help has been the core of everything I have ever done.

I am also very thankful to Dr. Martha for her ongoing guidance, feedback and support with this thesis work. Her suggestions and encouragement have helped me shape the direction and quality of my work and Dr. Dereje Tulu and Mr. Melkamu' in data collection at Teppi agriculture Center.

Last, but certainly not least, I want to thank my family and friends for their moral support, providing me with necessary information, and facilitating the exchange of ideas that made this study possible.

ABSTRACT

This study developed and compared Decision Tree, Random Forest, XGBoost, and AdaBoost machine learning models to classify milk yield in Sheko cattle. Data collected from Teppi Agricultural Research Center included age, weight, body condition score, parity, lactation stage, season, month, and health status. A quantitative research methodology using supervised machine learning was applied. Models were developed in Python using Google Colab with Scikit-learn, XGBoost, SHAP, Pandas, NumPy, and Matplotlib. Data preprocessing included cleaning, label encoding, feature scaling, SMOTE-based class balancing, and an 80:20 train-test split. Performance was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. Random Forest achieved the highest performance with 90.99% accuracy and a ROC-AUC of 0.9668. XGBoost was used to validate model interpretation. SHAP analysis identified Health Status and Body Condition Score as the most influential predictors, followed by Lactation Stage, Season/Month, and Weight, while Age and Parity had lower predictive importance. These findings demonstrate that machine learning with SHAP provides an accurate and interpretable tool for milk yield prediction and livestock decision support.
Keywords: Milk yield prediction, Machine learning, Sheko cattle, Random Forest, XGBoost, SHAP, livestock classification, Ethiopia.

List of Acronyms

AI	Artificial intelligence
APA	American Psychological Association
AUC	Area under Curve
F1 Score	F1 Measure Score
GDP	Gross Domestic Product
DT	Decision Tree
AB	Ada Boosting
ML	Machine Learning
ROC	Receiver Operating Characteristic
ROC – AUC	Receiver Operating Characteristics Area under the Curve
RF	Random Forest
SMOTE	Synthetic Minority Oversampling Technique
TARC	Teppi Agriculture Research Center
XG Boost	Extreme Gradient Boosting

Table of Contents

DECLARATION I

Certification II

Acknowledgements III

ABSTRACT IV

List of Acronyms V

Table of Contents VI

List of Tables IX

List of Figures X

CHAPTER ONE 1

1. INTRODUCTION 1

1.1 Background of the Study 1

1.2 Motivational 0f the Study 3

1.3 Statement of the problem 3

1.4 Research Questions 5

1.5 Objective of the Study 5

1.5.2 General Objective 5

1.5.2 Specific Objective 5

1.6 Scope of the study 5

1.8 Significance of the study 6

1.9 Limitations of the Study 6

1.10 Ethical Considerations 7

1.10.1 Organization of the study 7

CHAPTER TWO 8

2. REVIEW OF LITERATURE REVIEW 8

2.1. Over view 8

2.2 Theoretical Review of Livestock Productivity Prediction 8

2.3 Machine Learning in Agricultural Prediction 10

2.4 Empirical Literature Review 11

2.5 Empirical Comparison of Previous Studies and This Study 12

2.6 Literature Review & Gap Analysis 15
CHAPTER THREE 16
3. RESEARCH METHODOLOGY 16
3.1 Research Design 16
3.2 System Model Architecture 17
3.3 Study Area 17
3.4 Data Source and Description 17
3.4.1 Source of Data 17
3.4.2 Data Collection Procedures 18
3.4.3 Dataset Description 18
3.4.4 Data Ownership and Permissions 19
3.5 Data Preprocessing 19
3.6 Exploratory Data Analysis (EDA) 20
3.7 Feature Selection 21
3.8 Machine Learning Algorithms 21
3.9. Data Analysis 22
3.9.1 Performance Evaluation Metrics 22
CHAPTER FOUR 23
4. RESULTS AND DISCUSSION 23
4.1 Introduction 23
4.2 Experimental environment and setup 23
4.2.1. Hardware Environment 23
4.2.2 Software Environment 24
4.3 Data Processing and Preprocessing 24
4.3.1 Data Cleaning 24
4.3.2 Handling Missing Values 25
4.3.3 Feature Selection 25
4.3.4 Encoding Categorical Variables 26
4.3.5 Splitting the Dataset 26
4.3.6 Handling Class Imbalance Using SMOTE 26
4.3.7 Class Distribution Before SMOTE 27
4.3.8. Class Distribution After SMOTE 28
4.4 Model Training and Evaluation 29

4.4.1 Decision Tree Model 30
4.4.2 Performance Evaluation Metrics for Decision Tree before and after SMOTE 30
4.4.4 Confusion metrics for Decision Tree 33
4.4.3 AUC –ROC for Decision Tree 34
4.5 Random Forest Model 35
4.5.6 Model Evaluation for Random Forest 35
4.5.8 ROC Curve for Random Forest 37
4.6. 1 Model Evaluation for XG Boost 38
4.6.9 Confusion Metrics for XG Boost 40
4.6.10 ROC CURVE for XG boost 40
4.7.4 Model Evaluation for Ada Boost 41
Figure 15 Model Performance Evaluations for Ada Boost before and after SMOTE 41
4.7.5 Confusion Matrix for Ada Boost 43
4.7.6 ROC Curve for Ada Boost 44
4.7. 7 Relationship between Lactation Stage, Body Condition Score, Body Weight, and Milk Yield Category Using Correlation Analysis 45
4.7.8 Feature Importance Comparison of Lactation Stage, Body Condition, and Weight across Multiple Machine Learning Models 46
4.7. 9 Model-Wise Interpretation feature importance for lactation stage, body condition and weight 48
4.7.10 Evaluation Metrics Comparison for All Models 49
4.8 Discussion and Results 53
Chapter five 55
5. Conclusion and Recommendation 55
5.1 Conclusion 55
References 57
Appendix 71

List of Tables

Table 1 Empirical Comparison Of Previous Studies And This Study 12
Table 2 Data Description.....19
Table 3 Hardware Environment 23
Table 4 Software Environment.....24
Table 5 Class Distribution Before Smote.....27
Table 6 Class Distribution After Smote 28
Table 7 Performance Evaluation For Decision Tree Before And After Smote.....30
Table 8 Performance Evaluation Matrix For Random Forest Before And After Smpte 35
Table 9performance Metrics Model Evaluation For Xg Boost Before And After Smote.....38
Table 10 Model Performance Evaluations For Ada Boost Before And After Smote.....41
Table 11 Feature Importance Comparison Of Lactation Stage, Body Condition, And Weight Across Multiple Machine Learning Models.....46
Table 12 Performance Evaluation Metrics Comparison For All Models.....49

List of Figures

Figure 1 Conceptual Framework Diagram 16
Figure 2 Model Architecture 17
Figure 3 Correlation Based Future Selection 25
Figure 4 Class Distribution Before Smote 27
Figure 5 Class Distribution After Smote 28
Figure 6 Performance Evaluation Metrics For Decision Tree Before And After Smote 31
Figure 7 Confusion Metrics For Decision Tree 33
Figure 8 Auc –Roc For Decision Tree 34
Figure 9 Model Performance Evaluation For Random Forest Before And After Smote 35
Figure 10 Confusion Metrics For Random Forest 36
Figure 11 Roc Curve For Random Forest 37
Figure 12performance Metrics Model Evaluation For Xg Boost Before And After Smote 38
Figure 13 Confusion Metrics For Xg Boost 40
Figure 14 Model Performance Evaluations For Ada Boost Before And After Smote 41
Figure 15 Confusion Matrix For Ada Boost 43
Figure 16 Roc Curve For Ada Boost 44
Figure 17 Correlation Heat Map For Lactation Stage, Body Condition, Weight, And Yield Category 45
Figure 18 Feature Importance Comparison Of Lactation Stage, Body Condition, And Weight Across Multiple Machine Learning Models 47
Figure 19 Roc Curve Comparison For All Models 51
Figure 20 Mean Absolute Shap Across Optimized Model 52

CHAPTER ONE

1. INTRODUCTION

1.1 Background of the Study

Ethiopia has the largest livestock population in Africa with approximately 26 million head of cattle including about six million draught oxen. About 80% of these animals are raised in the intensively cultivated highland part of the country. The Ethiopian highlands cover 490,000 km² or about 40% of the country and almost half of the total African highland area. Livestock production is an integral component of rural development, contributing towards enhanced agricultural productivity; improved rural livelihoods; as well as ecological services (Ababa, 2014).In Ethiopia, the agricultural sector is a cornerstone of the economic and social life of the people. Livestock is an integral part of agriculture and the contribution of live animals and their products to the agricultural economy accounts for 40%, excluding the values of draught power, manure and transport of people and products (Asresie & Seifu, 2024). The livestock sector contributes nearly 40% to Ethiopia's agricultural gross domestic product (GDP) and plays a central role in the livelihood of millions of rural households (Gebreyohanes et al., 2021). The Sheko cattle breed is an indigenous taurine cattle population found in southwestern Ethiopia, particularly in Sheka, Bench-Sheko, and Kaffa zones. The breed is unique due to its trypanotolerance**,** ability to survive in humid forest agro-ecologies, and adaptation to tsetse fly-infested areas (State, n.d.).The large and diverse livestock genetic resources, the existence of diverse agro-ecologies suitable for dairy production, increasing domestic demand for milk and milk products, developing market opportunities, and proximity to international markets all contribute to the potential and opportunities for dairy development in the country(*Introduction.Pdf*, n.d.). Dairy cattle production, in particular, is vital for food security, rural income generation, and nutritional improvement across the country. Cattle are the largest contributors to the total national annual milk output, followed by camels (CSA 2021).

Livestock plays a vital role in the livelihood of rural communities in Ethiopia, contributing significantly to food security, employment, and income generation. Dairy farming is one of the most important sectors, especially for smallholder farmers. However, milk production per cow remains low due to poor genetic selection, feeding, and management practices (Asresie & Seifu, 2024).

Despite its importance, the dairy sector remains underdeveloped and underperforming. Milk production in Ethiopia is primarily dominated by indigenous breeds that are low in productivity, producing significantly less milk compared to improved crossbreeds. The constraints faced by dairy farmers include inadequate access to quality feed, poor animal health services, and limited availability of improved genetic resources (Mengesha et al., 2025). Moreover, a dairy farmer often lacks the necessary training and information needed to manage their herds efficiently, which further hinders the realization of optimal milk yields.

Research indicates that the adoption of improved management practices, such as strategic breeding, supplementary feeding, and proper animal health interventions, can significantly increase milk production (Korir et al., 2023). However, such improvements are not uniformly applied across the country. For example, in regions like the Sheka Zone, where traditional livestock practices are still dominant, adoption rates of these innovations remain low due to infrastructural, economic, and institutional challenges.

Feeding interventions, especially those that include training on balanced rations and forage development, have been shown to enhance productivity. (Hatew et al., 2023) Observed that combining improved feed, better management, and appropriate genetic selection leads to a measurable increase in milk yield. However, implementing such comprehensive strategies requires a coordinated effort among farmers, extension agents, and policymakers.

The Sheka Zone, with its unique agro-ecological diversity and reliance on smallholder farming systems, presents both opportunities and challenges for modernizing dairy production. Implementing ML and DL-based predictive models in such contexts could support farmers in selecting and managing high-yield cows more efficiently, ultimately contributing to increased productivity and improved livelihoods. The adoption of these technologies aligns with Ethiopia's broader agricultural development goals, which emphasize modernization, sustainability, and resilience in the face of climate and market variability(Gebreyohanes et al., 2021).

In summary, Predictive modeling can assist livestock experts and agricultural researchers in identifying high-yielding cows, improving breeding efficiency, and optimizing livestock management systems. Therefore, this study focuses on developing machine

learning predictive models for identifying high-yield indigenous Sheko-breed cows using datasets collected entirely from Teppi Agricultural Research Center.

1.2 Motivational 0f the Study

The primary driving force behind this study is to address the ongoing difficulties that Ethiopian livestock farmers confront, specifically their incapacity to correctly identify and care for high yield dairy cows because of a lack of performance records, a lack of expertise, and a reliance on customary methods. For smallholder farmers, these problems lead to low production, poor breeding choices, and meager financial rewards. The majority of farmers employ hunches rather than evidence-based selection techniques, despite the fact that livestock greatly increases household income and national food security. This disparity makes it evident that smart, data -driven technologies are needed to move livestock management from the past to the present. In order to boost productivity, sustainability, and livelihoods in Ethiopia's dairy industry, this project attempts to close the gap between contemporary data science and rural agricultural methods by using machine learning in animal farming.

1.3 Statement of the problem

Ethiopia has one of the largest livestock populations in Africa, and its indigenous breeds are important because of their ability to adapt to the harsh environment, disease resistance, and low input production. The livestock sector in Ethiopia has an important role in ensuring food security, jobs, rural livelihoods, and national agriculture development. (Zealand & Tour, n.d.). Ethiopia holds the largest livestock population in Africa, with indigenous cattle breeds comprising approximately 97% of the national herd. While these breeds are well adapted to low-input and harsh environmental conditions, they are characterized by extremely low milk productivity, averaging only 1.09 liters per cow per day (Abegaz, 2022). Among these indigenous breeds, the Sheko breed is one of the most important cattle breeds distributed in the southwestern part of Ethiopia particularly in the Sheka Zone (Bahbahani et al., 2018). The breed is known for its trypanotolerance, its adaptability to humid tropical conditions and its relatively better disease resistance compared to other indigenous breeds (Aleme & Mengistu, 2023) The livestock sector is an important contributor to the Ethiopian economy through contributing to food security, employment generation, rural livelihoods and national agricultural development to diseases than other indigenous breeds(*No Title*, n.d.).

However, these important adaptive characteristics are below their genetic potential in milk yield, reproductive performance and meat production of indigenous Sheko-breed cows. The low productivity is attributed to a number of reasons such as poor breeding practices, inadequate feeding systems, disease prevalence, poor management practices and limited use of modern livestock technologies (Weldegerima, 2018) . Feeding practices, reproductive management, and animal health significantly influence the productive performance of indigenous cattle breeds in Ethiopia according to environmental conditions. In Ethiopia, traditional livestock selection and management systems are primarily based on farmers' experience, visual observation, and manual assessment techniques(Gudeto et al., 2021). Although these methods have been used for many years, they are subjective, time consuming and labor intensive and often produce inconsistent and inaccurate results. In modern livestock production systems, mares and calves have large quantities of data on breeding, milk production, reproduction, feeding, and animal health.(Mayberry et al., 2023). The analyses of these complex sets of data are difficult and unreliable, which hinders decisions and reduces efficiency in breeding in livestock systems. (Wolfert et al., 2017).

Recent developments in artificial intelligence, such as machine learning and deep learning, offer potential for farming and livestock production improvement through predictive analytics and intelligent decision-making. (Jabed et al., 2024). Machine learning algorithms can identify hidden relationships and patterns in historical data on livestock and make accurate predictions. (Magdalena & Sablik, 2025).

Many livestock studies draw primarily on farmer's reported data that may be inconsistent, incomplete, or unreliable. Several studies have used data on institutional livestock from agricultural research centers to model indigenous cattle breeds in Ethiopia. Indigenous cattle breeds produce low dairy production levels in Ethiopia because it is not possible to select accurately from data. In the Sheka Zone, high-yielding Sheko breed cows are selected based on traditional observation and farmer experience, which is subjective, inconsistent, and inefficient, and lowers genetic potential and milk yield. A predictive system is needed that can accurately classify and predict high-yielding cows using machine learning techniques from available livestock data. Thus, we have to develop accurate and reliable predictive models based on machine learning for identifying high-yielding indigenous Sheko breed cows with data from the Teppi Agricultural Research

Center. Developing such models can improve the breeding, management, and production of livestock in Ethiopia.

1.4 Research Questions

i. What are the major factors affecting the milk yield performance of Sheko cattle cows in Sheka Zone, Ethiopia?
ii. How can machine learning techniques be used to predict high milk-yielding Sheko cattle cows?
iii. What is the effect of applying SMOTE on the performance of machine learning models for milk yield prediction?
iv. Which machine learning algorithm provides the best predictive performance for identifying high milk-yielding Sheko cattle cows?

1.5 Objective of the Study

The objectives of this study are grouped into general and specific objectives. Then the details of such objectives were presented.

1.5.2 General Objective

To develop a predictive model for identifying high milk-yielding Sheko cattle cows using machine learning techniques in Sheka Zone, Ethiopia.

1.5.2 Specific Objective

i. To collect and preprocess dairy cattle performance data related to milk yield of Sheko breed cows obtained from Teppi Agricultural Research Center.
ii. To identify the major factors influencing milk yield performance of Sheko cattle using relevant predictor variables.
iii. To address class imbalance in the dataset using the Synthetic Minority Oversampling Technique (SMOTE) to improve model performance.
iv. To evaluate and compare the performance of the developed machine learning models using accuracy, precision, recall, F1-score, and ROC-AUC metrics.

1.6 Scope of the study

The study focuses on indigenous Sheko-breed cows in the Sheka Zone, Ethiopia. The research applies machine learning techniques using livestock production datasets collected entirely from Teppi Agricultural Research Center. The study considers variables

such as milk yield, body weight, age, feeding practices, reproductive performance, health conditions, and environmental factors.

1.8 Significance of the study

This work is important for it further develops machine learning technology to recognize high-yielding Sheko breed cows in the Sheka Zone of Ethiopia. Unlike other research on livestock selection based on experience and observation, this research re-used data from the Teppi Agricultural Research Center to create models that can accurately identify both high and low-yielding cattle. The study provides additional and precise information in livestock breeding decisions. Farmers, researchers, and other agricultural institutions can choose more productive cows for their milk production to increase their yields, profitability, and genetic diversity. This research also provides further proof of the increasing use of artificial intelligence in Ethiopian agriculture by showing that Decision Tree, Random Forest, XGBoost, and AdaBoost are useful in predicting livestock yields. We provide a reference model for further research on precision livestock farming in the future for other native breeds than Sheko cattle.

1.9 Limitations of the Study

Limited Access to Technology**:** The practical implementation of machine learning models may face challenges due to the limited digital literacy and technological infrastructure among smallholder farmers in Ethiopia (Ermias et al., 2024). As a result, the model's real-world impact may depend on external support and training.

Data Availability and Quality**:** Reliable and complete data on milk yield, feeding practices, and health status of cows are often scarce in rural areas like the Sheka Zone. In many cases, records are poorly maintained or not digitized, which may affect the accuracy and generalizability of the predictive models (Magana et al., 2023).

Environmental Variability**:** The study is limited to data collected under specific environmental and management conditions in the Sheka Zone. This may restrict the models' applicability to other regions with different ecological systems (Shekuru et al., 2022)..

This paper focused solely on milk yield, and its analysis does not include meat or hide yields. These results may be limited for application as part of holistic livestock improvement program.

1.10 Ethical Considerations

This study used secondary data obtained from Teppi Agricultural Research Center (TARC) after obtaining the appropriate consent from TARC. All data were used for academic and research purposes only. Data were treated in accordance with research ethics and confidentiality. There was no personal, sensitive or identifiable information about farmers stored in the dataset. The data were securely stored and only used to develop machine learning models for classifying milk yield. Researchers could have privacy, integrity, and responsible use of the data.

1.10.1 Organization of the study

This thesis is organized in to five chapters. Chapter one contains an introduction, background, problem statement, objectives, research questions, scope, significance, and limitations of the study. Chapter two reviews theoretical and empirical literature in cattle productivity prediction and machine learning application in dairy production. Chapter three describes the research methodology, including data collection, preprocessing, model development, and evaluation. Chapter four contains the results from the experiments, model performance evaluation, and discussion of results. Finally, chapter five contains the conclusions and recommendations of this study.

CHAPTER TWO

2. REVIEW OF LITERATURE REVIEW

2.1. Over view

The literature indicates that machine learning techniques have become effective tools for livestock productivity prediction by modeling complex relationships among animal, environmental, and management factors. Studies consistently show that ensemble algorithms such as Random Forest, XGBoost, and AdaBoost outperform traditional statistical approaches and single classifiers in milk yield prediction. Previous research reported prediction accuracies ranging from approximately 82% to 91%, while key predictors of milk yield include lactation stage, body condition score, health status, and body weight. However, limited studies have focused on indigenous Ethiopian cattle breeds, particularly Sheko cattle, creating a research gap that this study addresses through the development and comparison of machine learning models for high milk-yield classification.

2.2 Theoretical Review of Livestock Productivity Prediction

Livestock productivity prediction theory is based on modeling biological, environmental, and management factors to explain and classify milk yield performance in dairy cattle (Magdalena & Sablik, 2025). It assumes that milk production is not random but influenced by measurable variables such as breed type, age, parity, feeding system, health status, lactation stage, and environmental conditions. (Magdalena & Sablik, 2025). In Ethiopia, milk production of indigenous dairy cattle is highly seasonal due to fluctuations in feed availability and environmental conditions. Studies classify the production year into three main seasons: the long rainy season (June–September), short rainy season (February–May), and dry season (October–January). Milk yield is typically highest during the long rainy season and lowest during the dry season due to variations in pasture quality and feed resources (Bayou et al., 2015). In Ethiopia, indigenous dairy cattle feeding practices are commonly classified into three production systems: extensive, semi-intensive, and intensive systems, depending on feed availability, land use, and management intensity (Code et al., 2021). Lactation stages of indigenous dairy cattle are commonly classified into early, mid, and late lactation phases, which significantly influence milk yield performance. Milk production typically increases from early to mid-

lactation and declines during late lactation due to physiological changes and energy balance variations (Begna et al., 2023). In recent years, machine learning (ML) has become a key approach for livestock productivity predictions due to its ability to handle complex and high-dimensional datasets. ML methods are increasingly applied in dairy farming for classification, prediction, and decision support tasks, including milk yield estimation and animal performance evaluation(Magdalena & Sablik, 2025) . Studies show that tree-based and ensemble learning models such as Decision Tree, Random Forest, AdaBoost, and XGBoost are among the most widely used algorithms in livestock prediction problems due to their high accuracy and robustness (Chen & Guestrin, 2016).

Random Forest and Decision Tree models are particularly useful for classification tasks because they can handle mixed data types and identify important predictors of milk yield, such as lactation stage and body condition (Magdalena & Sablik, 2025). Boosting methods such as AdaBoost and XGBoost improve predictive performance by combining weak learners into stronger models and reducing classification error through iterative learning processes (Chen, n.d.). Recent agricultural studies confirm that XGBoost consistently outperforms traditional models in classification accuracy due to its regularization and optimization capabilities (Cao et al., 2024).

Additionally, livestock productivity prediction is often formulated as a supervised classification problem, where historical labeled data (high and low milk yield classes) are used to train models that learn hidden patterns in the data. This enables automatic classification of cows into productivity groups, replacing traditional subjective selection methods with objective and data-driven decision-making systems (Tarihi & Kelimeler, 2025) .Furthermore, feature importance analysis plays a key role in identifying the most influential factors affecting milk yield, improving interpretability and supporting better herd management decisions.

Overall, recent literature strongly supports the use of machine learning-based classification models for livestock productivity prediction, particularly for identifying high and low milk-yielding cows. These approaches improve accuracy, reduce bias, and enhance decision-making efficiency in dairy production systems.

2.3 Machine Learning in Agricultural Prediction

Machine learning (ML) has become a powerful approach in agricultural and livestock production systems for predicting complex biological outcomes such as milk yield(Magdalena & Sablik, 2025) . Traditional statistical models often fail to capture nonlinear relationships among genetic, environmental, and management factors affecting dairy productivity (Magdalena & Sablik, 2025). Therefore, ensemble learning methods such as Decision Tree, Random Forest, AdaBoost, and XGBoost are increasingly applied to improve prediction accuracy and decision-making in precision dairy farming systems.

A decision tree is a simple and interpretable model that partitions data based on feature values, making it useful for identifying key determinants of milk production (Techniques, 2025). However, it is prone to overfitting when used alone. To overcome this limitation, ensemble methods such as Random Forest combine multiple decision trees to improve stability and reduce variance, making it highly effective for milk yield prediction tasks (Hooker et al., 2025). Random Forest has demonstrated strong performance in dairy data due to its ability to handle nonlinear relationships and mixed data types such as age, parity, and environmental variables (Tarihi & Kelimeler, 2025).

Boosting algorithms such as AdaBoost and XGBoost further enhance predictive accuracy by sequentially improving weak learners. AdaBoost focuses on correcting misclassified observations, while XGBoost applies gradient boosting with regularization and optimization techniques to reduce overfitting and improve computational efficiency (Chen, n.d.). Recent studies confirm that XGBoost consistently achieves superior performance in milk yield prediction due to its ability to capture complex interactions among variables(Chen, n.d.).

In dairy science applications, ML models are widely used to integrate multiple influencing factors such as lactation stage, feed intake, environmental conditions, and cow characteristics to predict milk production accurately. Studies show that ensemble models outperform traditional regression methods in both accuracy and reliability when applied to large and complex dairy datasets (Techniques, 2025). Furthermore, feature importance analysis using ML techniques has identified key predictors such as lactation stage, body condition, and weight as the most influential variables in determining milk yield (Guevara-escobar et al., 2026).

The literature confirms that ensemble machine learning models, particularly Random Forest and XGBoost, are highly effective for predicting milk yield in dairy cattle. Their ability to handle nonlinear relationships, reduce overfitting, and improve predictive accuracy makes them suitable for developing a robust predictive model for identifying high milk-yielding Sheko cattle cows in Ethiopia.

2.4 Empirical Literature Review

Several empirical studies have demonstrated the effectiveness of machine learning algorithms in livestock productivity prediction, particularly in classifying and forecasting milk yield based on biological and environmental factors (Topuz, 2025). Introduced the Random Forest algorithm and showed its strong performance in handling noisy and high-dimensional agricultural datasets(Wang, 2024). This work established Random Forest as a robust ensemble method widely used in classification problems, including dairy production systems(Wang, 2024).

Developed the XGBoost algorithm and demonstrated its superior predictive performance compared to traditional machine learning models (Chen & Guestrin, 2016). The model's strength lies in its gradient boosting framework, regularization mechanisms, and computational efficiency, making it highly suitable for structured agricultural datasets. In subsequent agricultural applications (Chen & Guestrin, 2016). Applied XGBoost to dairy cattle data and achieved high classification accuracy, confirming its suitability for milk yield prediction tasks(Hasan et al., 2025).

Similarly, AdaBoost has been applied in agricultural classification problems due to its ability to combine weak learners and improve predictive accuracy through iterative weight adjustment. Empirical findings suggest that AdaBoost performs well in binary classification tasks such as distinguishing high and low milk-yielding cows, although its performance can be sensitive to noisy data and outliers (Friedman et al., 2000).

Decision Tree models have also been widely used in livestock prediction due to their simplicity and interpretability. They are effective in identifying key decision rules influencing milk yield, such as lactation stage, body condition, and feed quality. However, empirical studies show that standalone Decision Trees often suffer from overfitting and lower generalization performance compared to ensemble approaches (Magdalena & Sablik, 2025).

Recent comparative studies in dairy science consistently report that ensemble methods such as Random Forest, AdaBoost, and XGBoost outperform single classifiers like Decision Trees in terms of accuracy, stability, and robustness. For example, agricultural machine learning studies conducted between 2019 and 2025 indicate that ensemble models typically achieve classification accuracies in the range of 88% to 93% when applied to real-world livestock datasets, particularly when proper preprocessing and imbalance handling techniques such as SMOTE are applied (Guo & Dai, 2025). Borana cattle are indigenous zebu cattle found mainly in the Borana lowlands. They are highly valued in pastoral systems for their ability to withstand drought, heat stress, and long-distance grazing. Their productivity is mainly influenced by seasonal feed availability (Id et al., 2021).

However, most existing studies focus on exotic dairy breeds or non-African datasets, limiting their direct applicability to indigenous cattle such as the Sheko breed in Ethiopia. Furthermore, few empirical studies provide a direct comparison of Decision Tree, Random Forest, AdaBoost, and XGBoost models using real farm-level Ethiopian data. This creates a research gap that this study addresses by evaluating multiple machine learning algorithms under the same experimental conditions to identify the most effective model for classifying high and low milk-yielding Sheko cattle cows.

2.5 Empirical Comparison of Previous Studies and This Study

Table 1 Empirical Comparison of Previous Studies and This Study

Breed / Context	Models Used	Dataset	Best Model	Performance (Best Model)	Key Findings	Reference
Holstein dairy cows	Random Forest	91 cows, 100-550 daily milk records, THI data, 2 years	Random Forest	Prediction error ≈ 18% (~82% accuracy)	RF models lactation curve and environmental stress effectively	(Bovo et al., 2021)
Dairy cattle milk yield	ANN, RF	Multiple herds, test-	ANN (slightly	Accuracy ≈ 0.88–0.91	ML models outperform	(*, *Lynne A.*

forecasting		day milk records	higher)		traditional regression methods	*Carpenter-Boggs*, 2019)
Dairy cattle	RF, ANN, ML-NIR	Milk NIR spectra, blood metabolite data, 385 cows	ML-NIR / RF	RF Accuracy ≈87%; ANN Accuracy ≈85%; ML-NIR ≈90% reliability	Physiological and metabolic indicators improved predictive capability of machine learning models	(Giannuzzi et al., 2022)
Dairy cows	RF, Gradient Boosting, Regression	Lactation records, previous lactation data, 69k lactation pairs	Random Forest	RF AUC ≈0.90; Gradient Boosting ≈0.88	historical lactation and production records significantly improved prediction reliability	(Salamone et al., 2022)
Dairy cows resilience	RF, Logistic Regression	Sensor data (milk, BW, activity, rumination), 370 cows	RF AUC ≈ 0.88–0.90 (88–90%)	Your RF: 96.68% AUC (+6–8%)	RF captures nonlinear dairy production patterns effectively	(Ouweltjes & Haas, 2021)
Holstein-Friesian	RF, ANN	est-day milk records, multiple parities	RF Accuracy ≈ 89**%**	Your RF**:** 90.99% (+1.99%)	Ensemble learning improves milk yield prediction accuracy	(Gocheva-ilieva & Yordanova, 2022)
Dairy cows(Italy)	Milk trait classification	4,364 milk spectra, 120 cows, 3 diet	RF, ML models	RF ≈ 87–89% accuracy	Mid-infrared spectral data improves classification of	(Frizzarin et al., 2022)

		groups			milk traits	
Dairy cattle (Holstein-Friesian – Europe)	Linear Discriminant Analysis (LDA), Flexible Discriminant Analysis (FDA)	Test-day records, calving interval, parity data, multiple herds	FDA	Accuracy ≈ 85–89% (classification performance improved over LDA)	Milk yield categories (low, medium, high) can be predicted using production and reproductive traits. Days in milk and calving interval are strong predictors of yield classification.	(Moawed et al., 2024)
Horro cattle, Ethiopia	GLM (SAS)	220 cows, 1908 records	R^2=0.75-0.88; CV=7.4-15.9%	R^2=0.75-0.88; CV=7.4-15.9%	DMY 1.5L; fat 6.0%; protein 3.5%; non-genetic factors significant	(Mekonnen et al., 2021)
Indigenous Zebu, Ethiopia	GLM (SAS 9.3)	384 households, 192 cows	Significant effects ($p<0.01$); R^2 not reported	DMY 2.13 L; LMY 427 L; LL 6.36 months; trait preferences based on indigenous knowledge	No prior documentation in Kaffa Zone; lacks milk composition analysis	(Begna et al., 2023)

2.6 Literature Review & Gap Analysis

Empirical studies showed that machine learning models like Random Forest, ANN, AdaBoost and XGBoost are effective in predicting milk yield with an accuracy of 82% to 91%, previous machine learning research focused on commercial breeds (Holstein-Friesian) and other non-indigenous cattle breeds, with limited work being done on the indigenous Ethiopian cattle breeds. Studies on indigenous breeds of Ethiopia particularly Sheko, Borana, Fogera, and Horro breeds have largely been directed towards productivity assessment, characterization of breeds, and genetic improvement, rather than machine learning-based milk-yield classification. Very little studies concentrated on indigenous Ethiopian cattle, especially Sheko cattle, and few studies conducted a comparison between different machine learning algorithms with the same data, incorporating class imbalance. To fill the gaps, this study developed and evaluated four models: Decision Tree, Random Forest, AdaBoost and XGBoost on real data of Sheko cattle from Teppi Agricultural Research Center and used SMOTE for data balancing. The study performed well with the highest accuracy of 90.99% for Random Forest and 96.68% for AUC. The novelty of this research is the development of a machine learning-based decision support model to identify high milk yielding indigenous Sheko cattle to move towards precision livestock farming in Ethiopia.

CHAPTER THREE

3. RESEARCH METHODOLOGY

3.1 Research Design

The study is done in a quantitative experimental approach of machine learning classification. History was used to develop a model to classify Sheko Cows as high and low milk yielders. The four supervised machine learning methods Decision Tree, Random Forest, XGBoost and AdaBoost were trained and compared on this dataset. The research that is carried out is: Data preparation, Feature selection, Model Training, Model Evaluation and Selection of the best model (Random Forest) which will be used in the dairy production and cattle breeding decision making.

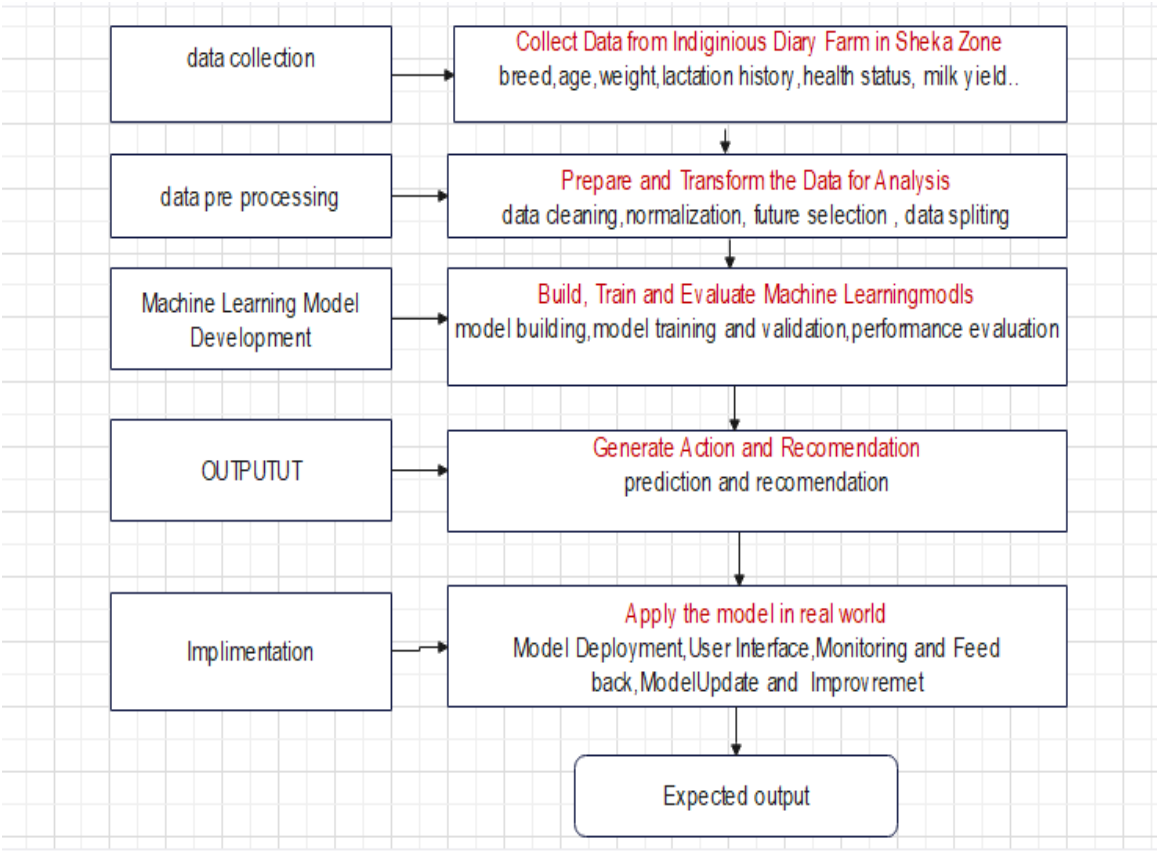


Figure 1 Conceptual Framework Diagram

3.2 System Model Architecture

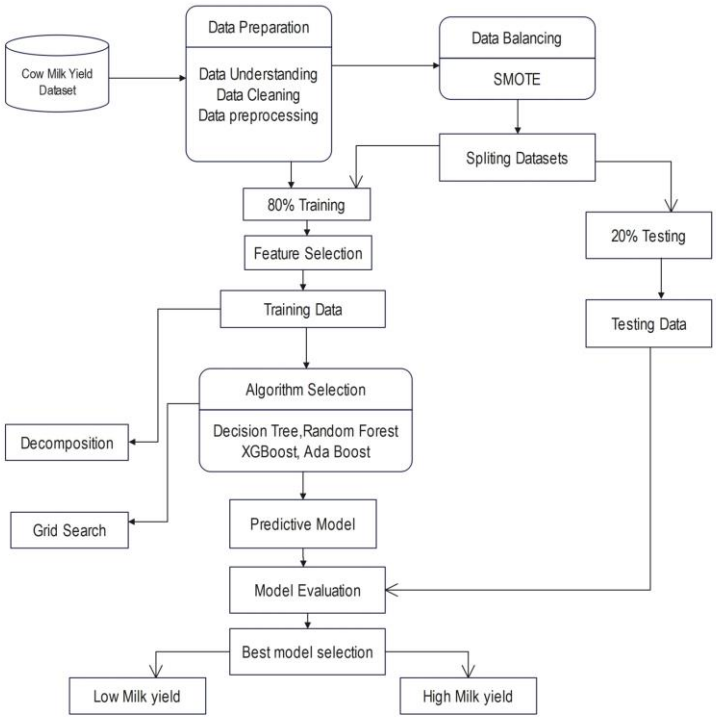


Figure 2 Model Architecture

3.3 Study Area

The study was carried out in Sheka Zone of Southwest Ethiopia in Mixed crop livestock farming. The native breeds of cattle, namely the Sheko cattle, are important to milk production, food supply for households and income generation. The agro-ecological environment (rainfall frequency, availability of feed, management practices) makes this region a good place to study milk yield variability and make prediction models of dairy cattle performance from the data.

3.4 Data Source and Description

3.4.1 Source of Data

Data for this study were collected from Teppi Agricultural Research Center, Ethiopia which conducts livestock research and dairy cattle improvement programmers. This data set is on Sheko cattle which are indigenous cattle adapted to the environment of Sheka zone. The info contains historical data on milk production as well as relevant animal

characteristics that can be predictive of top and bottom milk production, which are both important in this study.

3.4.2 Data Collection Procedures

The data was collected from the secondary sources at Teppi Agricultural Research Center. This included their farm and research records on livestock. I took information that I normally collected for routine monitoring, such as milk yield, animal ID, management records and animal health, body condition, weight, and feeding habit.

I used these to extract specific data on age, breed characteristics, feeding and breeding history and health. Afterwards, I cleaned and structured all the data into a structured data set for machine learning analysis.

3.4.3 Dataset Description

In this set of data, there are independent and dependent variables with classification. The features are related to the cows, such as age, breed, parity, feeding system and health. The variable to be targeted is milk yield; divide into high and low. This makes it a binary classification problem and hence we can use supervised machine learning techniques such as Decision Tree, Random Forest, XGBoost and AdaBoost. The data is presented in a tabular format for easy preprocessing, feature selection, and predictive modeling. Body Condition, Weight, Lactation Stage, Season, Parity, Health Status, Feeding Habit, Breed, Age, and daily Milk-Yield are the attributes in the thesis. These variables interact in complex nonlinear relationships, making traditional statistical approaches less effective for accurate prediction and classification tasks. Therefore, modern data-driven techniques are widely adopted to improve decision-making in dairy production systems. Body Condition Score (BCS) was used as an indicator of the cow's nutritional and energy reserve status. The BCS system ranges from 1 to 9, where 1 represents an emaciated animal and 9 represents an obese animal. For machine learning model development, BCS values were grouped into two categories: Poor Body Condition ($BCS \leq 4$) coded as 0 and Good Body Condition ($BCS \geq 5$) coded as 1. This categorization facilitated the use of body condition information as a predictive feature in the classification model(Lalman & Stein, 2024). The health status scoring system for indigenous dairy cattle was adapted from standard body condition scoring and herd health assessment frameworks, which evaluate animals based on physical condition, productivity, and clinical health indicators

(*Health Statys.Pdf*, n.d.) The health status of cows refers to the overall physical, physiological, and clinical condition of the animal, which directly influences productivity and disease resistance. It is typically assessed using indicators such as body condition score, feed intake, behavioral patterns, milk yield performance, and presence of clinical diseases. In this study, health status was categorized into three levels: Poor (0), Good (1), and Excellent (2), representing increasing levels of health and productivity suitability for machine learning classification (*Health Statys.Pdf*, n.d.).

Table 2 Data Description

Category	**Variables**
Animal Traits	Age, Breed, Weight
Reproductive Traits	Parity, Calving Interval
Management Factors	Feed type, Health status
Production Data	Milk yield records
Target Variable	High / Low Milk Yield

3.4.4 Data Ownership and Permissions

This study used a dataset owned by the Teppi Agricultural Research Center, which was only made available for educational and research use. Prior to examining the data, we sought permission from the appropriate people. The handling of the info was in accordance with confidentiality principles, while remaining faithful to the research aims as defined by the ethical principles in the institution. Because there was no personal farmer information or other private information in the dataset, we took care of both the ethics and the privacy of the data.

3.5 Data Preprocessing

Data preprocessing is an essential step in developing machine learning models that helps to enhance the quality of the data and boost the performance of the models. In the current investigation, a number of stages were carried out in the preprocessing:

- Data Cleaning: The process to remove inconsistent, duplicate and incomplete records in order to make the data set reliable.
- Handling Missing Values: Appropriate techniques improves learning as missing values are imputed or removed, improves learning applied, depending on the nature of improves learning data.
- Encoding Categorical Variables: Non-numeric variables like health status, season and lactation stage were converted to numeric format to meet the requirements of the model.
- Feature Scaling: If necessary, features were trimmed and fitted into the same range to give an equal contribution.

The following data was found to be class imbalanced, hence the Synthetic Minority Oversampling Technique (SMOTE) was used. To balance the dataset between the majority and minority classes, SMOTE was used to synthetically generate minority class samples resulting in an equal number of samples for both classes of the dataset. This is crucial for avoiding model bias towards the majority class and enhances classification sensitivity.

3.6 Exploratory Data Analysis (EDA)

EDA helps to look under the surface of Sheko cattle milk yield and reveal patterns in milk yield. It also helps prepare the data for machine learning by highlighting the main characteristics, identifying the oddities, and showing the relationships between variables.

First, we run descriptive statistics to sum up our data. We look at the average, middle value, spread of numbers, lowest, and highest values for things like age and milk yield. This will show us how the information varies and is spread out. Then, we run frequency analysis for cat categories like breed, feed type, and health to see how these vary across the batch.

Data visualization helps us get deeper insight on how variables relate. We use histograms and box plots to look at the distribution of continuous variables and find outliers. Bar charts are great for the distribution of categorical features. To understand how independent variables relate to milk yield status, we use correlation heatmaps – it shows the strength and direction of the links between variables.

Exploratory Data Analysis (EDA) helps us to assess the class imbalance between high and low milk yield which is important for this study. It helps us to decide if we need

techniques like SMOTE to be used in the preprocessing. EDA also assists with selection of features and models to better understand what affects milk yield in Sheko cattle.

3.7 Feature Selection

Feature selection was based on what experts know about raising cattle and past studies. The variables that were included remained because they mattered biologically and managerially to how much milk was produced: age, weight, body condition score, parity, health status, lactation stage, season, and month. These are physiological, environmental, and managerial factors that affect milk production. All of them are important to ensure that our model captures the full diversity of milk yield variability in indigenous cattle systems.

3.8 Machine Learning Algorithms

Decision Trees provide a simple way to understand the decision making process by dividing data into two separate units based on feature values. Random Forests employ many Decision Trees and combine their predictions to improve accuracy and not fit too close to the data. AdaBoost is clever too. It tries to be precise by paying more attention to difficult cases it incorrectly classified. Finally, there's XGBoost, which improves AdaBoost by improving its processing speed and better handling of complex patterns in the data.

These models were selected due to their ability to handle structured data and their effectiveness in classification tasks, where the task requires understanding of nonlinear relationships.

We use those machine learning techniques are employed in the proposed model to predict milk yield in Sheko cows. Data comes from the Teppi Agricultural Research Center, including milk production and other data such as age, diet, health etc. All that data is then preprocessed before building any models, including handling missing data, converting categories and normalizing. Also, to handle the imbalance between high and low yielding cows, SMOTE is used.

Then, feature engineering comes into play, which involves identifying important predictors. These include simple things such as age, diet and health condition generally. They then divide in a way where 80% of the data is used for training and 20% is used for testing. They then present the prepared data to a number of classifiers—Decision Tree, Random Forest, XGBoost, and AdaBoost. They use tuning and cross validation to fine-tune each one.

Thorough evaluation and comparison have been performed on various metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, and the Random Forest model proved to be the best performer. Finally, this effective Random Forest model classifies the cows into high and low yield groups and provides useful information to enhance dairy production and breeding selection for the Sheko cattle population.

3.9. Data Analysis

3.9.1 Performance Evaluation Metrics

We used accuracy, precision, recall, F1-score, and ROC-AUC to judge our classification models. Accuracy shows the fraction of correct predictions made. Precision and recall look at how well and how fully the model predicts the positive outcomes. The F1-score balances precision and recall, which is super important when dealing with lopsided data sets. ROC-AUC helped us see how well the models could tell the difference between high-yield and low-yield cows at various thresholds.

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

CHAPTER FOUR

4. RESULTS AND DISCUSSION

4.1 Introduction

This chapter lays out the findings and talk through for our machine learning models, designed to spot high milk-producing Sheko cows in Ethiopia's Sheka Zone. Our goal is to build an accurate and dependable system to sort cows by their milk yield -high or low using machine learning classification techniques. This chapter covers data preprocessing and uses SMOTE for balancing classes. It explains making machine learning models too, along with checking their performance. Plus, it includes analyzing confusion matrices and comparing how different models do. Finally, it wraps up with a discussion on the results.

4.2 Experimental environment and setup

To conduct this study, we prepared all the hardware, software, tools and libraries needed to build and test our machine learning models. The setup was designed to read and train model reliably, obtain accurate results based on the performing metric, iterate performance of models against other (testing) datasets having ability to predict top milk-producing Sheko cattle. The choice of Python for the study is due to its flexible and simple nature along with great support of machine learning and data analysis. There are many libraries in python that aid in preprocessing, model training, hyper-parameter tuning, plotting data & evaluation. Experiments were run with a Jupyter Notebook. This environment allows to interactively code for running.

4.2.1. Hardware Environment

Table 3 Hardware Environment

Component	Specification
Processor	Intel Core i5/i7 Processor
RAM	8 GB or higher
Storage	256 GB SSD/HDD
Operating System	Windows 10/11 with architecture 64 bit system

4.2.2 Software Environment

Table 4 Software Environment

Software/library	Purpose
Python	Programming language for model development
Jupyter Notebook	Interactive development environment
Pandas	Data loading and preprocessing
Humpy	Numerical computation
Matplotlib	Data visualization
Seaborn	Statistical visualization
Scikit-learn	Machine learning algorithms and evaluation
Imbalanced-learn (SMOTE)	Handling class imbalance
XGBoost	Gradient boosting classification
Joblib/Pickle	Model saving and deployment

4.3 Data Processing and Preprocessing

In order to improve the dataset quality for machine learning, we performed data cleaning, missing value imputation, duplicate removal and deletion of variables with high correlation and uninformative features, feature selection, categorical variable encoding feature normalization or scaling ,training set/testing set splits and balanced our datasets with SMOTE. The aggregation also increased the actual consistency of the info and lowered interpretive bias when it came to predicting superior performing people in high-milk garnering Sheko cattle. The models did a much better job with preprocessing steps.

4.3.1 Data Cleaning

We looked at the Teppi Agricultural Research Center's dataset to spot problems like duplicates, mistakes, and unnecessary info. Then, we cleaned the data to improve its quality and get it ready for machine learning analysis.

4.3.2 Handling Missing Values

Missing values pop up all the time in livestock datasets because of incomplete recording or plain errors. These issues can seriously hurt a model's accuracy if not fixed. To deal with this, we kept the data intact by preserving all records with complete info, rather than cutting those with missing bits. This way, we dodge info loss and improve our dataset's overall completeness.

4.3.3 Feature Selection

Features are useful for selecting the variables that really matter in predicting milk yield and, thus, improve model accuracy, reduce overfitting, reduce computation time and make the models understandable. We chose features based on our understanding of dairy science, a review by the FDRE, and key attributes associated with milk production. They were age, weight, body condition score, parity, lactation stage, feed type, health status, and daily milk yield. We made the target variable a binary variable where the most producing cows got a 1, while the least yielders got a 0. This strategy made our machine learning models far more accurate since they included the biggest factors that affect milk output.

4.3.3.1 Correlation based future selection

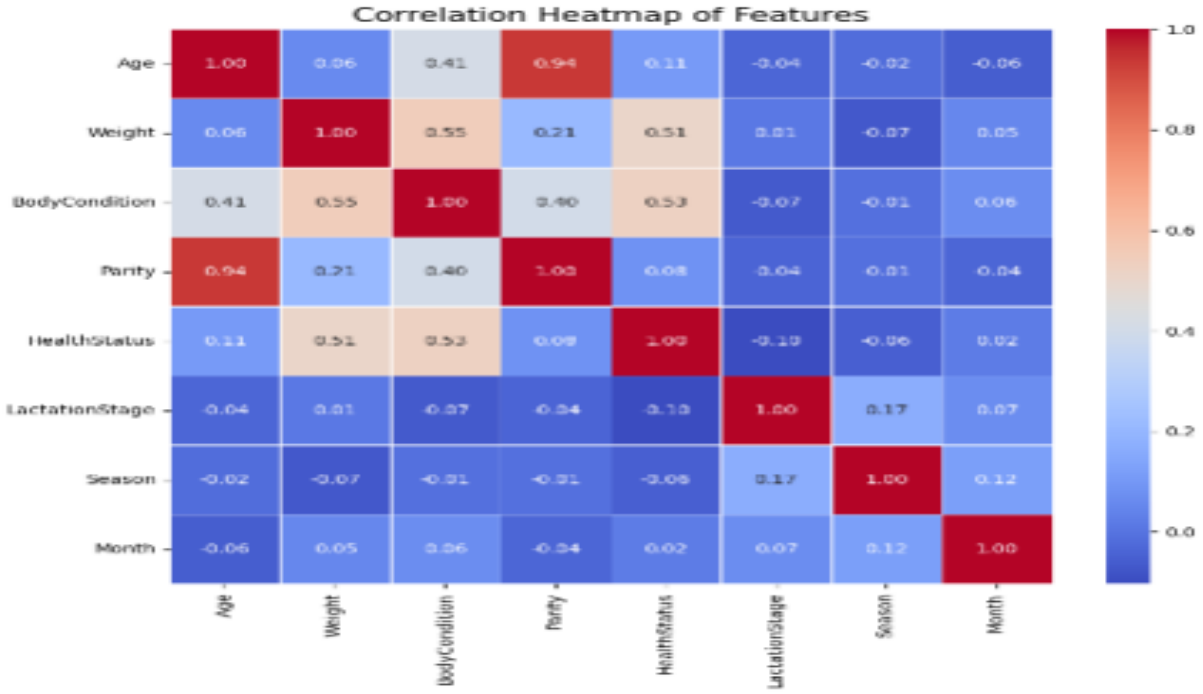


Figure 3 Correlation based future selection

There are a few important features to choose from so as to maximize the accuracy of machine learning models, avoid overfitting, reduce computational work, and make the models easier to understand. We selected features based on our knowledge of dairy science and the review done by the FDRE as well as factors related to milk production. We chose age, weight, body condition score, parity, lactation stage, feed type, health status, and daily milk yield. We made the target variable binary, i.e., the higher yielders get a 1 and the lower-producing cows got a 0. Ultimately, we improved the model by targeting the biggest factors affecting the output.

4.3.4 Encoding Categorical Variables

Machine learning algorithms require numerical input variables. Encoding transformed categorical variables into a machine-readable numerical format while preserving the information contained in the original variables.

4.3.5 Splitting the Dataset

The dataset was divided into training and testing subsets to evaluate model generalization capability.

- Training dataset: used for model learning
- Testing dataset: used for performance evaluation
- 80:20 ratio was used

Dataset splitting helped evaluate how well the trained models perform on unseen data and reduced overfitting risk.

4.3.6 Handling Class Imbalance Using SMOTE

The original dataset showed an imbalance between high milk-yielding and low milk-yielding cows. Imbalanced data can cause machine learning models to become biased toward the majority class.

To address this problem, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training dataset.

4.3.7 Class Distribution Before SMOTE

Yield category

1.0 6120

0.0 3418

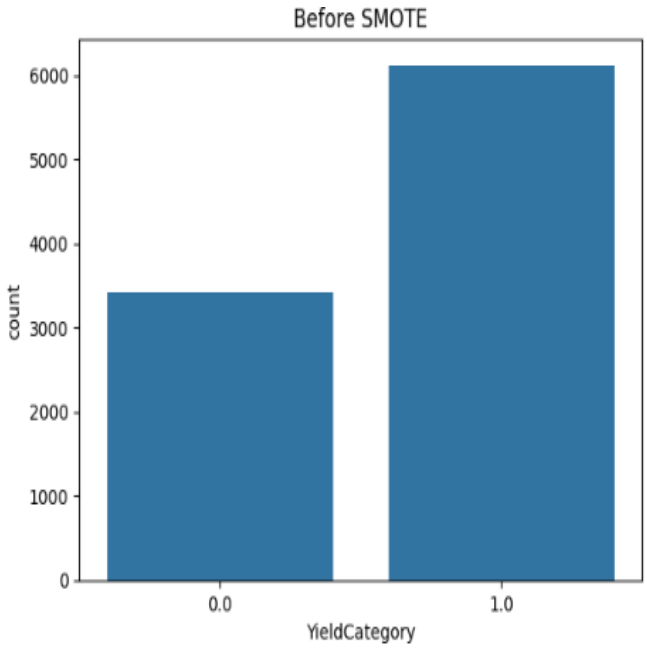


Figure 4 Class Distribution Before SMOTE

Before applying the Synthetic Minority Oversampling Technique (SMOTE), the dataset showed an imbalanced distribution between the two milk-yield categories. The majority class consisted of high milk-yielding cows, while the minority class represented low milk-yielding cows.

Table 5 Class Distribution Before SMOTE

Yield Category	Number of Instances
1.0 (High Milk Yield)	6120
0.0 (Low Milk Yield)	3418

This class imbalance could negatively impact machine learning model performance since models tend to favor the majority class during training. SMOTE was used to balance the data by creating synthetic samples for the minority class before model training.

Imbalance Ratio= $\frac{6120}{3418} \approx 1.79:1$

This indicates that the majority class was approximately 1.79 times larger than the minority class before applying SMOTE.

4.3.8. Class Distribution After SMOTE

Yield category

1.0 6120

0.0 6120

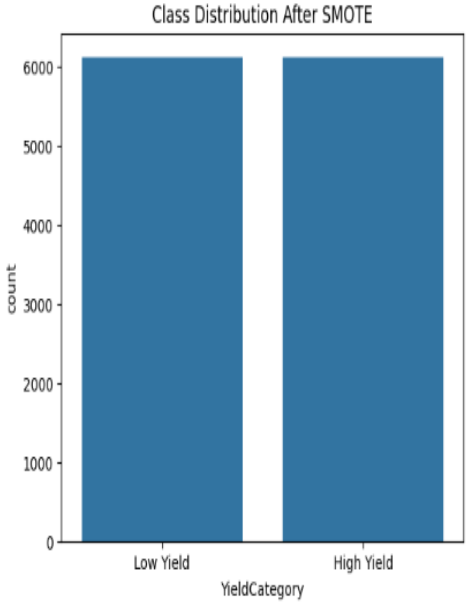


Figure 5 Class Distribution After SMOTE

The distribution of classes in the dataset was balanced after the application of the Synthetic Minority Oversampling Technique (SMOTE). Using SMOTE, synthetic samples were produced for the minority class (low-milk-yielding cows) until both classes had equal number of instances. The imbalance in the original dataset has been removed.

Table 6 Class Distribution After SMOTE

Yield Category	Number of Instances
1.0 (High Milk Yield)	6120
0.0 (Low Milk Yield)	6120

Balanced data enable machine learning algorithms to learn equally from both classes of data and reduce biased learning toward the majority class. This results in more accurate models that correctly classify both high and low milk producing cows. Balancing the dataset is especially important in classification problems as a poorly balanced dataset may lead to inaccurate classification and poor predictions of the minority class.

SMOTE creates synthetic examples of minority class using the nearest neighboring sample rather than duplicating existing records. The model generalization is better and the risk of overfitting is lower. After balancing the data set, the machine learning models were trained on the processed data to produce better, accurate predictions.

The class ratio after SMOTE = $\frac{6120}{6120}$ = 1:1

4.4 Model Training and Evaluation

After data preprocessing and weighting the dataset through the Synthetic Minority Oversampling Technique (SMOTE), the data set was used to train and evaluate machine learning models to detect high-milking Sheko cattle using the input features.

The dataset was split into training and testing subsets using the train-test split technique. In the current study, 80% of the dataset was used for training, while 20% was used for testing and validation. The training dataset was used to train the machine learning models, whereas the testing dataset was used to test the predictive power of the models on unviewed data. Several supervised machine learning classification algorithms were implemented and compared during the experimentation process. These included Random Forest, Decision Tree, AdaBoost, and XGBoost; the models were selected because of their effectiveness in solving binary classification problems and their ability to handle agricultural and livestock production datasets efficiently.

The classification of cows into low and high milk-yielding categories was based on the following threshold rule: Yield Category $\begin{cases} = & 0, \text{ if } \text{Milk Yield} < 3 \\ & 1, \text{ if } \text{M ilk Yield} >= 3 \end{cases}$

Where 0 indicates no milk-producing cows and 1 indicates no milk-producing cows. To improve model performance, we tuned hyperparameters using Grid Search Cross

Validation to identify the ideal settings for each machine learning algorithm. We also used K Fold Cross Validation to check model stability and generalization ability by training and validating the models on different data subsets.

Model performance was evaluated using accuracy, precision, recall, F1-score, ROC-AUC score, and confusion matrix analysis. Accuracy measures overall classification correctness, precision measures correctly predicted high milk-yielding cows, recall measures the ability to identify actual high milk-yielding cows, and F1-score provides a balance between precision and recall. ROC-AUC evaluates the discrimination capability of the classifier, while the confusion matrix visualizes correctly and incorrectly classified instances. The best-performing model was selected based on overall evaluation and cross-validation results and was used for developing the predictive prototype system for identifying high milk-yielding Sheko cattle cows.

4.4.1 Decision Tree Model

The Decision Tree algorithm was implemented to classify milk yield categories by creating tree-like decision rules based on input features. The model is simple, interpretable, and effective for nonlinear relationships.

4.4.2 Performance Evaluation Metrics for Decision Tree before and after SMOTE

Table 7 Performance Evaluation for Decision Tree before and After SMOTE

Model	Stage	Accuracy	Precision	Recall	F1-score	ROC-AUC
Decision Tree	Before SMOTE	0.9266	0.9195	0.9706	0.9444	0.9656
Decision Tree	After SMOTE	0.9046	0.9189	0.9714	0.9297	0.9657

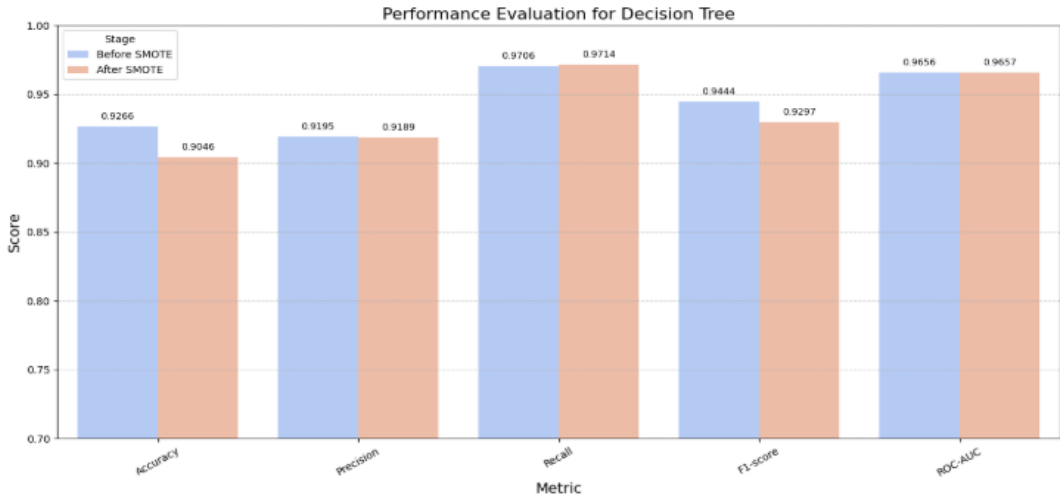


Figure 6 Performance Evaluation metrics for decision Tree before and after SMOTE

Before applying SMOTE, the Decision Tree model achieved an accuracy of 0.9266, with a precision of 0.9195, a recall of 0.9706, an F1-score of 0.9444, and a ROC-AUC of 0.9656. This indicates that the model performed well in identifying both classes, but the imbalance in the dataset could still influence learning toward the majority class.

After applying SMOTE, the performance remained stable with a slight change in precision (0.9189) and recall (0.9714), while accuracy and F1-score remained consistent at 0.90 and 0.9297 respectively. The ROC-AUC slightly improved to 0.9657, showing that SMOTE helped improve class balance without significantly changing overall performance, but improved model reliability in classification.

Hyperparameter tuning was applied to optimize the performance of the Decision Tree classifier and improve its generalization ability. Since Decision Tree performance depends on parameter settings, Grid Search Cross-Validation was used to identify the best combination of hyperparameters that maximize accuracy.

The tuned parameters include criterion (Gini impurity and Entropy), max depth (controls tree depth and prevents overfitting), min samples split (minimum samples required to split a node), and min samples leaf (minimum samples required at a leaf node).

A Grid Search with 5-fold cross-validation (k=5) was used to evaluate different parameter combinations. The model was trained on each combination, and the best parameters were selected based on the highest cross-validation accuracy.

Gini impurity after each split Model chooses split that maximizes class separation and accuracy. Used to build optimal tree structure for classification decisions.

K-fold cross-validation is used to evaluate the performance of the Decision Tree model by dividing the dataset into K equal subsets, where K is set to 5 in this study. The model is trained using K−1 folds and tested on the remaining fold, and this process is repeated five times so that each fold is used once as the testing set. The final model performance is calculated by taking the average of all evaluation results from the five folds. This method provides a more reliable estimate of model performance and helps to reduce overfitting during model evaluation

The Decision Tree model was trained using the processed and balanced dataset to classify Sheko cattle cows into high and low milk-yielding categories. After data preprocessing, feature selection, and handling class imbalance using SMOTE, the dataset was divided into training and testing sets. The training set was used to learn patterns and relationships between the input features (age, parity, body condition score, lactation stage, and weight) and the target variable (milk yield category).

The trained model learned decision rules that allow it to accurately predict whether a cow belongs to the high or low milk-yielding category based on its characteristics. After training, the model was evaluated using the testing dataset to measure its predictive performance.

The differences between “Before SMOTE” and “After SMOTE” are small but meaningful.

SMOTE does not drastically change the model’s overall performance but rebalances how the model behaves across classes.

The most important change is a shift toward better minority class detection (Recall improvement).

Based on the result SMOTE improves the model’s fairness and sensitivity toward the minority class, which is more important than accuracy in imbalanced livestock datasets.

4.4.4 Confusion metrics for Decision Tree

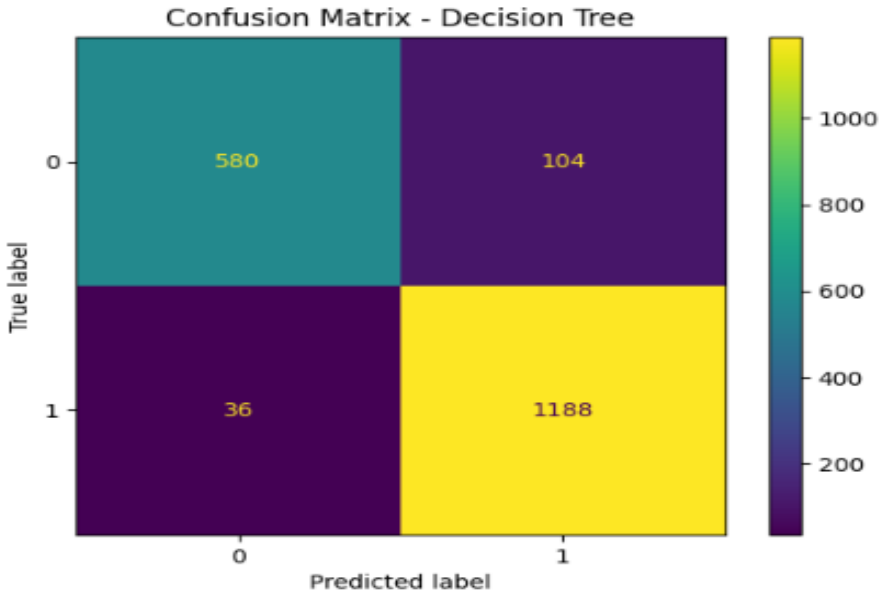


Figure 7 Confusion Metrics for Decision Tree

The confusion matrix evaluates the performance of the Decision Tree classification model in predicting low and high milk-yielding cows.

	Predicted Low (0)	Predicted High (1)
Actual Low (0)	580	104
Actual High (1)	36	1188

True Negative (TN) =580, False Positive (FP) =104, False Negative (FN) =36, and True Positive (TP) =1188

The confusion matrix shows that the Decision Tree model performed very well in classifying milk yield categories. The model produced a large number of correct predictions (TP and TN) and relatively few misclassifications (FP and FN).

Total correct predictions: TP + TN = 1188 + 580 = 1768

Total incorrect predictions: FP + FN = 104 + 36 = 140

The high number of True Positives and True Negatives demonstrates that the model effectively learned the relationship between input features and milk yield categories.

The low False Negative value particularly indicates that the model is highly effective at identifying high milk-yielding cows, which is important for dairy productivity improvement.

The confusion matrix result revealed that the Decision Tree model achieved strong classification performance in predicting milk yield categories of Sheko cattle cows. The model correctly classified 1188 high milk-yielding cows and 580 low milk-yielding cows. However, 104 low-yielding cows were incorrectly classified as high-yielding, while 36 high-yielding cows were misclassified as low-yielding. Overall, the model produced a high number of correct predictions with relatively few classification errors, demonstrating its effectiveness and reliability for identifying high milk-yielding cows.

4.4.3 AUC –ROC for Decision Tree

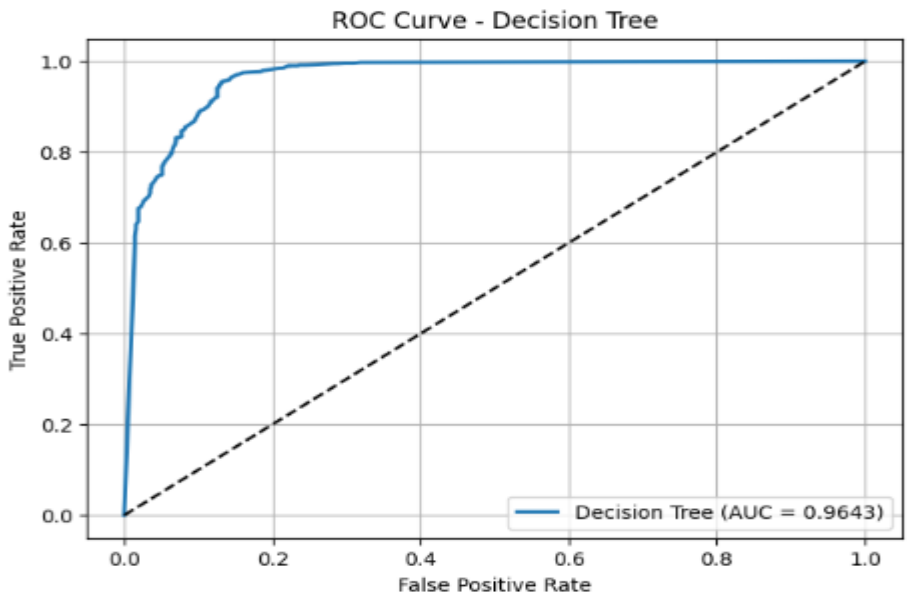


Figure 8 AUC –ROC for Decision Tree

The ROC curve of the Decision Tree model is located in the upper left corner of the graph and rises above the diagonal base for random classification. This indicates that the model is very sensitive to identification of positive cases and has a low false positive rate. The abrupt rise in the curve indicates that the model is highly sensitive to the distinction between high-yield cattle and low-yield cattle.

The model has an AUC-ROC of 0.9643 (96.43%), which is considered very good according to machine learning evaluations. This high AUC score implies that the

Decision Tree model is discriminative and is capable of reliably classifying yield categories of cattle. Therefore, this model is capable of making decisions related to predicting high-yield cattle.

4.5 Random Forest Model

Random Forest is an ensemble machine learning algorithm that combines multiple Decision Trees to improve classification performance and prediction accuracy. The algorithm reduces overfitting by aggregating predictions from several trees using majority voting. In this study, Random Forest was applied to predict high milk-yielding Sheko cattle cows based on physiological and management-related attributes. The model demonstrated strong predictive capability and improved generalization performance compared to individual Decision Tree models.

4.5.6 Model Evaluation for Random Forest

Table 8 Performance Evaluation Matrix for Random Forest before and After SMPTE

Model	Stage	Accuracy	Precision	Recall	F1-score	ROC-AUC
Random Forest	Before SMOTE	0.9266	0.9195	0.9706	0.9444	0.9659
Random Forest	After SMOTE	0.9099	0.9297	0.9297	0.9297	0.9668

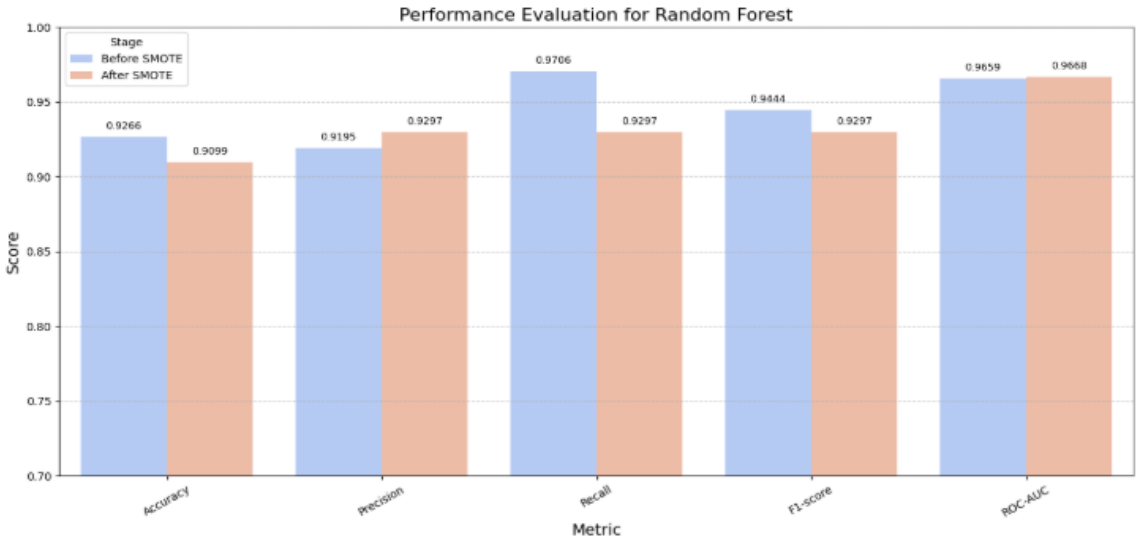


Figure 9 Model Performance Evaluation for Random Forest before and after SMOTE

The Random Forest model shows strong and stable performance both before and after applying SMOTE, with only slight variations in evaluation metrics. Before SMOTE, the model achieved higher overall accuracy (0.9266) and recall (0.9706), indicating strong ability to correctly identify the majority class and maintain high predictive performance. The F1-score (0.9444) and ROC-AUC (0.9659) also confirm excellent classification capability and strong class separation.

After applying SMOTE, the model accuracy slightly decreased to 0.9099, which is common due to balancing previously dominant classes. However, recall and precision became more balanced (both 0.9297), showing improved fairness in predicting both classes equally rather than favoring the majority class. The F1-score also stabilized at 0.9297, indicating a more balanced trade-off between precision and recall. Importantly, the ROC-AUC improved slightly to 0.9668, showing enhanced ability to distinguish between classes.

Before SMOTE the model was more accurate but slightly biased toward the majority class, while after SMOTE it became more balanced and fair in classification, with improved generalization across classes.

4.5.7 Confusion Metrics for Random Forest

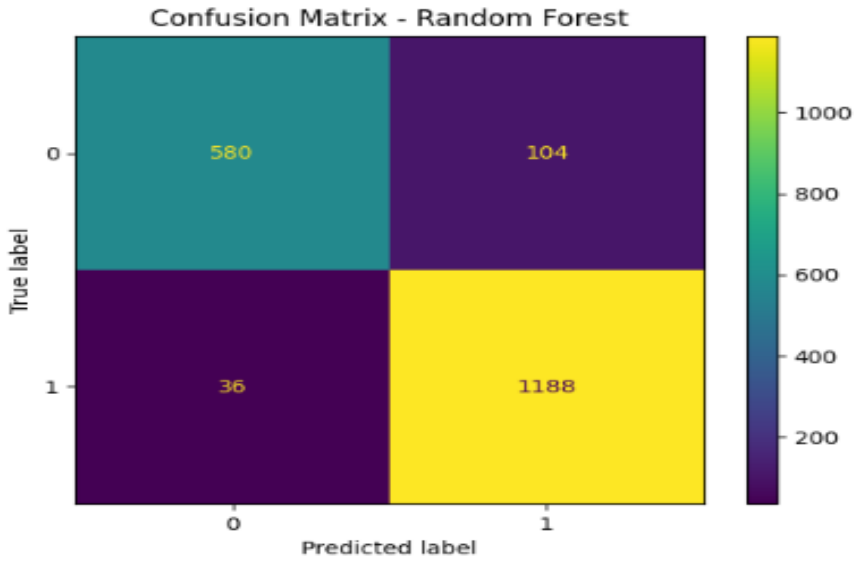


Figure 10 Confusion Metrics for Random Forest

The confusion matrix illustrates the classification performance of the Random Forest model in predicting milk yield categories. The matrix contains four important components: True Negatives (TN), False Positives (FP), False Negatives (FN), and True Positives (TP). In this model, class "0" represents low milk-yielding cows, while class "1" represents high milk-yielding cows.

The model correctly classified 580 low milk-yielding cows as class 0, representing the True Negatives (TN). It also correctly identified 1188 high milk-yielding cows as class 1, representing the True Positives (TP). These high correct classification values indicate that the Random Forest model performs very well in distinguishing between low and high milk yield categories.

However, the model incorrectly classified 104 low milk-yielding cows as high milk-yielding cows, which are False Positives (FP). In addition, 36 high milk-yielding cows were incorrectly predicted as low milk-yielding cows, representing False Negatives (FN). The lower number of false negatives compared to false positives indicates that the model is more effective at identifying high milk-yielding cows.

Overall, the confusion matrix demonstrates that the Random Forest model achieved strong classification performance with a high number of correct predictions and relatively few misclassifications. The model particularly showed excellent capability in detecting high milk-yielding cows, which is important for supporting dairy management, breeding selection, and production improvement decisions.

4.5.8 ROC Curve for Random Forest

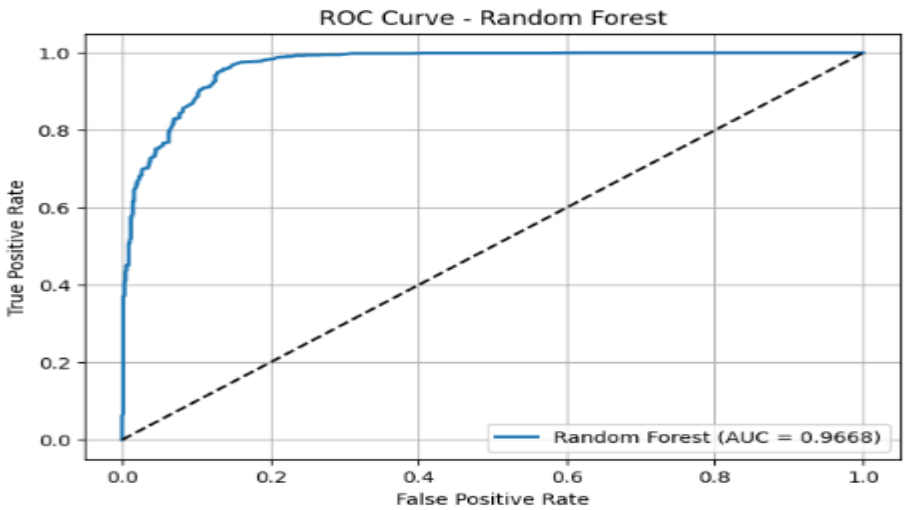


Figure 11 ROC Curve for Random Forest

The ROC curve of the Random Forest model is very close to the top-left corner of the graph and is still significantly above the diagonal reference line that represents the random classification line. It is seen that the model has a high True Positive Rate (TPR) and a low False Positive Rate (FPF) for various classification thresholds. The sharp initial increase of the curve shows the excellent discriminatory power and good classification ability of the model in the correct classification of high-yield cattle with few misclassifications.

This excellent level of predictive accuracy, measured by the AUC-ROC value of 0.9668 (96.68%) in machine learning classification tasks, was achieved by the model. AUC near the 1.0 value means that this model is good at differentiating high-yielding and low-yielding cows, at different decision thresholds. The prediction performance of the ensemble learning approach is superior to that of the Decision Tree model, XGBoost model and Ada Boost model (AUC = 0.9643), indicating that they have a slight improvement in discriminative ability. In this regard, the Random Forest model can be considered as a reliable and successful model to support the prediction of the yield of cattle and decisions making.

4.6. 1 Model Evaluation for XG Boost

Table 9performance metrics Model Evaluation for XG Boost before and after SMOTE

Model	Stage	Accuracy	Precision	Recall	F1-score	ROC-AUC
XGBoost	Before SMOTE	0.9266	0.9189	0.9714	0.9444	0.9673
XGBoost	After SMOTE	0.9078	0.9295	0.9265	0.9280	0.9660

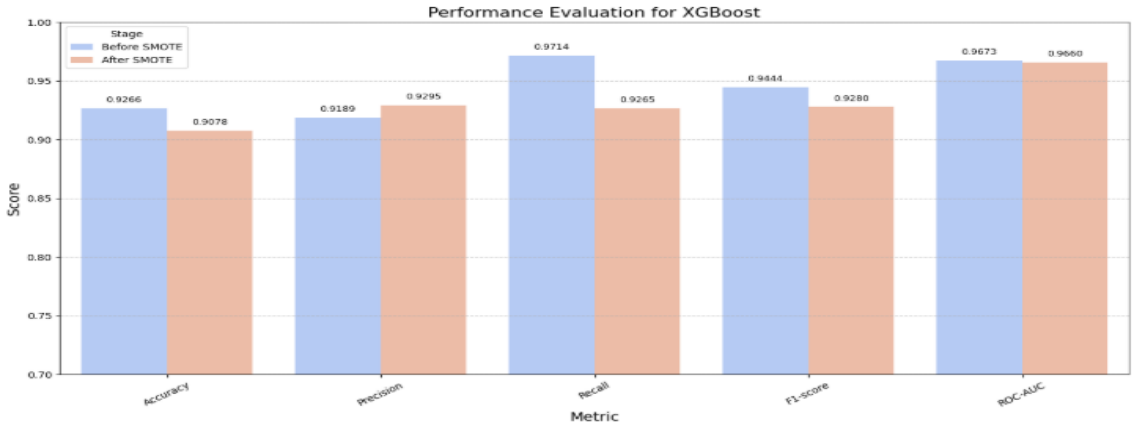


Figure 12performance metrics Model Evaluation for XG Boost before and after SMOTE

The performance of the XGBoost classification model was evaluated before and after applying the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance in the dataset. The results demonstrate noticeable changes in evaluation metrics, highlighting the impact of data balancing on model behavior.

Before applying SMOTE, the XGBoost model achieved an accuracy of 0.9266**,** precision of 0.9189**,** recall of 0.9714**,** F1-score of 0.9444**,** and ROC-AUC of 0.9673**.** These results indicate strong overall performance, particularly in terms of recall, which suggests the model was highly effective in identifying positive classes. However, this high recall may also indicate a bias toward the majority class, which is a common issue in imbalanced datasets.

After applying SMOTE, the model performance changed slightly. The accuracy decreased to 0.9078**,** while precision increased to 0.9295**,** recall decreased to 0.9265**,** F1-score slightly decreased to 0.9280, and ROC-AUC remained stable at 0.9660**.** The improvement in precision indicates that the model became more reliable in predicting positive classes, reducing the number of false positive predictions. However, the reduction in recall suggests that the model missed a slightly higher proportion of actual positive cases compared to the model trained on the original imbalanced dataset.

The F1-score, which represents the balance between precision and recall, showed a minor decline after SMOTE, reflecting the trade-off between these two metrics. Despite these variations, the ROC-AUC value remained consistently high in both cases, indicating that the model retained strong discriminatory ability between classes regardless of data balancing.

The application of SMOTE improved the balance of the dataset and enhanced the model's predictive reliability, particularly in terms of precision. However, this improvement came at the cost of a slight reduction in recall and overall accuracy. The results suggest that while SMOTE contributes to fairer class representation, it introduces a trade-off between sensitivity and precision in the XGBoost model.

4.6.9 Confusion Metrics for XG Boost

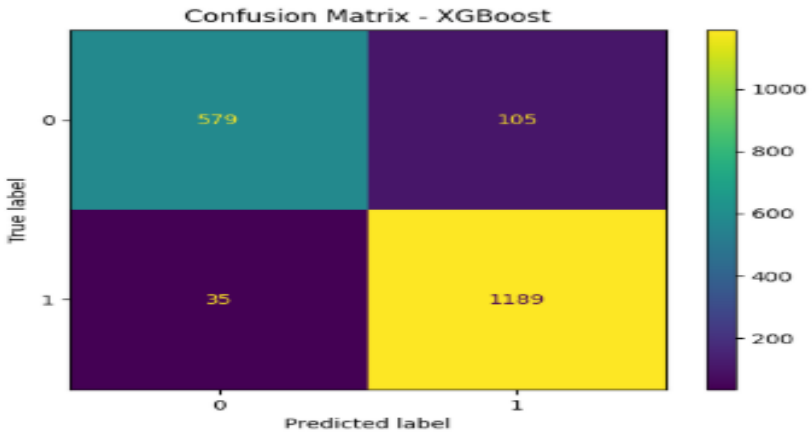


Figure 13 Confusion Metrics for XG Boost

The confusion matrix shows that the XGBoost model performs strongly in classifying both classes.

True Negatives (TN): 579 correct predictions of class 0**,** True Positives (TP): 1189 correct predictions of class 1**,** False Positives (FP): 105 classes 0 incorrectly predicted as class 1**,** and False Negatives (FN): 35 classes 1 incorrectly predicted as class 0

The model correctly classifies the majority high number of True Positive and True Negatives. The low number of false negatives (35) indicates strong ability to detect positive cases, while the moderate false positives (105) show a small level of misclassification in predicting class 1. Overall, the model demonstrates high classification performance with strong sensitivity toward class 1 predictions**.**

4.6.10 ROC CURVE for XG boost

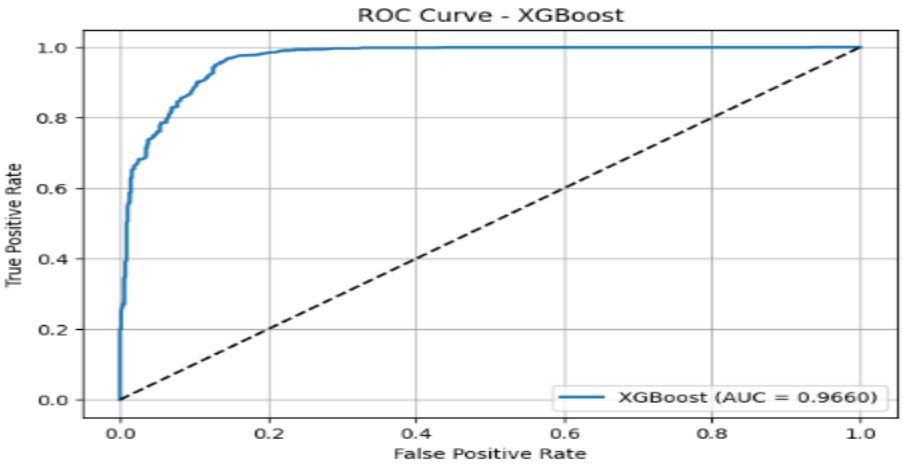


Figure 14 ROC CURVE for XG Boost

The ROC curve of the XGBoost model shows excellent classification performance in predicting milk yield. The curve is close to the upper-left corner of the graph, which means that the model has a high True Positive Rate (sensitivity) with a low False Positive Rate. The blue ROC curve is well above the diagonal reference line, which is the performance of a random classifier. The result indicates that the XGBoost model has a good performance in distinguishing the different milk yield classes and correctly classifies the positive cases without making too many classification errors.

The value of the Area under the Curve (AUC) is 0.9660, which further reinforces the excellent predictive power of the model. A value of 1.0 is very close to having excellent discrimination power; a value of 0.5 would mean that random prediction would have been used. The calculated AUC of 0.9660 thus implies that the XGBoost model has the capability of correctly ranking and classifying milk yield outcomes with about 96.6% accuracy in separating the classes. Based on this high AUC value, the model is robust, reliable, and very effective in milk yield classification and could be used as a good machine learning model to assist in decision making and productivity improvement in milk production systems of Sheko cattle.

4.7.4 Model Evaluation for Ada Boost

Table 10 Model Performance Evaluations for Ada Boost before and after SMOTE

Model	Stage	Accuracy	Precision	Recall	F1- Score	ROC-AUC
Ada Boost	Before SMOTE	0.8160	0.8129	0.9265	0.8660	0.8625
Ada Boost	After SMOTE	0.7710	0.8877	0.7361	0.8048	0.8693

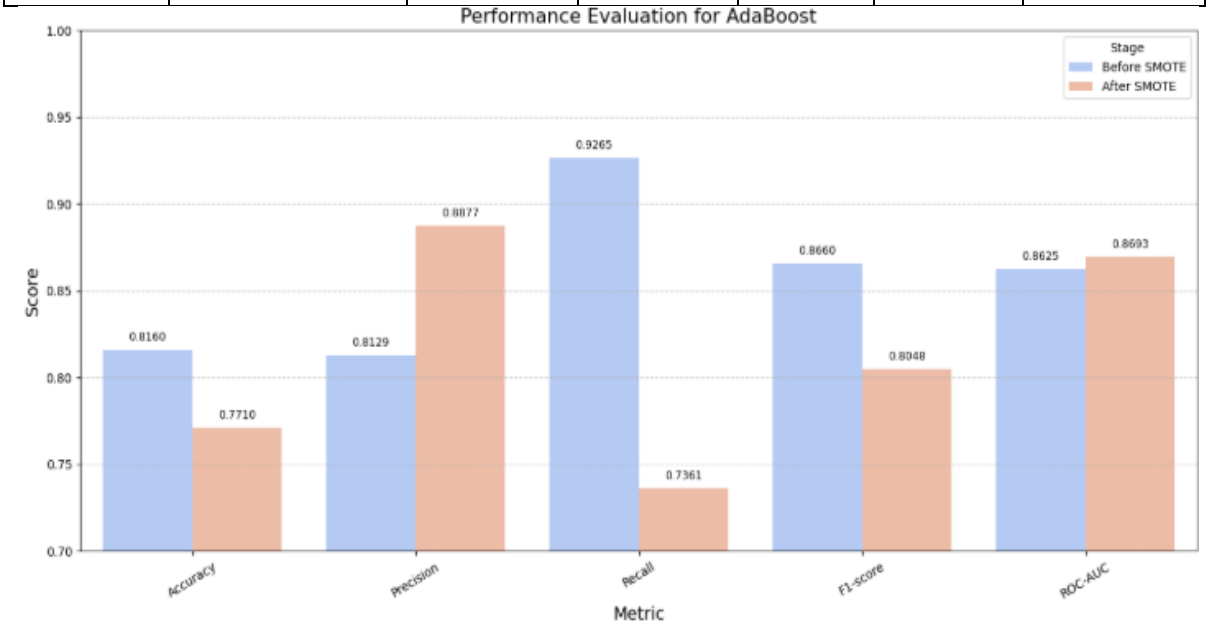


Figure 15 Model Performance Evaluations for Ada Boost before and after SMOTE

Performance Metrics for AdaBoost, before and after SMOTE presents a comparative evaluation of the AdaBoost classification model before and after applying the Synthetic Minority Oversampling Technique (SMOTE)**.** The graph compares evaluation metrics: Accuracy, Precision, Recall, F1-score, and AUC ROC**.** The blue bars represent the model performance before SMOTE, while the green bars represent the performance after balancing the dataset using SMOTE. This comparison helps evaluate how class balancing affects the effectiveness of the AdaBoost model in identifying high-yield cows.

The results show that the Accuracy slightly decreased after applying SMOTE, dropping from approximately 81% to 77%. This indicates that the overall proportion of correctly classified instances became slightly lower after balancing the dataset. However, accuracy alone is not sufficient for evaluating imbalanced classification problems because it may favor the majority class. In contrast, Precision improved significantly from around 82% to 89%, showing that the AdaBoost model became better at correctly identifying positive instances with fewer false positives after SMOTE. This improvement demonstrates that balancing the dataset enhanced the reliability of positive predictions.

The Recall value decreased slightly from approximately 77% to 73%, meaning the model identified slightly fewer actual positive cases after SMOTE. Although this reduction suggests a minor decline in sensitivity, it is relatively small and may result from the model becoming more conservative in classification decisions. The F1-score, which balances Precision and Recall, improved slightly from around 79% to 80%, indicating a better overall balance between identifying positive cases and minimizing incorrect predictions. This suggests that the AdaBoost model achieved a more stable and balanced classification performance after oversampling.

The ROC-AUC score improved marginally from approximately 86% to 87% after applying SMOTE. ROC-AUC measures the model's ability to distinguish between classes across different threshold values. The improvement indicates that the AdaBoost classifier achieved slightly better discriminative capability after balancing the minority class. A ROC-AUC value close to 1 represents excellent classification performance, and values above 0.85 generally indicate a strong predictive model.

The graph demonstrates that applying SMOTE improved several important classification metrics, particularly Precision, F1-score, and ROC-AUC, even though Accuracy and Recall showed slight decreases. This indicates that SMOTE helped AdaBoost better handle class imbalance and produce more reliable predictions for the minority class. Therefore, the post-SMOTE AdaBoost model can be considered more balanced and effective for classification tasks such as identifying high-yield indigenous Sheko-breed cows in Ethiopia.

4.7.5 Confusion Matrix for Ada Boost

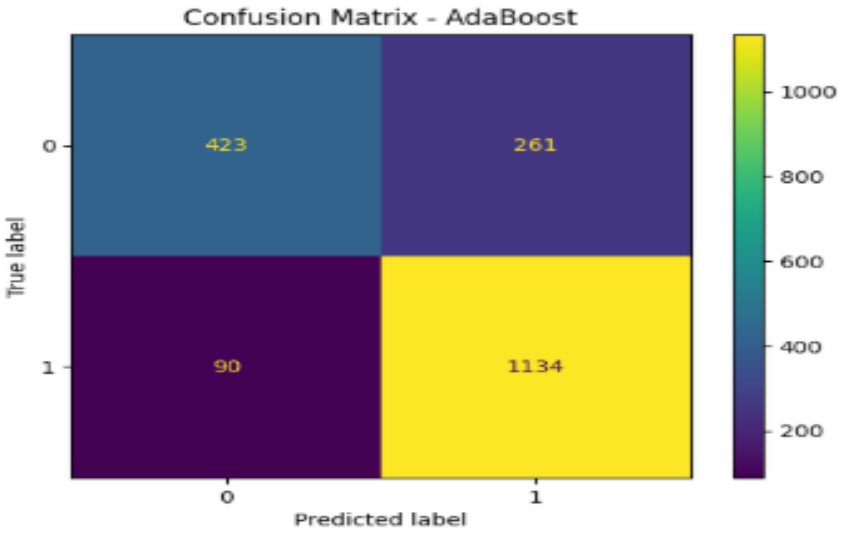


Figure 16 Confusion Matrix for Ada Boost

The confusion matrix result shows the performance of the developed classification model in predicting the target classes. The model correctly predicted 423 instances as class 0, which are referred to as True Negatives (TN). This indicates that the model was able to accurately identify negative cases without misclassification. Similarly, the model correctly predicted 1134 instances as class 1, known as True Positives (TP), showing that the model successfully identified a large number of positive cases. The high number of true positive predictions demonstrates that the model has strong capability in detecting the target class accurately.

However, some prediction errors were also observed. The model produced 261 False Positives (FP), meaning that 261 instances belonging to class 0 were incorrectly classified as class 1. This indicates that the model sometimes overestimates positive cases. In addition, the model generated 90 False Negatives (FN), where class 1 instances were incorrectly predicted as class 0. False negatives are particularly important because they represent actual positive cases missed by the model. Comparing the error values, false negatives is lower than false positives, indicating that the model is more effective at detecting positive cases than negative cases.

The confusion matrix demonstrates that the model achieved strong classification performance because the number of correct predictions (TN and TP) is significantly higher than the number of incorrect predictions (FP and FN). The high true positive value and relatively low false negative value suggest that the model is reliable and effective for

prediction tasks. These results confirm that the developed machine learning model has good predictive capability and can support accurate decision-making in the study

4.7.6 ROC Curve for Ada Boost

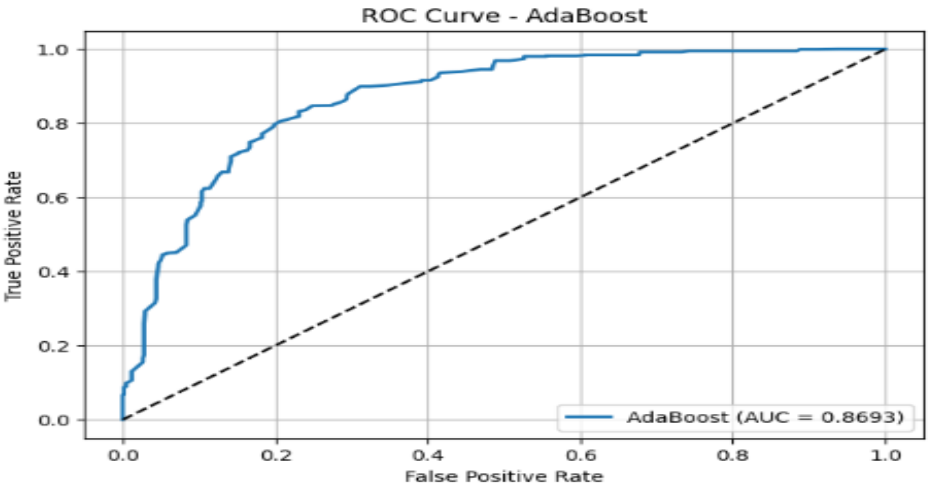


Figure 17 Roc Curve for Ada Boost

The ROC curve of the AdaBoost model shows the model's performance on differentiating the milk yield classes according to the classification thresholds. The curve is much above the dashed diagonal line, which corresponds to the performance of a random classifier, meaning that the model has a good discriminatory ability. As the FPR goes up, the TPR goes up quickly, indicating that the model is able to classify a high percentage of positive cases correctly with a relatively low false classification rate. Overall shape of the curve provides clear evidence that AdaBoost is much better than random guessing and is efficient in the classification of milk yield results.

The Area Under the Curve (AUC) value of 0.8693 also indicates that the AdaBoost model has a good prediction performance. The AUC value of 0.8693 means that the model has about 86.93% chance of correctly separating various milk yield classes. Despite this model having a good classification capability, its ROC curve is not as close to the upper-left corner as other models like Random Forest, Deision Tree, and XGBoost, which has a better discrimination power. Despite this, the AUC value achieved in this study falls within the acceptable range of 0.80 indicating that the AdaBoost model is reliable and robust enough to be used for milk yield classification for Sheko cattle production system and can be used for prediction and decision making process.

4.7. 7 Relationship between Lactation Stage, Body Condition Score, Body Weight, and Milk Yield Category Using Correlation Analysis

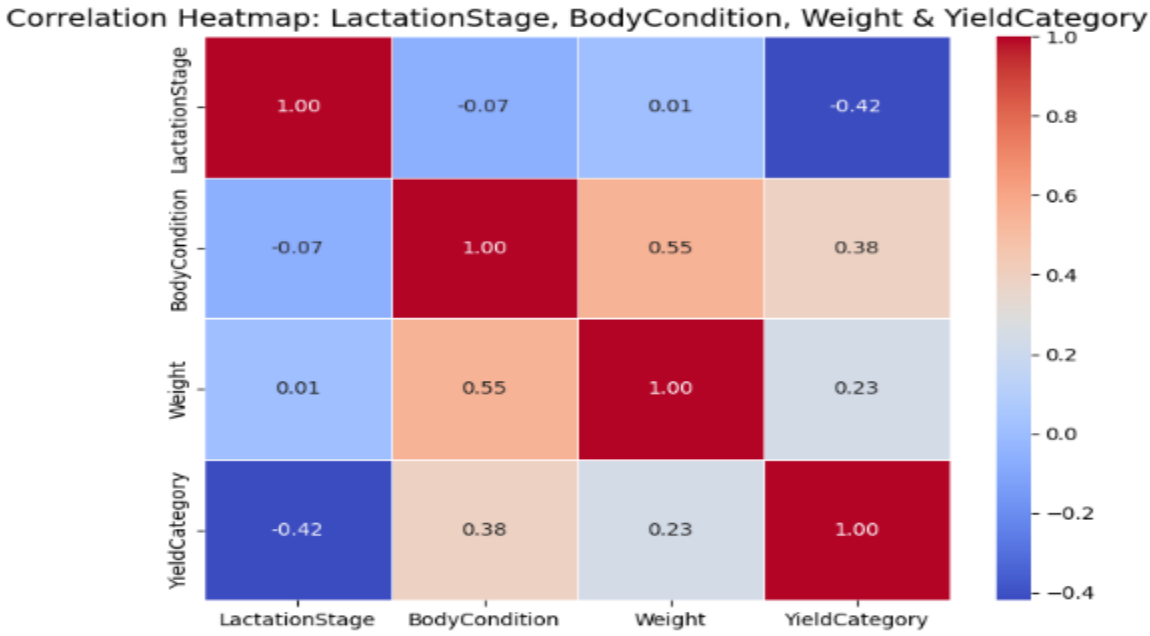


Figure 18 Correlation heat map for lactation stage, body condition, weight, and yield category

This heatmap shows the linear relationships among four key variables: Lactation Stage, Body Condition, Weight, and Milk Yield Category. The correlation analysis of indigenous cow lactation stage, body condition, body weight, and milk yield shows several significant correlations. A moderate positive correlation exists between body condition and weight (correlation coefficient of +0.55). This makes biological sense since the body condition score reflects the animal's fat and energy reserves, so ideally cows with high body condition scores also have high body weights. Lactation stage and yield also, this indicates moderate negative correlation of -0.42 so, that the lactation stages increase, the yield category generally decreases. This means that cows in the early stages of lactation typically have higher milk yields than do cows in the advanced stages of lactation. Weak or negligible associations exist between the variables lactation stage and body condition. The correlation coefficient (-0.07) calculated for these two variables indicates their relationship is almost completely non-linear (or very weakly linear). The relationship between lactation stage and weight may be established as there is no substantive correlation between these two variables (correlation coefficient = 0.01) so it is reasonable to assume that body weight does not depend upon the lactation stage from which a cow is producing milk.

When compared to the target variable (yield category), there is a moderate to weak positive correlation between body condition and yield category (correlation coefficient = 0.38). Therefore, cows with better body conditions are likely to be placed into higher categories of milk yield, making body condition a reasonable predictor of milk production. A similar positive correlation occurs between weight and yield category (correlation coefficient = 0.23) where heavier cows produce more milk than lighter ones to a lesser extent than the correlation previously described between body condition and yield category.

The degree of separation between lactation stage, body condition, and body weight in relation to milk yield also provides insight into the relative importance of these three features (physiological versus physical) when considering indigenous cow production. A moderate negative correlation exists between lactation stage and yield category (r = -0.42), which indicates that lactation stage is the most influential variable in regard to correlation strength within the three features mentioned earlier. Both lactation stage and body condition have moderate effects on the yield category of milk produced while body weight has a weaker effect than body condition. The findings of this study suggest that physiological factors (i.e., lactation stage and body condition) carry greater weight in determining total milk production than does physical size. Although there is a weak-moderate correlation between lactation stage, body condition, and yield category of milk production, the data indicate that linear paradigms alone cannot adequately model the complexity of predicting milk production from lactation stage, body condition, or body weight therefore supporting the rationale to utilize advanced machine learning models (e.g., Random Forests and XGBoost) in order to generate more accurate classifications.

4.7.8 Feature Importance Comparison of Lactation Stage, Body Condition, and Weight across Multiple Machine Learning Models

Table 11 Feature Importance Comparison of Lactation Stage, Body Condition, and Weight across Multiple Machine Learning Models

	Lactation Stage	Body Condition	Weight
Decision tree	0.2682	0.1542	0.1431
Random Forest	0.2714	0.0732	0.1844
XG Boost	0.0551	0.7894	0.0162
Ada Boost	0.1349	0.0243	2558

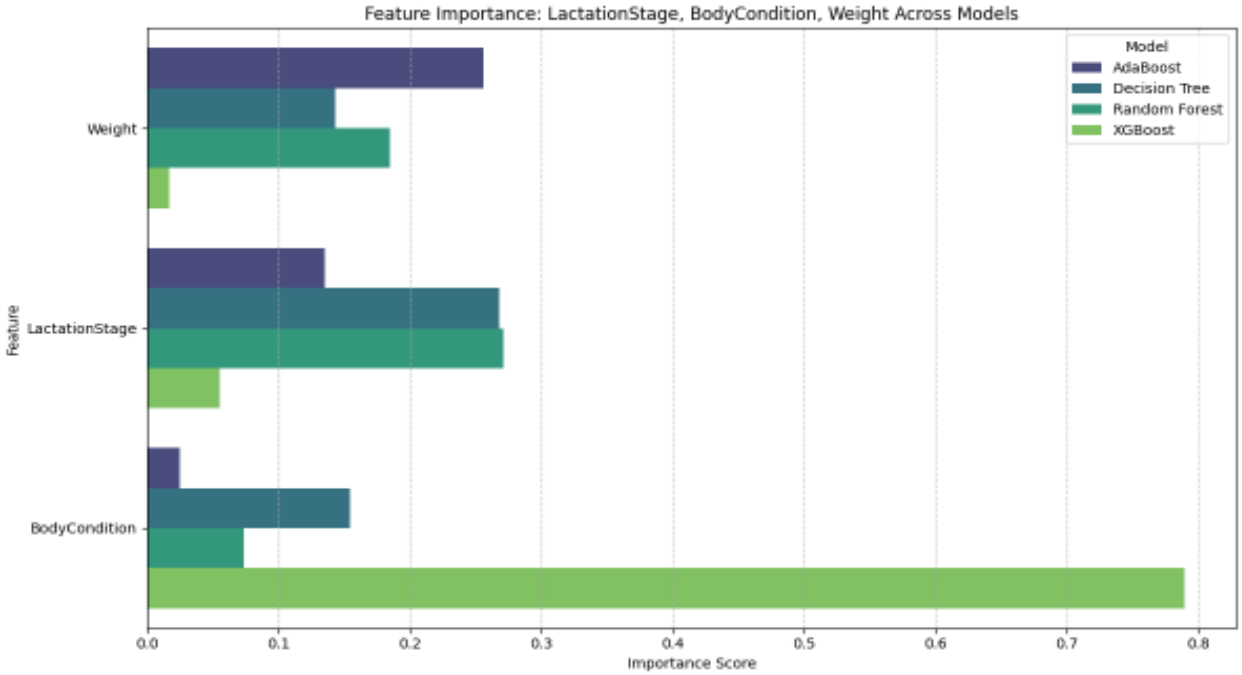


Figure 19 Feature Importance Comparison of Lactation Stage, Body Condition, and Weight across Multiple Machine Learning Models

The way in which AdaBoost, Decision Tree, Random Forest, and XGBoost rank the three predictors of Body Condition, Lactation Stage and Weight when predicting milk yield category is the subject of this graph. A significant takeaway from the analysis of this graph is that the importance of a given feature will vary depending on the model being used. As such, there are several ways in which each algorithm interprets the same data set. Based on the feature importance analysis from the three models, we see that XGBoost gives a very high body condition feature importance score of roughly .80. This indicates that body condition is the most important feature (predictor) for milk yield classification using the XGBoost boosted model. Because body condition is by far the highest predictor through XGBoost, it plays the most significant role compared to all other variables in the XGBoost model. Lactation stage has different levels of importance across the different models examined. In the Decision Tree and Random Forest, lactation stage had a very large importance score of approximately 0.27 to 0.28, meaning that it contributed to predicting milk yield fairly well; in AdaBoost, however, this importance was approximately 0.15 to 0.18, so about half as much influence. XGBoost features lactation stage's low importance score of approximately 0.1 or less. Although lactation stage is consistently part of the model predictions, it loses its importance through various levels of ensemble, specifically XGBoost; as lactation stage may have some overlapping effects

with the other features, particularly with body condition. The significance of weight varies among different modeling options as well. It has a fairly modest level of importance with a score of approximately 0.20 (moderate) to 0.25 with Decision Trees and AdaBoost Modeling; Random Forest modeling also scores at a somewhat similar level (0.15-0.20). Weight scores as having a very low level of predictive importance in XGBoost Modeling with an importance score of approximately 0.02 to 0.05; therefore, it can be inferred that weight has little or no direct effect on the prediction being made. In conclusion, weight does not appear to be a strong independent predictor among the performance of any of the sextuple modeling options; its contribution (to prediction) may be captured through body condition (indirectly).

4.7. 9 Model-Wise Interpretation feature importance for lactation stage, body condition and weight

According to the results of the AdaBoost model, there appears to be a more uniform distribution across all features, meaning that no variable dominates the prediction process. The lactation stage, along with the weight of the animal, plays a moderate role in predicting the category of milk yield in the AdaBoost model; however, this model employs a relatively simple learning method through reduced dominance of features and a more even spread of weights among predictor variables. On the other hand, the Decision Tree model has a fairly uniform distribution of feature importance, with lactation as the dominating feature when compared to the rest of the features; thus, lactation is likely to have a stronger effect when splitting decisions for the purpose of the tree structure. By contrast, Random Forest models produce a nearly identical representation of feature importance; however, the Random Forest correctly represents the importance of lactation stage without being overly strong as a single weighted value. This is due to the ensemble averaging process, which spreads the importance value throughout the other features.

XGBoost was found to be the most informative model; however, it was primarily reliant upon importance values for the condition of the animals as it had an average importance score near 0.80. In addition, the importance of weight in this model was given little consideration as opposed to the importance of the lactation stage in other models. Thus, it appears that XGBoost is very good at producing good feature selection and optimization because it is able to classify milk yield .This also indicates that the boosting algorithm of XGBoost has effectively modeled the complex relationships within the dataset and has provided the most accurate characterization of the best predictor for classifying milk

yield. An analysis of feature importance on the body condition as the strongest predictor of milk yield class showed that body condition is the best predictor by far under XGBoost modeling; however, lactation stage provides a moderate level of predictability; and weight has the least amount of influence on predicting milk yield. In addition, the variation/configuration of these data sets amongst the models demonstrates the benefit of an ensemble learning approach to determining different views of feature importance. Finally, the results indicate that physiological health (body reserves) is more important than physical size in determining the productivity of indigenous cows.

4.7.10 Evaluation Metrics Comparison for All Models

Table 12 Performance Evaluation Metrics Comparison for All Models

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Decision Tree	0.9046	0.9292	0.9216	0.9253	0.9643
Random Forest	0.9099	0.9297	0.9297	0.9297	0.9668
XG Boost	0.9078	0.9295	0.9265	0.9280	0.9660
Ada Boost	0.7710	0.8877	0.7361	0.8048	0.8693

Based on an assessment of SMOTE-processed optimized ML model performance, results showed that nearly every ensemble-based model except AdaBoost scored reasonably well on classification ability. The models were evaluated using Accuracy, Precision, Recall, F1 -Score & ROC-AUC to determine which model provided the best predictions of cows with high milk-yield.

The Decision Tree Model had an Accuracy Rate of 90.46%, Precision Rate of 92.92%, Recall Rate of 92.16%, and F1-Score of 92.53% and ROC-AUC of 96.43%. This performance indicates that the model was successful in accurately classifying both positive and negative classes. The high Recall Rate and F1-Score demonstrate that the model was strong in identifying high-yielding cows and produced balanced performance across its various classifications.

The Random Forest model outperformed all other classification methods, producing accurate predictions. It achieved high levels of accuracy (90.99%), precision (92.97%), recall (92.97%), and F1 score (92.97%), along with an exceptional ROC-AUC value

(96.68%). The fact that all four measures had equal values (precision, recall, F1 score) indicates that Random Forest was able to make balanced predictions without significant bias. Moreover, the Random Forest model achieved the highest ROC-AUC value among the classifiers tested (96.68%) demonstrating that Random Forest was the most capable classifier for distinguishing between the two classes being studied. These results indicate that the Random Forest classifier provided consistency, accuracy, and dependability in its classification performance following optimization and the use of SMOTE balancing. Although the XGBoost model performed slightly lower than the Random Forest model overall, the results obtained from the XGBoost model remain very similar with regards to the differences between both models; however, the key takeaway here is that XGBoost is still an extremely robust machine learning approach to predicting milk yield. The results suggest that the XGBoost model has excellent discriminatory ability and balanced classification accuracy. Overall, the results from the XGBoost model demonstrated similar levels of quality assessment as would normally be associated with Random Forest; therefore, it would be accurate to state that both models can be used interchangeably when predicting durably. By comparison, AdaBoost had lower performance figures indicating its lower effectiveness and reliability for the prediction task given the following performance measures: accuracy = 77.10%, precision = 88.77%, recall = 73.61%, F1 = 80.48% and ROC-AUC = 86.93%; indicating that it did not correctly classify a substantial number of positive classifications (as indicated by the low recall value), and that it has moderate discriminatory ability compared to the other optimized models (as indicated by ROC -AUC value below 90%). The results of the comparison indicate that the classification performance improved considerably due to the implementation of SMOTE and hyperparameter tuning using ensemble learning methods. The highest accuracy, F1 score, recall, and ROC-AUC values were observed for Random Forest and hence Random Forest was chosen as the most optimal final model. Random Forest was the model with the most balanced and stable performance demonstrating the best ability to generalize and predict whether dairy cows will have been high yielding milking cows. As a result, Random Forest is deemed the best model to use to assist with decision making and management of dairy production in this research.

4.7.10.1 ROC Curve Comparison for all Model

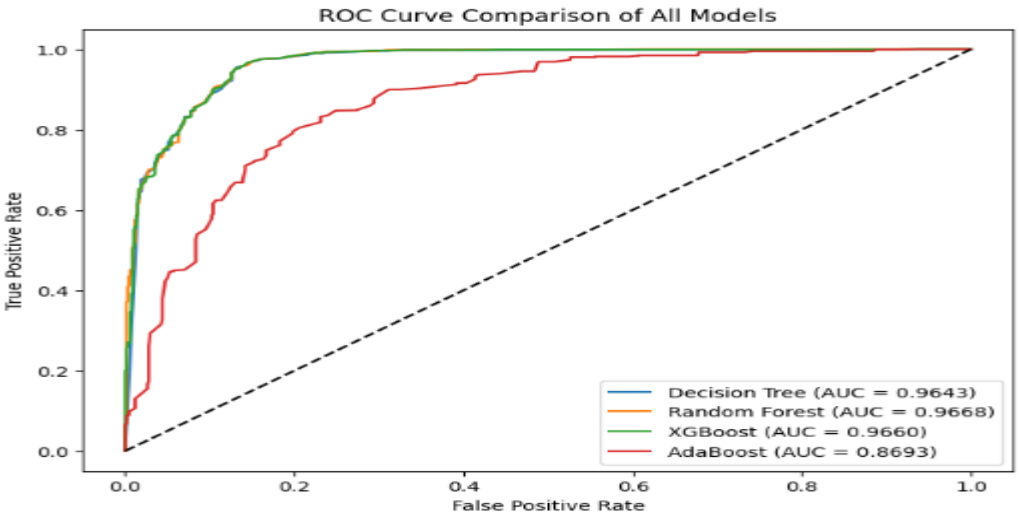


Figure 20 ROC Curve Comparison for all Models

The ROC comparison shows that all four machine learning models had a significant improvement over the random classification; all of them have curves that are above the diagonal reference line. Random Forest obtained the highest AUC value (0.9668) with XGBoost and Decision Tree coming very close behind (0.9660 and 0.9643 respectively). The ROC curves of these three models are grouped around the top left corner of the graph, which shows a good discrimination ability, high sensitivity and low false-positive rates. They showed almost the same curves indicating that all three models work well in distinguishing milk yield classes.

AdaBoost, on the other hand, had the lowest AUC value (0.8693), and the corresponding ROC curve is further from the upper left corner than the other models. AdaBoost still has good classification performance and is well above the random classifier line, but its power to separate milk yield classes is less compared to the other classifiers. The slight difference in terms of the predictive performance between the models, Random Forest, XGBoost and Decision Tree, suggests that these models perform in a similar manner with Random Forest performing slightly better. The overall best model with the highest accuracy was the Random Forest with the highest AUC score followed by XGBoost and Decision Tree and then the AdaBoost model, Random Forest was the most reliable model for milk yield classification in this study.

4.7.10.2 Mean Absolute SHAP across Optimized Model

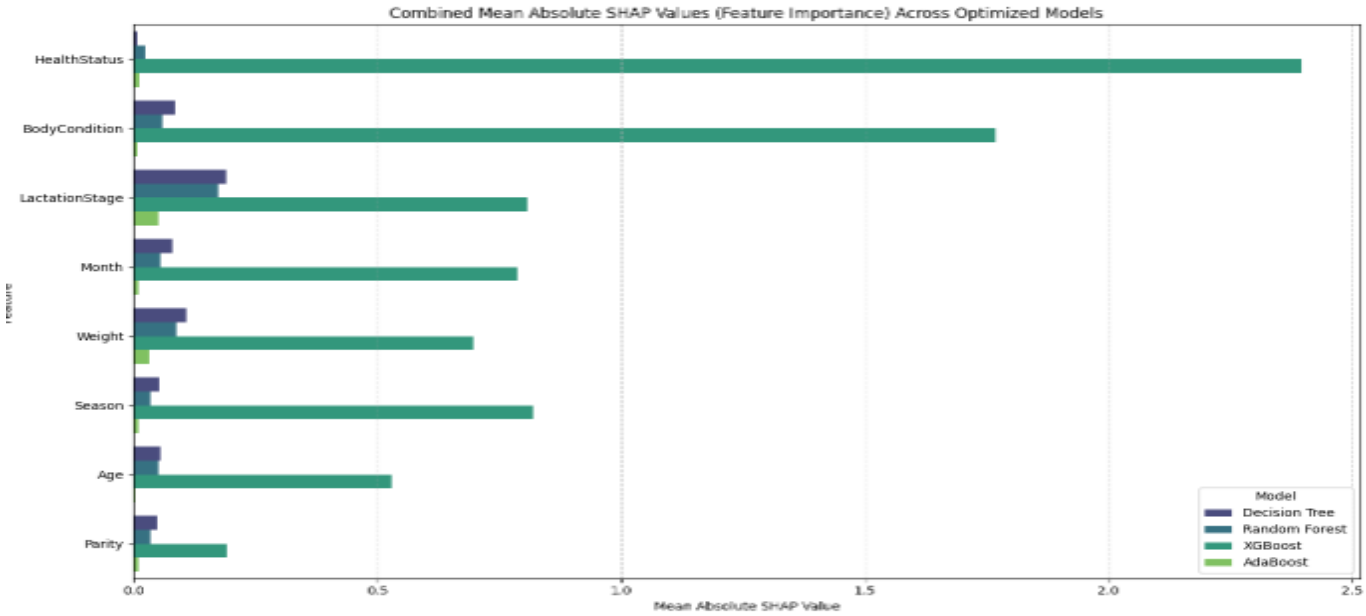


Figure 21 Mean Absolute SHAP across Optimized Model

The Mean Absolute SHAP (SHAP) values for the four models (Decision Tree, Random Forest, XGBoost and AdaBoost) are combined to demonstrate the relative importance of each of the features in the classification of milk yield. Higher SHAP values correspond to higher relative importance of the feature in the milk yield prediction. The most important predictor, and with the highest SHAP value (~2.4, most important in the XGBoost) is Health Status, which means that animal health is the most significant indicator of milk yield, with poor health indicating high risk of low milk yield. Body Condition is the second most important feature (~1.7-1.9), and indicates the stock of body energy reserves, and if cows are well conditioned, they will produce more milk; and if they are thin, cows are more likely to have low yield. The lactation stage has moderate to high importance (~0.8–0.9), with early lactation being related to high yield and late lactation to low yield. The other variables are moderate in importance, with Month and Season having values of ~0.75–0.85, suggesting that the amount of feed available and climatic conditions have a significant effect on milk yield, particularly during dry periods. Weight is a medium important factor (~0.6–0.7), that is heavier cows tend to have higher production capacity. The productivity curve shows that age contributes at lower to moderate importance (~0.45–0.55), whereas the remaining feature in the table Parity is found to be least important with ~0.1–0.2 contribution, which means it is not a strong predictor compared to the other features considered as related to health and nutrition.

The SHAP analysis shows that there are three levels of risk factors which are essential for the prediction of milk yield. Health Status and Body Condition are the most important factors affecting milk yield outcomes and are considered to be high-risk. Medium-risk drivers include the Lactation Stage, Season/Month and Weight which moderately influence production under physiological and environmental circumstances. The factors that appear to have relatively little effect on the prediction of milk yield are low risk factors such as Age and Parity.

The health status is the most important determinant of milk yield classification; Body Condition Score is the second most important predictor. Production is significantly influenced by environmental factors: season and month. Age and Parity have little effect in comparison to other physiological and management related variables. Random Forest has the highest classification accuracy, and XGBoost has the best feature interpretability

4.8 Discussion and Results

Author/year	Breed	Dataset used	Model used	Performance	Key finding	Gap
Abdelrahman et al. (2022)	Holstein dairy cows	91 cows (100–550	Random Forest	≈82% Accuracy	Random Forest effectively predicted milk yield under environmental stress.	Small dataset and only Holstein cattle.
González-Recio et al. (2020)	Dairy cattle	69,000 lactation records	Random Forest, Gradient Boosting	RF AUC ≈0.90	Historical lactation records improved prediction.	Focused on commercia l dairy breeds only.
Pzczola et al. (2021)	Dairy cows	370 cows	Random Forest, Logistic Regression	AUC ≈88–90%	Random Forest captured nonlinear production patterns.	Predicted resilience rather than milk-yield class.
Negussie et al. (2021)	Horro cattle, Ethiopia	220 cows, 1,908 records	GLM (SAS)	R^2 = 0.75–0.88	Non-genetic factors significantly influenced milk yield.	No machine learning models were applied.

As mentioned in section 1.4, we carefully formulated research questions to address the problem and, for now, have found answers to those research questions in the following paragraphs.

Q1. What are the major factors affecting the milk yield performance of Sheko Cattle Cows in Sheka Zone, Ethiopia?

According to SHAP analysis (Figure 20), the dominant factor influencing milk yield is Health Status, followed by Body Condition Score (2nd dominant factor), then Lactation Stage, Season/Month, Weight, Age and Parity.

Q2. How can machine learning techniques be used to predict high milk-yielding Sheko cattle cows?

Using preprocessing (Clearing data, Encoding data, Feature selection, Class balancing with SMOTE), splitting data into 80-20 ratio, training classifiers (Decision Tree, Random Forest, XGBoost, AdaBoost) with cross-validation and evaluating with accuracy, precision, recall, F1-score and ROC-AUC.

Q3. What is the effect of applying SMOTE on the performance of machine learning models for milk yield prediction?

SMOTE changes the ratio of the dataset from 1.79:1 to 1:1, reducing majority class bias and boosting precision (e.g., AdaBoost is now 81.29% more precise than before and 88.77%) and ROC-AUC, but it reduces overall accuracy but improves minority class detection and model fairness.

4. Which machine learning algorithm provides the best predictive performance for identifying high milk-yielding Sheko cattle cows?

Random Forest scored highest in accuracy (90.99%), the highest ROC-AUC (96.68%), and the highest balance in precision, recall, and F1-score (92.97% each), so it is the best, stable, and reliable model for this classification task.

Chapter five

5. Conclusion and Recommendation

5.1 Conclusion

The aim of this study was to overcome the problem of subjective selection and farmer experience in the Sheko Zone, Ethiopia, where a farmer's main selection criteria for indigenous Sheko cattle is the subjective observation of the herd. For this purpose, a machine learning-based predictive framework was designed with the livestock production data collected at the Teppi agricultural research center. The dataset was preprocessed extensively by implementing data cleaning, feature selection, categorical encoding, train-test split and class balancing via SMOTE. Four supervised machine learning algorithms were developed, trained, and tested with the commonly used performance metrics: Decision Tree, Random Forest, XGBoost, and AdaBoost.

The experimental findings revealed that machine learning methods could benefit in the classification of milk yield of indigenous cattle population. The highest predictive performance was obtained by the model Random Forest (accuracy = 90.99%, f1-score = 92.97% and ROC-AUC = 96.68%) among the models evaluated. Lactation stage, body condition score and body weight were the most important variables influencing milk yield, according to feature importance and SHAP explainability analysis. The results of the research addressed all the research questions, and it was concluded that data driven approach will be helpful for identifying high milk yielding cows, for the making of informed breeding decisions, and for improving the productivity of the dairy industry. Hence, the study brings a significant contribution toward the development of precision livestock farming in Ethiopia by developing an accurate, interpretable and reliable machine learning model for boosting the productivity and management of indigenous Sheko cattle.

5.2 Recommendation

The newly created livestock prediction model should be validated using larger and more varied livestock datasets as well as considering additional factors like genetics, nutrition, and environmental conditions to improve prediction accuracy and the appropriateness of the prediction model. Farmers and extension workers should utilize data-driven approaches when making breeding and herd management decisions. Greater attention

should be given to key factors affecting milk production, such as lactation stage, body weight, body condition, and animal health, to enhance dairy productivity. Government and livestock development agencies should promote the integration of artificial intelligence and machine learning technologies into livestock development programs. Support should also be provided for the establishment of digital livestock databases and decision-support systems to improve dairy production and breeding efficiency.

Future studies should evaluate advanced Machine learning and deep learning techniques, use datasets from different regions and cattle breeds, and develop practical mobile or web-based applications for milk yield prediction and livestock management.

References

*, *Lynne A. Carpenter-Boggs*. (2019). 1–50.

Ababa, A. (2014). *TRAINING PACKAGE ON LIVESTOCK HUSBANDRY UNDER SEMI-INTENSIVE MANAGEMENT*. *December*.

Abegaz, S. B. (2022). Milk production status and associated factors among indigenous dairy cows in Raya Kobo district, north eastern Ethiopia. *Veterinary Medicine and Science*, *8*(2), 852–863. https://doi.org/10.1002/vms3.740

Aleme, M., & Mengistu, G. (2023). *Management practice of the Sheko cattle breed in Ethiopia : A review*. *4*, 64–70. https://doi.org/10.46265/genresj.IBNU2035

Asresie, A., & Seifu, L. Z. (2024). *Contribution of Livestock Sector in Ethiopian Economy : A Review Contribution of Livestock Sector in Ethiopian Economy : A Review*. *February*, 78–91.

Bahbahani, H., Afana, A., & Wragg, D. (2018). *Genomic signatures of adaptive introgression and environmental adaptation in the Sheko cattle of southwest Ethiopia*. 1–14. https://doi.org/10.1126/science.1167936.PubMed

Bayou, E., Haile, A., Gizaw, S., & Mekasha, Y. (2015). Evaluation of non - genetic factors affecting calf growth , reproductive performance and milk yield of traditionally managed Sheko cattle in southwest Ethiopia. *SpringerPlus*. https://doi.org/10.1186/s40064-015-1340-9

Begna, R., Asfaw, Y., & Masho, W. (2023). Heliyon Evaluation of trait preferences and effect of parity , season and lactation stage on production performance of indigenous dairy cow in kaffa zone , southwest Ethiopia. *Heliyon*, *9*(12), e22380. https://doi.org/10.1016/j.heliyon.2023.e22380

Bovo, M., Agrusti, M., Benni, S., Torreggiani, D., & Tassinari, P. (2021). Random forest modelling of milk yield of dairy cows under heat stress conditions. *Animals*, *11*(5). https://doi.org/10.3390/ani11051305

Cao, Y., Dai, J., Zhang, G., & Xia, M. (2024). *Combinations of Feature Selection and Machine Learning Models for Object-Oriented " Staple-Crop-Shifting " Monitoring Based on Gaofen-6 Imagery*.

Chen, T. (n.d.). *XGBoost : A Scalable Tree Boosting System*.

Chen, T., & Guestrin, C. (2016). *XGBoost : A Scalable Tree Boosting System*. 785–794.

Code, L. G., Code, T., Apr, A. G. R., & Apr, A. G. R. (2021). *Animal production Level III Based on March 2018 , Version 3 Occupational standard Activities LG Code : TTLM Code : 0*(March 2018).

Ermias, N., Mesfin, B., Chaka, D., Gezahegn, A., Bikila, S., & Teshome, W. (2024). Transforming agriculture with Machine Learning , Deep Learning , and IoT : perspectives from Ethiopia — challenges and opportunities. In *Discover Agriculture*. Springer International Publishing. https://doi.org/10.1007/s44279-024-00066-7

Friedman, B. J., Hastie, T., & Tibshirani, R. (2000). *Special invited paper*. *28*(2), 337–407.

Frizzarin, M., Visentin, G., Ferragina, A., Hayes, E., Dhariyal, B., Domijan, K., Khan, H., Ifrim, G., Nguyen, T. Le, Meagher, J., Menchetti, L., & Singh, A. (2022). *Classification of cow diet based on milk Mid Infrared Spectra : a data analysis competition at the " International Workshop of Spectroscopy and Chemometrics 2022 ."* 1–27.

Gebreyohanes, G., Yilma, Z., Moyo, S. and, & A., M. O. (2021). Dairy industry development in Ethiopia : Current status , major challenges and potential interventions for improvement. *ILRI POSITION PAPER Dairy*, 1–39.

Giannuzzi, D., Flavio, L., Mota, M., Pegolo, S., Gallo, L., Schiavon, S., Tagliapietra, F., Katz, G., Fainboym, D., & Minuti, A. (2022). In - line near - infrared analysis of milk coupled with machine learning methods for the daily prediction of blood metabolic profile in dairy cattle. *Scientific Reports*, 1–13. https://doi.org/10.1038/s41598-022-11799-0

Gocheva-ilieva, S., & Yordanova, A. (2022). *Predicting the 305-Day Milk Yield of Holstein-Friesian Cows Depending on the Conformation Traits and Farm Using Simplified Selective Ensembles*.

Gudeto, A., Mirkena, T., & Kebede, T. (2021). *Traditional cattle breeding practices and rate of inbreeding in the mid rift valley of Oromia , Ethiopia*. *7*, 77–82.

Guevara-escobar, A., Lemus-ramírez, V., & García-muñiz, J. G. (2026). *Efficient Prediction of Milk Yield with Machine Learning Models Using Cow Identification or Milk Quality Traits*. 1–26.

Guo, R., & Dai, Y. (2025). *Leveraging IGOOSE-XGBoost for the Early Detection of Subclinical Mastitis in Dairy Cows*. 1–18.

Hasan, M. F., Celik, N., Williams, Y., Williams, S. R. O., & Marett, L. C. (2025). An effective way to incorporate temperature – humidity index to study effect of heat stress on milk yield by an XGBoost machine learning model. *Journal of Dairy Science*, *108*(12), 13995–14017. https://doi.org/10.3168/jds.2025-27193

Hatew, B., Peñagaricano, F., Balehegn, M., Jones, C. S., Dahl, G. E., & Adesogan, A. T. (2023). Synergies of feed, management trainings, and genetics on milk production of dairy cows in the tropics: The case of Ethiopian smallholder farmers. *Frontiers in Animal Science*, *4*(March), 1–12. https://doi.org/10.3389/fanim.2023.1119786

health statys.pdf. (n.d.).

Hooker, J., Medeiros, B. B. De, Saha, C., Abdulrahman, T., & Alves, A. A. C. (2025). Evaluating the contribution of behavioral , milking system , and environmental data to short-term milk yield prediction in commercial dairy cows using machine learning. *Journal of Dairy Science*, *108*(10), 11133–11150. https://doi.org/10.3168/jds.2025-26724

Id, M. B., Yigrem, S., Betsha, S., & Tolera, A. (2021). *Production , reproduction and some adaptation characteristics of Boran cattle breed under changing climate : A systematic review and meta-analysis*. 1–26. https://doi.org/10.1371/journal.pone.0244836

Introduction.pdf. (n.d.).

Jabed, A., Azrifah, M., & Murad, A. (2024). Heliyon Crop yield prediction in agriculture : A comprehensive review of machine learning and deep learning approaches , with insights for future research and sustainability. *Heliyon*, *10*(24), e40836. https://doi.org/10.1016/j.heliyon.2024.e40836

Korir, L., Manning, L., Moore, H. L., Lindahl, J. F., Gemechu, G., Mihret, A., Berg, S., Wood, J. L. N., & Nyokabi, N. S. (2023). Adoption of dairy technologies in smallholder dairy farms in Ethiopia. *Frontiers in Sustainable Food Systems*, *7*(May). https://doi.org/10.3389/fsufs.2023.1070349

Lalman, D., & Stein, D. (2024). *Body Condition Scoring of Cows*. 1–4.

Magana, J., Gavojdian, D., Menahem, Y., Lazebnik, T., Zamansky, A., & Adams-Progar,

A. (2023). Machine learning approaches to predict and detect early-onset of digital dermatitis in dairy cows using sensor data. *Frontiers in Veterinary Science*, *10*(November), 1–9. https://doi.org/10.3389/fvets.2023.1295430

Magdalena, J., & Sablik, P. (2025). *The Use of Selected Machine Learning Methods in Dairy Cattle Farming : A Review*. 1–38.

Mayberry, D., Jemberu, W., Schrobback, P., Herrero, M., Chaters, G., Knight-jones, T., & Rushton, J. (2023). *Characterizing Ethiopian cattle production systems for disease burden analysis*. *September*, 1–9. https://doi.org/10.3389/fvets.2023.1233474

Mekonnen, A., Haile, A., Dessie, T., & Mekasha, Y. (2021). *Original Research*. *10*(3), 1–13.

Mengesha, E., Didanna, H. L., & Ayza, A. (2025). Assessing strategy to improve dairy production and milk quality under smallholder dairying: Evidence from Ethiopia. *Acta Scientiarum - Animal Sciences*, *47*(1). https://doi.org/10.4025/actascianimsci.v47i1.70290

Moawed, S. A., Mahrous, E., Elaswad, A., Gouda, H. F., & Fathy, A. (2024). *Milk yield prediction in Friesian cows using linear and flexible discriminant analysis under assumptions violations*. 1–12.

No Title. (n.d.). *August 2025*.

Ouweltjes, W., & Haas, Y. De. (2021). A data-driven prediction of lifetime resilience of dairy cows using commercial sensor data collected during first lactation. *Journal of Dairy Science*, *104*(11), 11759–11769. https://doi.org/10.3168/jds.2021-20413

Salamone, M., Adriaens, I., Vervaet, A., Opsomer, G., Atashi, H., Fievez, V., Aernouts, B., & Hostens, M. (2022). Animal The international journal of animal biosciences Prediction of first test day milk yield using historical records in dairy cows. *Animal*, *16*(11), 100658. https://doi.org/10.1016/j.animal.2022.100658

Shekuru, A. H., Berlie, A. B., & Bizuneh, Y. K. (2022). Variability and trends of temperature and rainfall over three agro-ecological zones in North Shewa, Central Ethiopia. *Arabian Journal of Geosciences*, *15*(18). https://doi.org/10.1007/s12517-022-10674-6

State, T. H. E. (n.d.). *THE STATE OF THE WORLD ' s ANIMAL GENETIC*

RESOURCES FOR FOOD AND AGRICULTURE.

Tarihi, K., & Kelimeler, A. (2025). *Makine Öğrenmesinin Süt ve Süt Ürünlerinde Kullanımına Yönelik Meta -Analitik Değerlendirme*. *9*(4), 1223–1232.

Techniques, M. L. (2025). *International Journal of Computer Techniques – Volume 12 Issue 3 , May - June - 2025 Predictive Modeling of Milk Production Using Artificial Intelligence and International Journal of Computer Techniques – Volume 12 Issue 3 , May - June - 2025*. *12*(3), 1–7.

Topuz, D. (2025). *Hybrid Ensemble Model for Lactation Milk Yield Prediction of Holstein Cows*. *31*(5), 603–611. https://doi.org/10.9775/kvfd.2025.34031

Wang, T. (2024). Improved random forest classification model combined with C5 . 0 algorithm for vegetation feature analysis in non - agricultural environments. *Scientific Reports*, 1–13. https://doi.org/10.1038/s41598-024-60066-x

Weldegerima, T. M. (2018). *Characterization of productive and reproductive performances , morphometric and challenges and opportunities of indigenous cattle breeds of Ethiopia : A review*. *9*(March), 29–41. https://doi.org/10.5897/IJLP2017.0426

Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M. (2017). Big Data in Smart Farming – A review. *Agricultural Systems*, *153*, 69–80. https://doi.org/10.1016/j.agsy.2017.01.023

Zealand, N., & Tour, S. (n.d.). *Transforming Ethiopia ' s Livestock Sector*.

*, *Lynne A. Carpenter-Boggs*. (2019). 1–50.

Ababa, A. (2014). *TRAINING PACKAGE ON LIVESTOCK HUSBANDRY UNDER SEMI-INTENSIVE MANAGEMENT*. *December*.

Abegaz, S. B. (2022). Milk production status and associated factors among indigenous dairy cows in Raya Kobo district, north eastern Ethiopia. *Veterinary Medicine and Science*, *8*(2), 852–863. https://doi.org/10.1002/vms3.740

Aleme, M., & Mengistu, G. (2023). *Management practice of the Sheko cattle breed in Ethiopia : A review*. *4*, 64–70. https://doi.org/10.46265/genresj.IBNU2035

Asresie, A., & Seifu, L. Z. (2024). *Contribution of Livestock Sector in Ethiopian Economy : A Review Contribution of Livestock Sector in Ethiopian Economy : A*

Review. *February*, 78–91.

Bahbahani, H., Afana, A., & Wragg, D. (2018). *Genomic signatures of adaptive introgression and environmental adaptation in the Sheko cattle of southwest Ethiopia*. 1–14. https://doi.org/10.1126/science.1167936.PubMed

Bayou, E., Haile, A., Gizaw, S., & Mekasha, Y. (2015). Evaluation of non - genetic factors affecting calf growth , reproductive performance and milk yield of traditionally managed Sheko cattle in southwest Ethiopia. *SpringerPlus*. https://doi.org/10.1186/s40064-015-1340-9

Begna, R., Asfaw, Y., & Masho, W. (2023). Heliyon Evaluation of trait preferences and effect of parity , season and lactation stage on production performance of indigenous dairy cow in kaffa zone , southwest Ethiopia. *Heliyon*, *9*(12), e22380. https://doi.org/10.1016/j.heliyon.2023.e22380

Bovo, M., Agrusti, M., Benni, S., Torreggiani, D., & Tassinari, P. (2021). Random forest modelling of milk yield of dairy cows under heat stress conditions. *Animals*, *11*(5). https://doi.org/10.3390/ani11051305

Cao, Y., Dai, J., Zhang, G., & Xia, M. (2024). *Combinations of Feature Selection and Machine Learning Models for Object-Oriented " Staple-Crop-Shifting " Monitoring Based on Gaofen-6 Imagery*.

Chen, T. (n.d.). *XGBoost : A Scalable Tree Boosting System*.

Chen, T., & Guestrin, C. (2016). *XGBoost : A Scalable Tree Boosting System*. 785–794.

Code, L. G., Code, T., Apr, A. G. R., & Apr, A. G. R. (2021). *Animal production Level III Based on March 2018 , Version 3 Occupational standard Activities LG Code : TTLM Code : 0*(March 2018).

Ermias, N., Mesfin, B., Chaka, D., Gezahegn, A., Bikila, S., & Teshome, W. (2024). Transforming agriculture with Machine Learning , Deep Learning , and IoT : perspectives from Ethiopia — challenges and opportunities. In *Discover Agriculture*. Springer International Publishing. https://doi.org/10.1007/s44279-024-00066-7

Friedman, B. J., Hastie, T., & Tibshirani, R. (2000). *Special invited paper*. *28*(2), 337–407.

Frizzarin, M., Visentin, G., Ferragina, A., Hayes, E., Dhariyal, B., Domijan, K., Khan, H.,

Ifrim, G., Nguyen, T. Le, Meagher, J., Menchetti, L., & Singh, A. (2022). *Classification of cow diet based on milk Mid Infrared Spectra : a data analysis competition at the " International Workshop of Spectroscopy and Chemometrics 2022 ."* 1–27.

Gebreyohanes, G., Yilma, Z., Moyo, S. and, & A., M. O. (2021). Dairy industry development in Ethiopia : Current status , major challenges and potential interventions for improvement. *ILRI POSITION PAPER Dairy*, 1–39.

Giannuzzi, D., Flavio, L., Mota, M., Pegolo, S., Gallo, L., Schiavon, S., Tagliapietra, F., Katz, G., Fainboym, D., & Minuti, A. (2022). In - line near - infrared analysis of milk coupled with machine learning methods for the daily prediction of blood metabolic profile in dairy cattle. *Scientific Reports*, 1–13. https://doi.org/10.1038/s41598-022-11799-0

Gocheva-ilieva, S., & Yordanova, A. (2022). *Predicting the 305-Day Milk Yield of Holstein-Friesian Cows Depending on the Conformation Traits and Farm Using Simplified Selective Ensembles*.

Gudeto, A., Mirkena, T., & Kebede, T. (2021). *Traditional cattle breeding practices and rate of inbreeding in the mid rift valley of Oromia , Ethiopia*. *7*, 77–82.

Guevara-escobar, A., Lemus-ramírez, V., & García-muñiz, J. G. (2026). *Efficient Prediction of Milk Yield with Machine Learning Models Using Cow Identification or Milk Quality Traits*. 1–26.

Guo, R., & Dai, Y. (2025). *Leveraging IGOOSE-XGBoost for the Early Detection of Subclinical Mastitis in Dairy Cows*. 1–18.

Hasan, M. F., Celik, N., Williams, Y., Williams, S. R. O., & Marett, L. C. (2025). An effective way to incorporate temperature – humidity index to study effect of heat stress on milk yield by an XGBoost machine learning model. *Journal of Dairy Science*, *108*(12), 13995–14017. https://doi.org/10.3168/jds.2025-27193

Hatew, B., Peñagaricano, F., Balehegn, M., Jones, C. S., Dahl, G. E., & Adesogan, A. T. (2023). Synergies of feed, management trainings, and genetics on milk production of dairy cows in the tropics: The case of Ethiopian smallholder farmers. *Frontiers in Animal Science*, *4*(March), 1–12. https://doi.org/10.3389/fanim.2023.1119786

health statys.pdf. (n.d.).

Hooker, J., Medeiros, B. B. De, Saha, C., Abdulrahman, T., & Alves, A. A. C. (2025). Evaluating the contribution of behavioral , milking system , and environmental data to short-term milk yield prediction in commercial dairy cows using machine learning. *Journal of Dairy Science*, *108*(10), 11133–11150. https://doi.org/10.3168/jds.2025-26724

Id, M. B., Yigrem, S., Betsha, S., & Tolera, A. (2021). *Production , reproduction and some adaptation characteristics of Boran cattle breed under changing climate : A systematic review and meta-analysis*. 1–26. https://doi.org/10.1371/journal.pone.0244836

Introduction.pdf. (n.d.).

Jabed, A., Azrifah, M., & Murad, A. (2024). Heliyon Crop yield prediction in agriculture : A comprehensive review of machine learning and deep learning approaches , with insights for future research and sustainability. *Heliyon*, *10*(24), e40836. https://doi.org/10.1016/j.heliyon.2024.e40836

Korir, L., Manning, L., Moore, H. L., Lindahl, J. F., Gemechu, G., Mihret, A., Berg, S., Wood, J. L. N., & Nyokabi, N. S. (2023). Adoption of dairy technologies in smallholder dairy farms in Ethiopia. *Frontiers in Sustainable Food Systems*, *7*(May). https://doi.org/10.3389/fsufs.2023.1070349

Lalman, D., & Stein, D. (2024). *Body Condition Scoring of Cows*. 1–4.

Magana, J., Gavojdian, D., Menahem, Y., Lazebnik, T., Zamansky, A., & Adams-Progar, A. (2023). Machine learning approaches to predict and detect early-onset of digital dermatitis in dairy cows using sensor data. *Frontiers in Veterinary Science*, *10*(November), 1–9. https://doi.org/10.3389/fvets.2023.1295430

Magdalena, J., & Sablik, P. (2025). *The Use of Selected Machine Learning Methods in Dairy Cattle Farming : A Review*. 1–38.

Mayberry, D., Jemberu, W., Schrobback, P., Herrero, M., Chaters, G., Knight-jones, T., & Rushton, J. (2023). *Characterizing Ethiopian cattle production systems for disease burden analysis*. *September*, 1–9. https://doi.org/10.3389/fvets.2023.1233474

Mekonnen, A., Haile, A., Dessie, T., & Mekasha, Y. (2021). *Original Research*. *10*(3), 1–13.

Mengesha, E., Didanna, H. L., & Ayza, A. (2025). Assessing strategy to improve dairy production and milk quality under smallholder dairying: Evidence from Ethiopia. *Acta Scientiarum - Animal Sciences*, *47*(1). https://doi.org/10.4025/actascianimsci.v47i1.70290

Moawed, S. A., Mahrous, E., Elaswad, A., Gouda, H. F., & Fathy, A. (2024). *Milk yield prediction in Friesian cows using linear and flexible discriminant analysis under assumptions violations*. 1–12.

No Title. (n.d.). *August 2025*.

Ouweltjes, W., & Haas, Y. De. (2021). A data-driven prediction of lifetime resilience of dairy cows using commercial sensor data collected during first lactation. *Journal of Dairy Science*, *104*(11), 11759–11769. https://doi.org/10.3168/jds.2021-20413

Salamone, M., Adriaens, I., Vervaet, A., Opsomer, G., Atashi, H., Fievez, V., Aernouts, B., & Hostens, M. (2022). Animal The international journal of animal biosciences Prediction of first test day milk yield using historical records in dairy cows. *Animal*, *16*(11), 100658. https://doi.org/10.1016/j.animal.2022.100658

Shekuru, A. H., Berlie, A. B., & Bizuneh, Y. K. (2022). Variability and trends of temperature and rainfall over three agro-ecological zones in North Shewa, Central Ethiopia. *Arabian Journal of Geosciences*, *15*(18). https://doi.org/10.1007/s12517-022-10674-6

State, T. H. E. (n.d.). *THE STATE OF THE WORLD ' s ANIMAL GENETIC RESOURCES FOR FOOD AND AGRICULTURE*.

Tarihi, K., & Kelimeler, A. (2025). *Makine Öğrenmesinin Süt ve Süt Ürünlerinde Kullanımına Yönelik Meta -Analitik Değerlendirme*. *9*(4), 1223–1232.

Techniques, M. L. (2025). *International Journal of Computer Techniques – Volume 12 Issue 3 , May - June - 2025 Predictive Modeling of Milk Production Using Artificial Intelligence and International Journal of Computer Techniques – Volume 12 Issue 3 , May - June - 2025*. *12*(3), 1–7.

Topuz, D. (2025). *Hybrid Ensemble Model for Lactation Milk Yield Prediction of Holstein Cows*. *31*(5), 603–611. https://doi.org/10.9775/kvfd.2025.34031

Wang, T. (2024). Improved random forest classification model combined with C5 . 0 algorithm for vegetation feature analysis in non - agricultural environments.

Scientific Reports, 1–13. https://doi.org/10.1038/s41598-024-60066-x

Weldegerima, T. M. (2018). *Characterization of productive and reproductive performances , morphometric and challenges and opportunities of indigenous cattle breeds of Ethiopia : A review*. *9*(March), 29–41. https://doi.org/10.5897/IJLP2017.0426

Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M. (2017). Big Data in Smart Farming – A review. *Agricultural Systems*, *153*, 69–80. https://doi.org/10.1016/j.agsy.2017.01.023

Zealand, N., & Tour, S. (n.d.). *Transforming Ethiopia ' s Livestock Sector*.

*, *Lynne A. Carpenter-Boggs*. (2019). 1–50.

Ababa, A. (2014). *TRAINING PACKAGE ON LIVESTOCK HUSBANDRY UNDER SEMI-INTENSIVE MANAGEMENT*. *December*.

Abegaz, S. B. (2022). Milk production status and associated factors among indigenous dairy cows in Raya Kobo district, north eastern Ethiopia. *Veterinary Medicine and Science*, *8*(2), 852–863. https://doi.org/10.1002/vms3.740

Aleme, M., & Mengistu, G. (2023). *Management practice of the Sheko cattle breed in Ethiopia : A review*. *4*, 64–70. https://doi.org/10.46265/genresj.IBNU2035

Asresie, A., & Seifu, L. Z. (2024). *Contribution of Livestock Sector in Ethiopian Economy : A Review Contribution of Livestock Sector in Ethiopian Economy : A Review*. *February*, 78–91.

Bahbahani, H., Afana, A., & Wragg, D. (2018). *Genomic signatures of adaptive introgression and environmental adaptation in the Sheko cattle of southwest Ethiopia*. 1–14. https://doi.org/10.1126/science.1167936.PubMed

Bayou, E., Haile, A., Gizaw, S., & Mekasha, Y. (2015). Evaluation of non - genetic factors affecting calf growth , reproductive performance and milk yield of traditionally managed Sheko cattle in southwest Ethiopia. *SpringerPlus*. https://doi.org/10.1186/s40064-015-1340-9

Begna, R., Asfaw, Y., & Masho, W. (2023). Heliyon Evaluation of trait preferences and effect of parity , season and lactation stage on production performance of indigenous dairy cow in kaffa zone , southwest Ethiopia. *Heliyon*, *9*(12), e22380.

https://doi.org/10.1016/j.heliyon.2023.e22380

Bovo, M., Agrusti, M., Benni, S., Torreggiani, D., & Tassinari, P. (2021). Random forest modelling of milk yield of dairy cows under heat stress conditions. *Animals*, *11*(5). https://doi.org/10.3390/ani11051305

Cao, Y., Dai, J., Zhang, G., & Xia, M. (2024). *Combinations of Feature Selection and Machine Learning Models for Object-Oriented " Staple-Crop-Shifting " Monitoring Based on Gaofen-6 Imagery*.

Chen, T. (n.d.). *XGBoost : A Scalable Tree Boosting System*.

Chen, T., & Guestrin, C. (2016). *XGBoost : A Scalable Tree Boosting System*. 785–794.

Code, L. G., Code, T., Apr, A. G. R., & Apr, A. G. R. (2021). *Animal production Level III Based on March 2018 , Version 3 Occupational standard Activities LG Code : TTLM Code : 0*(March 2018).

Ermias, N., Mesfin, B., Chaka, D., Gezahegn, A., Bikila, S., & Teshome, W. (2024). Transforming agriculture with Machine Learning , Deep Learning , and IoT : perspectives from Ethiopia — challenges and opportunities. In *Discover Agriculture*. Springer International Publishing. https://doi.org/10.1007/s44279-024-00066-7

Friedman, B. J., Hastie, T., & Tibshirani, R. (2000). *Special invited paper*. *28*(2), 337–407.

Frizzarin, M., Visentin, G., Ferragina, A., Hayes, E., Dhariyal, B., Domijan, K., Khan, H., Ifrim, G., Nguyen, T. Le, Meagher, J., Menchetti, L., & Singh, A. (2022). *Classification of cow diet based on milk Mid Infrared Spectra : a data analysis competition at the " International Workshop of Spectroscopy and Chemometrics 2022 ."* 1–27.

Gebreyohanes, G., Yilma, Z., Moyo, S. and, & A., M. O. (2021). Dairy industry development in Ethiopia : Current status , major challenges and potential interventions for improvement. *ILRI POSITION PAPER Dairy*, 1–39.

Giannuzzi, D., Flavio, L., Mota, M., Pegolo, S., Gallo, L., Schiavon, S., Tagliapietra, F., Katz, G., Fainboym, D., & Minuti, A. (2022). In - line near - infrared analysis of milk coupled with machine learning methods for the daily prediction of blood metabolic profile in dairy cattle. *Scientific Reports*, 1–13. https://doi.org/10.1038/s41598-022-11799-0

Gocheva-ilieva, S., & Yordanova, A. (2022). *Predicting the 305-Day Milk Yield of Holstein-Friesian Cows Depending on the Conformation Traits and Farm Using Simplified Selective Ensembles*.

Gudeto, A., Mirkena, T., & Kebede, T. (2021). *Traditional cattle breeding practices and rate of inbreeding in the mid rift valley of Oromia , Ethiopia*. *7*, 77–82.

Guevara-escobar, A., Lemus-ramírez, V., & García-muñiz, J. G. (2026). *Efficient Prediction of Milk Yield with Machine Learning Models Using Cow Identification or Milk Quality Traits*. 1–26.

Guo, R., & Dai, Y. (2025). *Leveraging IGOOSE-XGBoost for the Early Detection of Subclinical Mastitis in Dairy Cows*. 1–18.

Hasan, M. F., Celik, N., Williams, Y., Williams, S. R. O., & Marett, L. C. (2025). An effective way to incorporate temperature – humidity index to study effect of heat stress on milk yield by an XGBoost machine learning model. *Journal of Dairy Science*, *108*(12), 13995–14017. https://doi.org/10.3168/jds.2025-27193

Hatew, B., Peñagaricano, F., Balehegn, M., Jones, C. S., Dahl, G. E., & Adesogan, A. T. (2023). Synergies of feed, management trainings, and genetics on milk production of dairy cows in the tropics: The case of Ethiopian smallholder farmers. *Frontiers in Animal Science*, *4*(March), 1–12. https://doi.org/10.3389/fanim.2023.1119786

health statys.pdf. (n.d.).

Hooker, J., Medeiros, B. B. De, Saha, C., Abdulrahman, T., & Alves, A. A. C. (2025). Evaluating the contribution of behavioral , milking system , and environmental data to short-term milk yield prediction in commercial dairy cows using machine learning. *Journal of Dairy Science*, *108*(10), 11133–11150. https://doi.org/10.3168/jds.2025-26724

Id, M. B., Yigrem, S., Betsha, S., & Tolera, A. (2021). *Production , reproduction and some adaptation characteristics of Boran cattle breed under changing climate : A systematic review and meta-analysis*. 1–26. https://doi.org/10.1371/journal.pone.0244836

Introduction.pdf. (n.d.).

Jabed, A., Azrifah, M., & Murad, A. (2024). Heliyon Crop yield prediction in agriculture : A comprehensive review of machine learning and deep learning

approaches , with insights for future research and sustainability. *Heliyon*, *10*(24), e40836. https://doi.org/10.1016/j.heliyon.2024.e40836

Korir, L., Manning, L., Moore, H. L., Lindahl, J. F., Gemechu, G., Mihret, A., Berg, S., Wood, J. L. N., & Nyokabi, N. S. (2023). Adoption of dairy technologies in smallholder dairy farms in Ethiopia. *Frontiers in Sustainable Food Systems*, *7*(May). https://doi.org/10.3389/fsufs.2023.1070349

Lalman, D., & Stein, D. (2024). *Body Condition Scoring of Cows*. 1–4.

Magana, J., Gavojdian, D., Menahem, Y., Lazebnik, T., Zamansky, A., & Adams-Progar, A. (2023). Machine learning approaches to predict and detect early-onset of digital dermatitis in dairy cows using sensor data. *Frontiers in Veterinary Science*, *10*(November), 1–9. https://doi.org/10.3389/fvets.2023.1295430

Magdalena, J., & Sablik, P. (2025). *The Use of Selected Machine Learning Methods in Dairy Cattle Farming : A Review*. 1–38.

Mayberry, D., Jemberu, W., Schrobback, P., Herrero, M., Chaters, G., Knight-jones, T., & Rushton, J. (2023). *Characterizing Ethiopian cattle production systems for disease burden analysis*. *September*, 1–9. https://doi.org/10.3389/fvets.2023.1233474

Mekonnen, A., Haile, A., Dessie, T., & Mekasha, Y. (2021). *Original Research*. *10*(3), 1–13.

Mengesha, E., Didanna, H. L., & Ayza, A. (2025). Assessing strategy to improve dairy production and milk quality under smallholder dairying: Evidence from Ethiopia. *Acta Scientiarum - Animal Sciences*, *47*(1). https://doi.org/10.4025/actascianimsci.v47i1.70290

Moawed, S. A., Mahrous, E., Elaswad, A., Gouda, H. F., & Fathy, A. (2024). *Milk yield prediction in Friesian cows using linear and flexible discriminant analysis under assumptions violations*. 1–12.

No Title. (n.d.). *August 2025*.

Ouweltjes, W., & Haas, Y. De. (2021). A data-driven prediction of lifetime resilience of dairy cows using commercial sensor data collected during first lactation. *Journal of Dairy Science*, *104*(11), 11759–11769. https://doi.org/10.3168/jds.2021-20413

Salamone, M., Adriaens, I., Vervaet, A., Opsomer, G., Atashi, H., Fievez, V., Aernouts, B., & Hostens, M. (2022). Animal The international journal of animal biosciences Prediction of first test day milk yield using historical records in dairy cows. *Animal*, *16*(11), 100658. https://doi.org/10.1016/j.animal.2022.100658

Shekuru, A. H., Berlie, A. B., & Bizuneh, Y. K. (2022). Variability and trends of temperature and rainfall over three agro-ecological zones in North Shewa, Central Ethiopia. *Arabian Journal of Geosciences*, *15*(18). https://doi.org/10.1007/s12517-022-10674-6

State, T. H. E. (n.d.). *THE STATE OF THE WORLD ' s ANIMAL GENETIC RESOURCES FOR FOOD AND AGRICULTURE*.

Tarihi, K., & Kelimeler, A. (2025). *Makine Öğrenmesinin Süt ve Süt Ürünlerinde Kullanımına Yönelik Meta -Analitik Değerlendirme*. *9*(4), 1223–1232.

Techniques, M. L. (2025). *International Journal of Computer Techniques – Volume 12 Issue 3 , May - June - 2025 Predictive Modeling of Milk Production Using Artificial Intelligence and International Journal of Computer Techniques – Volume 12 Issue 3 , May - June - 2025*. *12*(3), 1–7.

Topuz, D. (2025). *Hybrid Ensemble Model for Lactation Milk Yield Prediction of Holstein Cows*. *31*(5), 603–611. https://doi.org/10.9775/kvfd.2025.34031

Wang, T. (2024). Improved random forest classification model combined with C5 . 0 algorithm for vegetation feature analysis in non - agricultural environments. *Scientific Reports*, 1–13. https://doi.org/10.1038/s41598-024-60066-x

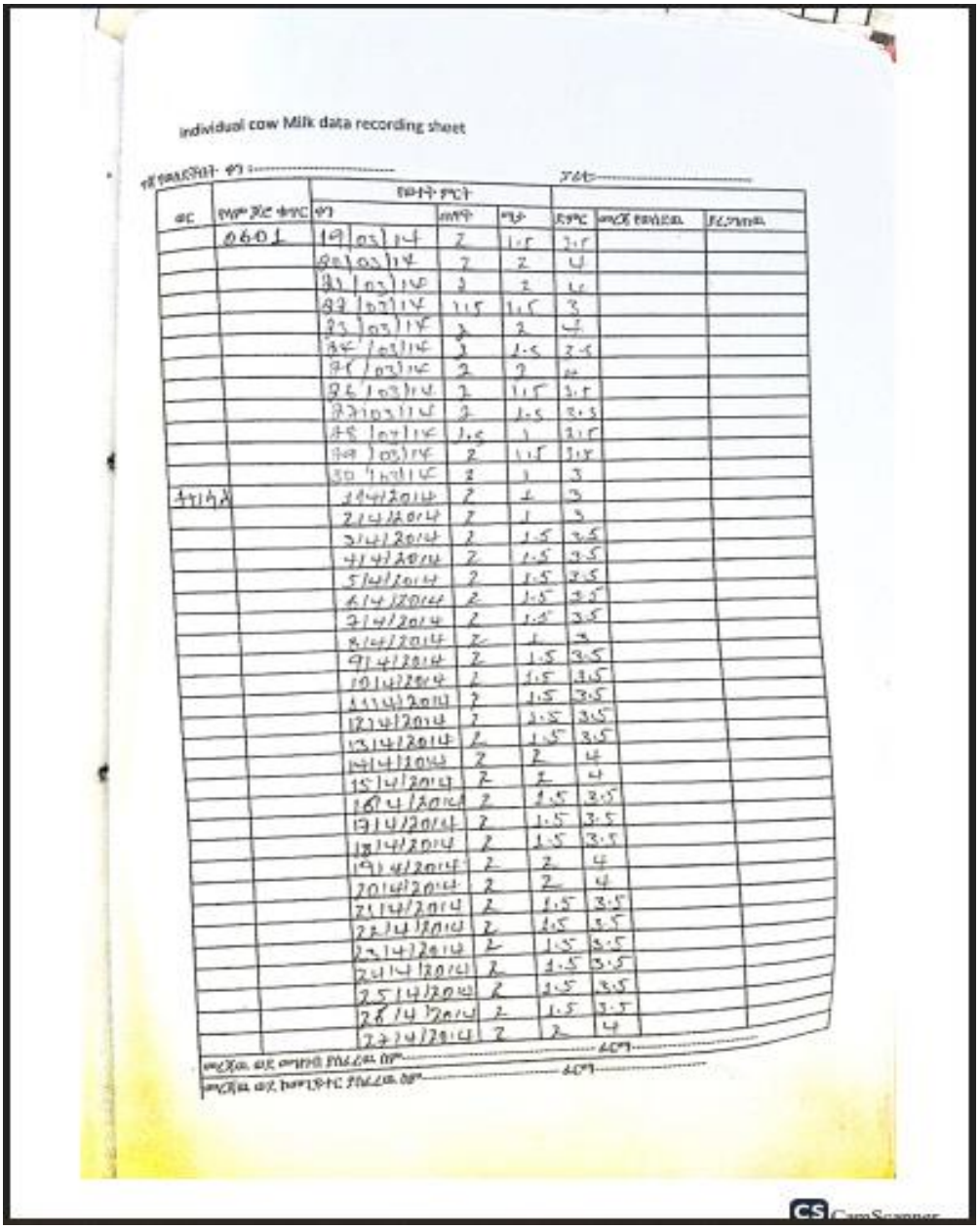
Weldegerima, T. M. (2018). *Characterization of productive and reproductive performances , morphometric and challenges and opportunities of indigenous cattle breeds of Ethiopia : A review*. *9*(March), 29–41. https://doi.org/10.5897/IJLP2017.0426

Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M. (2017). Big Data in Smart Farming – A review. *Agricultural Systems*, *153*, 69–80. https://doi.org/10.1016/j.agsy.2017.01.023

Zealand, N., & Tour, S. (n.d.). *Transforming Ethiopia ' s Livestock Sector*.

Appendix

A) Dataset Sample



individual cow Milk data recording sheet

ቁጥር	የላም ጋር ቁጥር	ቀን	ጠዋት	ማታ	ድምር	ሙሉ የወተቱ መጠን	ያጠባችበት ቀን
	0601	19/03/14	2	1.5	3.5		
		20/03/14	2	2	4		
		21/03/14	2	2	4		
		22/03/14	1.5	1.5	3		
		23/03/14	2	2	4		
		24/03/14	2	1.5	3.5		
		25/03/14	2	2	4		
		26/03/14	2	1.5	3.5		
		27/03/14	2	1.5	3.5		
		28/03/14	1.5	1	2.5		
		29/03/14	2	1.5	3.5		
		30/03/14	2	1	3		
ተባዕ		1/4/2014	2	1	3		
		2/4/2014	2	1	3		
		3/4/2014	2	1.5	3.5		
		4/4/2014	2	1.5	3.5		
		5/4/2014	2	1.5	3.5		
		6/4/2014	2	1.5	3.5		
		7/4/2014	2	1.5	3.5		
		8/4/2014	2	1	3		
		9/4/2014	2	1.5	3.5		
		10/4/2014	2	1.5	3.5		
		11/4/2014	2	1.5	3.5		
		12/4/2014	2	1.5	3.5		
		13/4/2014	2	1.5	3.5		
		14/4/2014	2	2	4		
		15/4/2014	2	2	4		
		16/4/2014	2	1.5	3.5		
		17/4/2014	2	1.5	3.5		
		18/4/2014	2	1.5	3.5		
		19/4/2014	2	2	4		
		20/4/2014	2	2	4		
		21/4/2014	2	1.5	3.5		
		22/4/2014	2	1.5	3.5		
		23/4/2014	2	1.5	3.5		
		24/4/2014	2	1.5	3.5		
		25/4/2014	2	1.5	3.5		
		26/4/2014	2	1.5	3.5		
		27/4/2014	2	2	4		

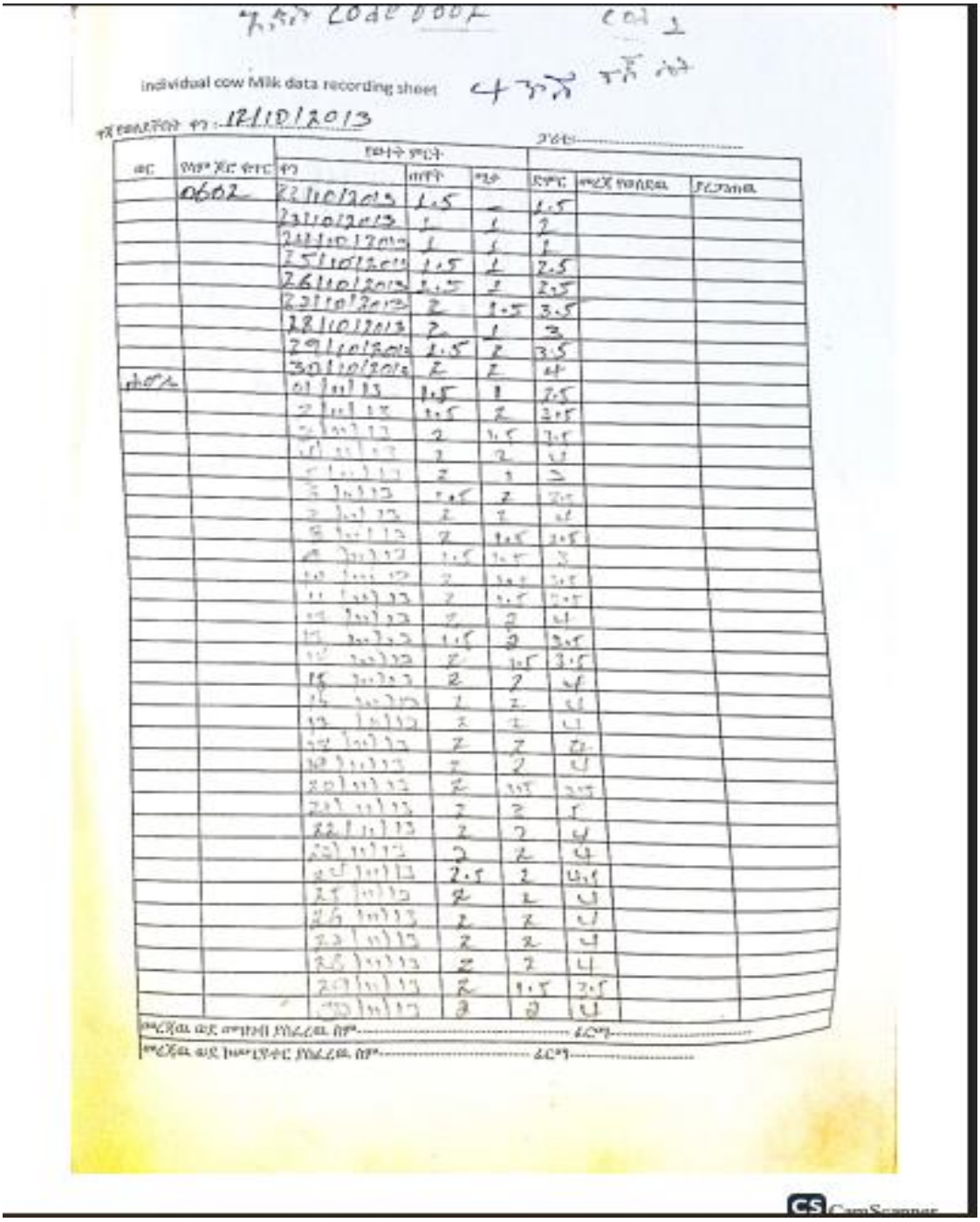
0602

individual cow Milk data recording sheet

ቀን	የላም ጋር ቁጥር	ጠዋት ምርት				መኖ የመስጠት	ያረጋገጠው
		ቀን	ጠዋት	ማታ	ድምር		
	ላም ቁጥር	1/12/2013	2	1.5	3.5		
		2/12/2013	2	1.5	3.5		
		3/12/2013	2	2	4		
		4/12/2013	1.5	2	3.5		
		5/12/2013	2	2	4		
		6/12/2013	2	2	4		
		7/12/2013	2	2	4		
		8/12/2013	2	2	4		
		9/12/2013	2	2	4		
		10/12/2013	2	1.5	3.5		
		11/12/2013	2	2	4		
		12/12/2013	2	2	4		
		13/12/2013	2	2	4		
		14/12/2013	2	2	4		
		15/12/2013	2	2	4		
		16/12/2013	2	2	4		
		17/12/2013	2	1.5	3.5		
		18/12/2013	2	1.5	3.5		
		19/12/2013	2.5	1.5	4		
		20/12/2013	2	1.5	3.5		
		21/12/2013	2	2	4		
		22/12/2013	2	1.5	3.5		
		23/12/2013	2	1.5	3.5		
		24/12/2013	2.5	1.5	3.5		
		25/12/2013	2	1.5	3.5		
		26/12/2013	2	2	2		
		27/12/2013	1.5	2	7.5		
		28/12/2013	2	2	3		
		29/12/2013	1.5	2	2.5		
		30/12/2013	1.5	2	7.5		
ላም ቁጥር		1/13/2013	1.5	1.5	3		
		2/13/2013	2	1.5	3.5		
		3/13/2013	2	1.5	3.5		
		4/13/2013	2	2	4		
		5/13/2013	2	1.5	3.5		
ላም ቁጥር		[illegible]	2	2	4		
		[illegible]	2.5	2	4.5		
		[illegible]	1.5	1.5	3		
		[illegible]	2	1.5	3.5		

[illegible] ..... [illegible] .....

[illegible] ..... [illegible] .....



Individual cow Milk data recording sheet

ተ.ቁ	የላም ጆር ቁጥር	የወተት መረጃ					
		ቀን	ጠዋት	ማታ	ድምር	መረጃ የወሰደው	ያረጋገጠው
	0602	23/10/2013	1.5	-	1.5		
		24/10/2013	1	1	2		
		25/10/2013	1	1	1		
		26/10/2013	1.5	1	2.5		
		27/10/2013	2	1.5	3.5		
		28/10/2013	2	1	3		
		29/10/2013	1.5	2	3.5		
		30/10/2013	2	2	4		
ሐመረ		31/10/13	1.5	1	2.5		
		1/11/13	1.5	2	3.5		
		2/11/13	2	1.5	3.5		
		3/11/13	2	2	4		
		4/11/13	2	1	3		
		5/11/13	1.5	2	3.5		
		6/11/13	2	2	4		
		7/11/13	2	1.5	3.5		
		8/11/13	1.5	1.5	3		
		9/11/13	2	1.5	3.5		
		10/11/13	2	2	4		
		11/11/13	2	1.5	3.5		
		12/11/13	2	2	4		
		13/11/13	1.5	2	3.5		
		14/11/13	2	1.5	3.5		
		15/11/13	2	2	4		
		16/11/13	2	2	4		
		17/11/13	2	2	4		
		18/11/13	2	2	4		
		19/11/13	2	1.5	3.5		
		20/11/13	2	2	4		
		21/11/13	2	2	4		
		22/11/13	2	2	4		
		23/11/13	2	2	4		
		24/11/13	2.5	2	4.5		
		25/11/13	2	2	4		
		26/11/13	2	2	4		
		27/11/13	2	2	4		
		28/11/13	2	2	4		
		29/11/13	2	1.5	3.5		
		30/11/13	2	2	4		

CamScanner

B) Loading Dataset

```
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import (
    train_test_split,
    StratifiedKFold,
    GridSearchCV,
    cross_val_score
)

from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import StandardScaler

from sklearn.metrics import (
    accuracy_score,
    precision_score,
    recall_score,
    f1_score,
    roc_auc_score,
    confusion_matrix,
    classification_report,
    roc_curve
)
```

```
)

from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import (
    RandomForestClassifier,
    AdaBoostClassifier
)

from xgboost import XGBClassifier

from imblearn.pipeline import Pipeline
from imblearn.over_sampling import SMOTE
```

```
from google.colab import drive
drive.mount('/content/drive', force_remount=True)
```

```
Mounted at /content/drive
```

```
1 import pandas as pd
2
3 df = pd.read_csv('/content/drive/MyDrive/New_Dataset.csv')
4 df.head()
```

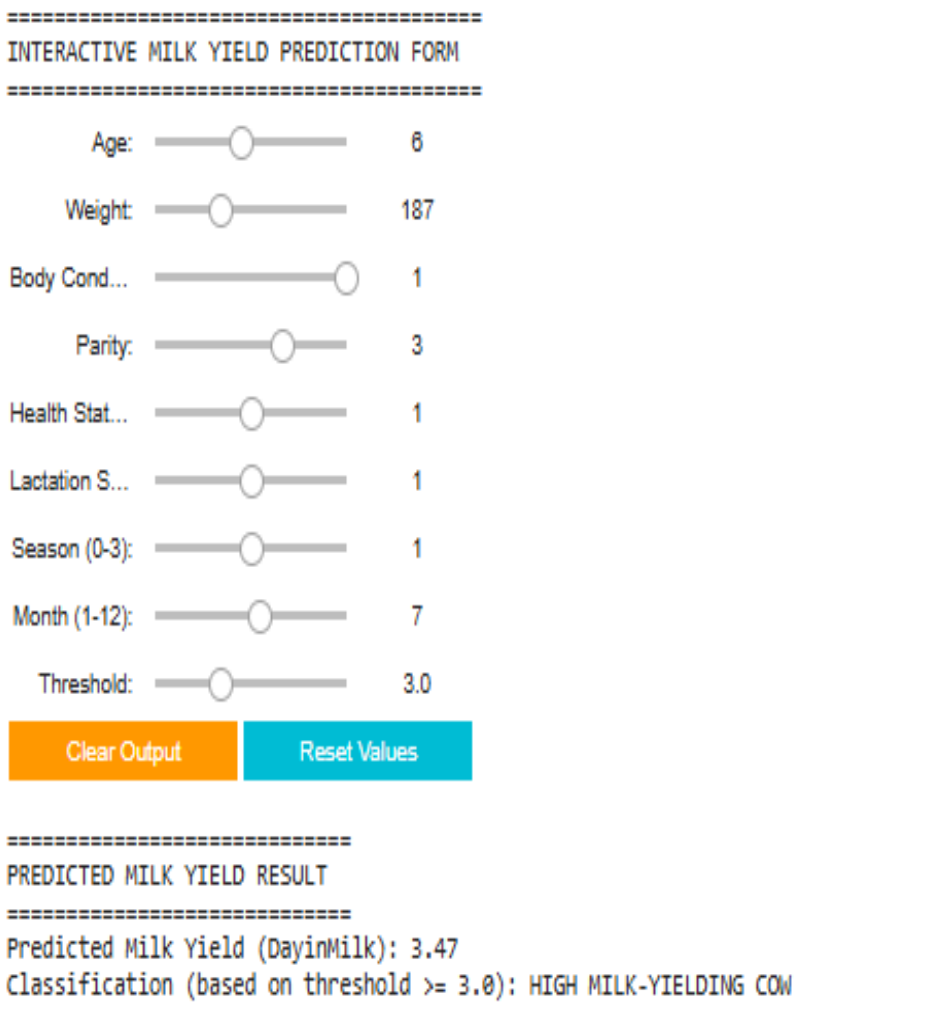
	Date	CowID	Age	Weight	BodyCondition	Parity	DayinMilk	HealthStatus	LactationStage	Season	YieldCategory
0	2014-08-26	600	7	187	1	3	4.0	1	0	1	1
1	2014-08-27	600	7	187	1	3	3.5	1	0	1	1
2	2014-08-28	600	7	187	1	3	3.5	1	0	1	1
3	2014-08-29	600	7	187	1	3	3.5	1	0	1	1
4	2014-08-30	600	7	187	1	3	3.5	1	0	1	1

```
Dataset Shape: (9538, 11)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9538 entries, 0 to 9537
Data columns (total 11 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Date            9538 non-null   object
 1   CowID           9538 non-null   int64
 2   Age             9538 non-null   int64
 3   Weight          9538 non-null   int64
 4   BodyCondition   9538 non-null   int64
 5   Parity          9538 non-null   int64
 6   DayinMilk       9538 non-null   float64
 7   HealthStatus    9538 non-null   int64
 8   LactationStage  9538 non-null   int64
 9   Season          9538 non-null   int64
 10  YieldCategory   9538 non-null   int64
dtypes: float64(1), int64(9), object(1)
memory usage: 819.8+ KB
None
Date              0
CowID             0
Age               0
Weight            0
BodyCondition     0
Parity            0
DayinMilk         0
HealthStatus      0
LactationStage    0
Season            0
YieldCategory     0
dtype: int64
```

```
X_train, X_test, y_train, y_test = train_test_split(
    X,
    y,
    test_size=0.2,
    random_state=42
)

print(X_train.shape)
print(X_test.shape)
```

```
(7630, 10)
(1908, 10)
```


INTERACTIVE MILK YIELD PREDICTION FORM
Age: 6
Weight: 187
Body Cond... 1
Parity: 3
Health Stat... 1
Lactation S... 1
Season (0-3): 1
Month (1-12): 7
Threshold: 3.0
Clear Output
Reset Values
PREDICTED MILK YIELD RESULT
Predicted Milk Yield (DayinMilk): 3.47
Classification (based on threshold >= 3.0): HIGH MILK-YIELDING COW

C) Aproval Sheet



በኢትዮጵያ የግብርና ምርምር ኢንስቲትዩት
የቴፒ ግብርና ምርምር ማዕከል
ስልክ ቁ. + 251-475-560356
ፋክስ + 251-475-560087
E-mail:TepiAgriculturalResearchCenter@eiar.gov.et
ፖ.ሣ.ቁ34 ቴፒ ኢትዮጵያ

Ethiopian Institute of Agricultural Research
Tepi Agricultural Research Center
Tel:- + 251-475-560356
Fax: + 251-475-560087
E-mail:TepiAgriculturalResearchCenter@eiar.gov.et
P.O. Box 34 Tepi, Ethiopia

Ref. 19/2062
Date 22/05/2026

To FDRE TVET Institute, Research and Dissemination Directorate
Addis Ababa

Subject: Response to Request for Collaboration and Data Access for Postgraduate Research

Dear Sir/Madam,

I am writing to formally inform you about the research conducted by Mr. Sultan Ahmed Hussen as part of his Master of Science (MSc) thesis entitled, "Predictive Modeling Using Machine Learning Techniques for Identifying High Milk-Yielding Sheko Breed Cows in Sheka Zone." The study focuses on the application of machine learning techniques to identify high milk-producing Sheko breed cows using biological and environmental characteristics. The research aims to support evidence-based livestock management and improve dairy productivity in the study area.

The dataset used for this research was officially obtained from the Livestock Research Department of Tepi Agricultural Research Center. Key variables included cow age, parity, lactation stage, health status, breed, season, body condition, and daily milk yield. These attributes were carefully recorded and used to develop predictive models for classifying cows with high milk-production potential. This letter confirms that the data were obtained from a legitimate institutional source and were used strictly for academic purposes.

Sincerely,

Dr. Doreje Tulu Robi
Center Director

የኢአምኢ ዋና መሥሪያቤት EIAR Headquarters
Tel: (251-11) 6462334/6465/4443/6462270Fax: (251-11) 6461294 P.O. Box: 2003
6461388/6460379http://www: eiar.gov. et Addis Ababa, Ethiopia

CamScanner