AN ENSEMBLE LEARNING MODEL FOR PREDICTION OF ARTIFICIAL INSEMINATION OUTCOMES IN KENYAN DAIRY COWS: A CASE STUDY.

BY

ERICK KIPKEMOI RUTTO

MASTER OF SCIENCE IN DATA ANALYTICS

KCA UNIVERSITY

2025

AN ENSEMBLE LEARNING MODEL FOR PREDICTION OF ARTIFICIAL INSEMINATION OUTCOMES IN KENYAN DAIRY COWS: A CASE STUDY.

BY

ERICK KIPKEMOI RUTTO

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTER OF SCIENCE IN DATA ANALYTICS IN THE SCHOOL OF TECHNOLOGY AT KCA UNIVERSITY.

JULY, 2025

DECLARATION

I declare that the work in this dissertation has not been previously published or submitted elsewhere for award of a degree. I also declare that this my own original work and contains no material written of published by other people except where due reference is made and author duly acknowledged.

Student Name:	**Erick Kipkemoi Rutto**	Reg. No	**19/06544**
Sign		Date:	**06/12/2025**

I do hereby confirm that I have examined the master's dissertation of:

Erick Kipkemoi Rutto

And have certified that all revisions that the dissertation panel and examiners recommended have been adequately addressed.

Sign:	Digitally signed by Prof. Dr. Lucy W. Mburu Date: 2026.01.09 22:52:17 +03'00'	Date: 09/01/2026	
Dissertation Supervisor Name:	**Prof. Dr. Lucy Mburu**		

ABSTRACT

Artificial insemination is a vital reproductive technology for smallholder dairy systems, yet its adoption in Kenya and East Africa remains low majorly due to poor success rates influenced by complex management, animal-related, and farmer-related factors. The ability to predict the outcome of each insemination based on the available influencing factors using a decision support tool will boost the adoption of this breeding technology. The background of such a tool is machine learning models. This case study sought to develop an ensemble machine learning model to predict the artificial insemination outcome of dairy cows in smallholder systems. This was addressed through four objectives: assessing influencing factors for artificial insemination outcome, evaluating existing machine learning models using smallholder farmer data, developing a tailored ensemble model for predicting insemination outcome in dairy cows in smallholder systems, and testing performance on heterogeneous data from small scale farmers. The study utilized data on pregnancy diagnosis outcome obtained from smallholder dairy farmers in Kakamega and Kisumu counties of Kenya. Additional data from regions with the same ecological zones in Tanzania and Ethiopia were incorporated to corroborate the findings in the east African context. A total of 1347 pregnancy diagnosis outcome records were used to test the predictive ability of five models such as Support Vector Machines (SVM), Naive Bayes (NB), K-Nearest Neighbors (KNN), Random Forest (RF) and Decision Trees (DT). The outcomes of these models were then fitted to Logistic Regression (LR) stacked ensemble model and its performance compared with XGBoost stacked ensemble. Results from feature importance analysis identified estrus synchronization (0.23), cow age (0.16), body condition score (0.13), membership to cooperative (0.09), and fodder growing (8.81%) as key predictors, while conventional health factors like vaccination and deworming showed minimal impact. Assessing existing base models for this type of data it was found that Random Forest (RF) outperformed all others (accuracy: 0.88). Logistic Regression-based ensemble achieved robust results (accuracy: 0.86) and exceeded XGBoost ensembled model but was outperformed by RF. The study highlights the importance of management decisions over traditional health interventions and the strengths of tree-based models in handling such type of data providing a framework for data-driven reproductive management.

Key words: Stacked ensemble, artificial insemination, dairy cows, smallholder systems, machine learning, precision livestock farming.

AKNOWLEDGEMENT

First and foremost, I would like to thank God for the good health, strength and wisdom He granted me during this study. Secondly, I would like to express my sincere gratitude to my supervisor Prof. Lucy Mburu, for her invaluable insights and direction in mentoring and supporting me to complete this project as well as my lecturers from the school of Technology at KCA University. I would also like to thank my work supervisors Dr. Domelevo Entfellner Jean-Baka and Dr. Mwai Okeyo from ILRI for their mentorship and support during this course.

ACRONYMS AND ABBREVIATIONS

DSS-Decision Support System

AI-Artificial Intelligence

SVM- Support Vector Machine

BCS-Body Condition Score

ROC-Receiver Operating Characteristics

KNN- K-Nearest Neighbours

ILRI- International Livestock Research Institute

FP- False Positives

TP-True Positives

PCA - Principle Component Analysis

RF- Random Forests

PLF - Precision Livestock Farming

GDP **-** Gross Domestic Product

ML- Machine learning

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

DEFINITION ON TERMS

Machine learning: is a subset of Artificial Intelligence that enables computers to learn from data and improve their performance on tasks by their own.

Ensemble learning: is a technique in machine learning that aggregate outputs from multiple models to produce a more accurate, robust, and generalizable prediction than any single model alone.

Synthetic Minority Oversampling Technique (SMOTE): is a method used to balance imbalanced datasets—especially when one class (usually the "positive" class) has far fewer examples than the others.

Gross Domestic Product: is the total value of all final goods and services in monetary terms that are produced or provided within a country's borders in a specific time period.

Artificial Insemination: is a reproductive technology where sperm is collected from a male animal, processed, and manually inserted into the female's reproductive tract to achieve pregnancy. It is widely used in dairy and beef cattle, goats, sheep, and pigs to improve genetics, productivity, and disease control.

Estrus: is a state of sexual receptivity during which the female will accept the male and is capable of conceiving.

Estrus synchronization: is a reproductive management technique used in agriculture to coordinate the estrous cycles of female animals, primarily cattle, so they come into heat (estrus) at or around the same time.

CHAPTER ONE: INTRODUCTION

1.1 Background of the study

The dairy sector remains a cornerstone of agricultural economies worldwide. Notably, smallholder dairy farming systems play a central role in food security, rural livelihoods, and poverty reduction across developing economies. In East Africa, the dairy industry is particularly vital, contributing both to household nutrition and national economies. The region currently consumes approximately 12.4 billion liters of milk annually, with demand projected to rise at a compound annual growth rate (CAGR) of 4–8% (FAO, 2023; AU-IBAR, 2022). Kenya leads in regional milk consumption, with one of the highest per capita averages of around 120 liters per person per year, underscoring the cultural and dietary significance of dairy products in the country. Economically, the sector is a major contributor, accounting for 3.8% of the total national Gross Domestic Product (GDP) and approximately 17% of agricultural GDP (KNBS, 2023). The sector's backbone is the 1.8 million smallholder dairy farmers who collectively produce nearly 80% of the total milk supply (ILRI, 2020; KDB, 2022).

Despite its importance, the livestock sector faces mounting challenges in meeting the region's rapidly growing demand for milk and other livestock products. Projections indicate that milk demand in East Africa could surpass 15 billion liters by 2030 (AU-IBAR, 2023). Kenya's Dairy Master Plan estimates a sharp rise in national milk demand, from 4.5 billion liters currently to 12.67 billion liters by 2030, driven by population growth, urbanization, and rising per capita consumption. To meet these targets, Kenya would need to increase milk production by approximately 150% over current levels (KDB, 2022). Achieving such growth requires a shift from extensive to intensive production systems, emphasizing improved genetics, better feeding regimes, and efficient herd management to enhance per-cow productivity.

Artificial insemination is recognized as a key reproductive technology with the potential to drive genetic improvement and boost dairy productivity. In developed countries, artificial insemination has been widely adopted and has significantly transformed livestock systems. However, in developing countries, including Kenya, its implementation remains limited and the outcomes achieved are often unsatisfactory (Omondi et al., 2017). Current adoption is constrained by variable conception rates, often ranging between 40–60%, which contribute to economic losses, delayed calving intervals, and reduced milk productivity (Ojango et al., 2019).

The primary bottleneck is inaccurate estrus (heat) detection, which remains the single most important determinant of insemination success (Walsh et al., 2011; Lucy, 2001). In smallholder systems, estrus detection is typically carried out through traditional visual observation methods. These are prone to errors due to limited labor availability, insufficient technical knowledge, and reduced observation periods, especially where farm families combine dairy with other livelihood activities (Mutiga et al., 2021). Consequently, many inseminations are mistimed, leading to lower conception rates and disincentivizing farmers from continued use of artificial insemination. This situation perpetuates the low adoption rates (often below 30%) and undermines the potential of artificial insemination to transform the sector.

Emerging evidence suggests that predictive modeling and machine learning (ML) approaches can provide solutions to these challenges. By analyzing multiple variables such as body condition score, cow age, parity, nutritional status, management practices, and environmental conditions ML algorithms can generate accurate forecasts of insemination success (Gupta et al., 2023). Such tools have the potential to complement or even substitute traditional estrus detection methods by offering data-driven, timely, and reliable recommendations. In resource-constrained smallholder settings, predictive models are particularly valuable, as they enable targeted interventions, optimize insemination timing, and improve conception rates while minimizing costs (Mwanga et al., 2020).

When locally calibrated, ML-based decision-support systems have the potential to revolutionize reproductive management in Kenya and other East African countries. By enhancing the efficiency, affordability, and scalability of artificial insemination services, these technologies can contribute to improved milk productivity, household incomes, and food security. Moreover, they align with broader agricultural modernization goals, including digital transformation, climate resilience, and inclusive innovation.

1.2 Historical Context of Artificial Insemination in East Africa

Artificial insemination was first introduced in East Africa in the mid-20th century as part of colonial and post-independence efforts to modernize agriculture (Mugisha et al., 2014; Bebe et al., 2020). Early programs focused on crossbreeding indigenous cattle with exotic breeds to improve milk yield and growth rates. However, these initiatives were largely top-down, with limited involvement of smallholder farmers and inadequate infrastructure to support widespread adoption (Ojango et al., 2019).

In Kenya, the Kenya Animal Genetic Resources Centre (KAGRC) was established to produce and distribute semen from high-performing bulls (Kibiego et al., 2020). Despite these efforts, artificial insemination adoption remained low due to logistical challenges, poor service delivery, and lack of farmer awareness (Kabunga et al., 2014; Ngeno et al., 2019). Similar patterns were observed in Uganda, Tanzania, and Ethiopia, where artificial insemination services were concentrated in urban and peri-urban areas, leaving rural smallholders underserved (Lemma et al., 2021; Mugabo et al., 2021).

Over the decades, various donor-funded projects have attempted to revitalize artificial insemination services, including the East Africa Dairy Development (EADD) project and the Africa Dairy Genetic Gains (ADGG) initiative by International Livestock Research Institute (ILRI) (Omondi et al., 2017; Mrode et al., 2020). These programs emphasized capacity building, infrastructure development, and farmer education. While they achieved some success, systemic barriers such as unreliable estrus detection, poor record-keeping, and limited access to trained technicians continued to hinder progress (EADD, 2013).

The emergence of digital agriculture and mobile-based extension services in the 2010s opened new possibilities for artificial insemination delivery. Mobile apps, SMS alerts, and cloud-based data platforms began to support estrus tracking, technician scheduling, and performance monitoring (GSMA, 2023). However, these tools often lacked predictive capabilities and were not tailored to the diverse conditions of smallholder farms (Buller et al., 2020).

The integration of machine learning into artificial insemination decision-making represents a new frontier in reproductive management. By leveraging local data and advanced analytics, ML models can provide personalized recommendations, reduce uncertainty, and enhance farmer confidence in artificial insemination technology (Zhang et al., 2023; Ouma et al., 2024). This shift from reactive to predictive breeding strategies marks a significant evolution in the region's approach to livestock improvement.

1.2 Problem statement

Although artificial insemination is widely recognized for improving dairy genetics and productivity, its adoption among smallholder farmers in East Africa remains low ranging from only 5% to 40% compared to over 80% in high-income countries (Kabunga et al., 2014; FAO, 2023). This limited uptake persists despite decades of promotion, largely due to costly insemination failures, poor estrus detection accuracy, and limited access to skilled technicians (Kibiego et al., 2020; Walsh et al., 2011). Conception rates in these systems often fall below 50%, deterring continued use and reducing profitability. Existing artificial insemination success prediction models, developed for high-input commercial systems, perform poorly in East African contexts where data are sparse, noisy, and heterogeneous (Gupta et al., 2023). This research addresses that gap by developing a locally adapted ensemble learning model capable of handling mixed, imbalanced datasets to predict artificial insemination outcomes more accurately. Such a tool is urgently needed to guide on-farm breeding decisions, reduce insemination failures, and enhance the adoption of artificial insemination technologies crucial for dairy productivity, food security, and rural livelihoods in Kenya and the wider region.

1.3 Main objective of the study

The overarching aim of this study was to develop a predictive model for determining the likelihood of artificial insemination success in dairy cows within smallholder farming systems in Kenya using ensemble machine learning techniques.

1.4 Specific objectives of the study

To achieve the main objective, the study was guided by the following specific objectives:

1. To identify and assess the key factors influencing the success or failure of artificial insemination in smallholder dairy farms.
2. To evaluate the performance of existing machine learning models in predicting artificial insemination outcomes in smallholder dairy systems.
3. To develop a stacked ensemble learning model tailored to smallholder dairy conditions for predicting artificial insemination success.
4. To validate the performance of the ensemble model using real-world, imbalanced, and mixed-type data from smallholder dairy farms.

1.5 Research Questions

To guide the investigation and ensure alignment with the study's objectives, the following research questions were formulated:

1. **What are the key factors influencing the success or failure of artificial insemination in smallholder dairy farms?**
   This question seeks to uncover the animal-related, farmer-related, managerial, and feed-related variables that significantly affect conception outcomes. It aims to provide a comprehensive understanding of the multifactorial nature of artificial insemination success in smallholder systems.
2. **Can existing machine learning models accurately predict artificial insemination outcomes in smallholder dairy systems?**

This question evaluates the applicability and performance of conventional machine learning algorithms such as Support Vector Machines, Decision Trees, Random Forests, Naïve Bayes, and K-Nearest Neighbors when applied to real-world data from smallholder farms.

3. **Does a stacked ensemble learning model improve prediction accuracy of artificial insemination success compared to individual machine learning models?** This question investigates whether combining multiple base learners through a meta-model enhances predictive performance, particularly in the context of heterogeneous and imbalanced datasets typical of smallholder systems.
4. **How well does the ensemble learning model perform when validated on real-world, imbalanced, and mixed-type data from smallholder dairy farms?** This question assesses the robustness, generalizability, and practical utility of the developed ensemble model, focusing on its ability to handle noisy, incomplete, and diverse data commonly encountered in field conditions.

1.6 Significance of the Study

This research is of substantial significance to Kenya and the broader East African smallholder dairy sector, as it directly addresses persistent challenges that have hindered the effective adoption and impact of artificial insemination. It is significant particularly to smallholder farmers, agricultural researchers, policymakers, and technology developers.

For Farmers: This study directly addresses the persistent barriers that limit smallholder dairy farmers' adoption of artificial insemination, particularly low conception rates, repeated insemination costs, and inefficiencies in reproductive management. By applying machine learning to predict insemination outcomes, farmers are provided with a pathway to reduce financial losses from failed services, a critical concern for resource-constrained households (Mrode et al., 2020). Through feature importance ranking, the study identifies key farmers, animal, management, and environmental factors affecting artificial insemination outcomes, enabling farmers and technicians to adopt more precise insemination strategies. In turn, this enhances reproductive efficiency, herd genetics, and productivity, ultimately improving milk supply, food security, and household resilience in rural communities.

For Policy Makers: The findings provide evidence-based insights that can guide breeding policies, extension services, and subsidy frameworks aimed at improving adoption (Gupta et al., 2023). By demonstrating the potential of ML in optimizing insemination outcomes, the study supports the sustainable intensification of dairy systems in alignment with the African Union's Agenda 2063, which emphasizes the role of science and innovation in agricultural transformation (African Union, 2015), and the East African Community's (EAC) livestock policies that prioritize improved breeding services (EAC, 2021). Policymakers can use these insights to design targeted interventions that make artificial insemination services more affordable, accessible, and effective for smallholder farmers.

For Technologists: The technological significance of this research lies in the development of a validated predictive architecture that can be scaled and adapted to low-resource smallholder farming systems. Leveraging the growing mobile penetration and digital infrastructure in East Africa (GSMA, 2023), ML-powered decision-support tools can be delivered through mobile and cloud-based platforms. Such farmer-friendly applications ensure accessibility in rural areas and contribute to digital agriculture innovations that foster inclusivity and sustainability. The study thus offers technologists a practical framework for bridging the gap between advanced machine learning models and everyday farm management.

For Researchers: From a scholarly perspective, this research fills a critical gap in the literature by focusing on artificial insemination outcome prediction in smallholder dairy systems, a context that has received limited attention in machine learning research. Most existing models are designed for high-input, commercial farms in developed countries and are poorly suited to the realities of East African agriculture. By using locally sourced data and tailoring the model to smallholder conditions, this study advances the field of context-aware predictive modeling in agriculture. Additionally, the study contributes to methodological innovation by comparing the performance of individual ML models with that of a stacked ensemble, offering insights into model selection, tuning, and validation in real-world agricultural datasets.

1.7 Motivation of the Study

The motivation for this study stems from the persistent underperformance of artificial insemination services in East Africa, particularly within smallholder dairy systems. Despite decades of promotion and the proven potential of artificial insemination to enhance genetic gains and milk productivity, adoption rates remain below 30% in many parts of the region (Ojango et al., 2019; KDB, 2022). This low uptake reflects a complex interplay of technical, economic, and institutional challenges that continue to undermine the effectiveness of artificial insemination as a reproductive technology.

At the heart of the problem is the high financial risk associated with failed inseminations. For smallholder farmers, each artificial insemination attempt represents a significant investment not only in terms of direct costs for semen and technician services but also in terms of opportunity costs related to delayed conception, extended calving intervals, and reduced milk yields. When insemination fails, the economic consequences can be severe, particularly for households that rely on dairy as a primary source of income and nutrition. These risks discourage repeated use of artificial insemination and contribute to a cycle of low adoption and poor reproductive performance.

Another key motivation is the inefficiency of traditional estrus detection methods, which are heavily reliant on visual observation and farmer experience. In smallholder systems, where labor is often shared across multiple livelihood activities, the time and expertise required for accurate heat detection are frequently lacking. This leads to mistimed inseminations and low conception rates, further eroding farmer confidence in artificial insemination services. While training programs and extension services have attempted to address this gap, their reach and impact remain limited, especially in remote and underserved areas.

The study is also motivated by the limited availability of context-specific decision-support tools that can guide AI use in smallholder settings. Most existing predictive models have been developed in high-input, commercial dairy systems and are ill-suited to the realities of East African agriculture. These models often assume access to high-quality data, consistent feeding regimes, and advanced veterinary care conditions that are rarely met in smallholder contexts. As a result, they fail to provide reliable guidance to farmers and technicians operating under resource constraints.

From a technological perspective, the study is driven by the growing potential of machine learning (ML) to transform agricultural decision-making. ML algorithms can analyze complex, multidimensional datasets to uncover patterns and generate accurate predictions, even in the presence of noisy or incomplete data. In the context of artificial insemination, ML offers a powerful tool for improving estrus detection, optimizing insemination timing, and increasing conception rates. However, the application of ML in smallholder dairy systems remains underexplored, and there is a pressing need for models that are tailored to local conditions and validated using real-world data.

Finally, the study is motivated by a broader vision of agricultural transformation in East Africa. By improving the efficiency and reliability of artificial insemination services, this research aims to contribute to increased milk productivity, enhanced household incomes, and improved food security. It also aligns with national and regional development goals that emphasize the role of science, technology, and innovation in driving inclusive and sustainable growth. In this sense, the study is not only a response to a technical challenge but also a contribution to the empowerment of smallholder farmers and the modernization of the dairy sector.

1.8 Scope and Limitations

The study focuses on dairy cattle in smallholder farming systems within East Africa, specifically in Kenya, using datasets containing reproductive, health, management, and environmental variables. The scope is limited to prediction of artificial conception success, using existing algorithms and developing an ensemble model for the same. Limitations include potential biases arising from incomplete records, variability in data quality, and constraints on real-time data acquisition.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

The literature review provides a comprehensive synthesis of existing research on artificial insemination in dairy cattle, with a particular focus on smallholder systems in Kenya and other parts of East Africa. It explores the theoretical underpinnings of artificial insemination adoption, the role of machine learning (ML) in reproductive prediction, and the empirical evidence supporting various modelling approaches. The review is structured to address four key areas:

1. Theoretical foundations of artificial insemination and ML in livestock reproduction.
2. Empirical studies on artificial insemination success prediction using ML.
3. Comparative analysis of ML models used in artificial insemination outcome prediction.
4. Identification of research gaps and justification for the current study.

This chapter lays the groundwork for the development of a context-specific ensemble learning model by highlighting the limitations of existing approaches and the potential of ML to transform reproductive decision-making in smallholder dairy systems.

2.2 Theoretical review

Artificial insemination has long been recognized as a key reproductive technology with the potential to significantly enhance milk productivity in the dairy sector. However, adoption among smallholder farmers in the country remains limited, primarily due to uncertainty surrounding conception outcomes (Shalloo et al., 2014; Cabrera, 2014). Economic constraints exacerbate this challenge: service fees range from KES 1,500 to 3,000 per insemination, which is a substantial investment for smallholders. These costs are further compounded by expenses for pregnancy diagnosis and the need for repeat services, often resulting from failed or low conception rates (Kibiego et al., 2020).

Conception rates as low as 30–40% have been documented, discouraging continued artificial insemination use by small-scale farmers. Several factors contribute to these low success rates, including inaccurate heat detection, poorly timed inseminations, and substandard semen quality (Ngeno et al., 2019). Such technical and management limitations, combined with the financial burden of repeated attempts, diminish farmers' confidence in the technology.

Enhancing farmers' ability to anticipate artificial insemination outcomes before committing resources could transform adoption rates and reproductive efficiency. Early and reliable prediction of conception probability would allow for better decision-making, targeted breeding interventions, and reduced economic risk (Buller et al., 2020). Achieving this requires the development of accessible decision-support tools that integrate multiple determinants of conception success into robust, predictive models (Zhang et al., 2023).

Machine learning provides a promising pathway toward this goal. By analyzing complex, multidimensional datasets such as animal health metrics, metabolic indicators, environmental conditions, and historical breeding records ML algorithms can generate accurate, data-driven predictions of artificial insemination outcomes (Ouma et al., 2024). As a core component of artificial intelligence systems, ML has the capacity to revolutionize smallholder dairy reproduction management, offering scalable, adaptable, and cost-effective solutions (Mwilawa et al., 2023).

This review therefore synthesizes existing literature on the biological, environmental, and management factors influencing artificial insemination success rates, alongside current advances in the application of machine learning for predictive modelling in smallholder dairy production systems.

2.3 Theoretical framework

A growing body of research demonstrates that machine learning technologies can play a pivotal role in enhancing artificial insemination adoption and success rates in dairy farming. These advancements are driven by ML's ability to process complex, multidimensional datasets and identify patterns that support more precise reproductive management. In particular, ML has shown significant potential in improving estrus detection accuracy, enabling genetic selection for desirable traits, providing reliable pregnancy success predictions, and facilitating automated reproductive monitoring.

This framework draws on established theories and empirical findings to explain how integrating ML into dairy reproduction management can address persistent challenges faced by smallholder farmers particularly uncertainty in conception outcomes, high costs associated with repeat inseminations, and limitations in on-farm reproductive expertise. The following subsections outline the theoretical perspectives underpinning these applications, as discussed in the literature.

2.3.1 Optimal Insemination Timing Prediction

This theory stipulates that machine learning can help in achieving high conception rates. The success of artificial insemination in dairy cows is highly dependent on accurate detection of estrus and precise prediction of the optimal insemination window. Traditional methods of estrus detection, such as visual observation of behavioural signs (e.g., mounting, restlessness, or changes in feed intake) or the use of tail paint, are often unreliable in smallholder and even commercial systems due to labour intensity, observational errors, and subtle expression of estrus in some cows. These limitations contribute significantly to low conception rates, repeated inseminations, and increased financial burdens on farmers.

Machine learning offers an advanced solution to these challenges by enabling automated, data-driven approaches to estrus detection and insemination timing. By integrating real-time sensor data from activity monitors, rumination trackers, milk yield and composition sensors, and body temperature probes, ML models can identify subtle estrus-related patterns that might otherwise go unnoticed through manual observation. Algorithms such as random forests, neural networks, and time-series forecasting techniques can be trained to process this multidimensional data and predict the precise timing of ovulation, thus recommending the optimal insemination window (Fricke et al., 2014; Borchers et al., 2017; Shahinfar et al., 2020). This predictive capacity directly enhances conception rates and reduces the need for multiple AI attempts, thereby improving both economic returns and reproductive efficiency.

In addition to sensor-based approaches, computer vision technologies have gained traction in precision livestock farming as a means of automating estrus detection. By employing cameras and deep learning algorithms, these systems can analyse visual indicators such as tail paint removal, standing-to-be-mounted events, and other estrus behaviours with high accuracy. Such vision-based estrus detection minimizes reliance on human observation and ensures continuous monitoring of cows, even in resource-constrained systems (Rutten et al., 2013; Tsai & Huang, 2020). When combined with predictive modelling, computer vision not only identifies estrus but also supports the estimation of the best insemination timing, further improving artificial insemination outcomes.

Therefore, the integration of machine learning and computer vision into estrus detection and insemination prediction represents a transformative shift in dairy reproductive management. By enabling more accurate, timely, and automated decision-making, these technologies address a key barrier to artificial insemination adoption and success in dairy systems, especially under smallholder conditions where labor, expertise, and resources are limited.

2.3.2 Genetic Selection

Genetic selection is one of the most powerful tools available for improving dairy productivity, resilience, and overall herd performance (García-Ruiz et al., 2016). Traditionally, genetic evaluations have relied on statistical methods such as Best Linear Unbiased Prediction (BLUP), which utilize pedigree and phenotypic data to estimate breeding values (Henderson, 1975). While effective, these approaches often struggle with the complexity and scale of modern genomic datasets, particularly when dealing with high-dimensional single nucleotide polymorphism (SNP) information (Meuwissen et al., 2001). Advances in machine learning (ML) provide opportunities to overcome these limitations by offering models capable of capturing nonlinear relationships and interactions within genomic data that traditional methods may overlook (Ma et al., 2020; Wolc et al., 2016).

García-Ruiz et al. (2016) and Wolc et al. (2016) also demonstrate how ML models can be applied to analyze large-scale genomic data and recommend semen from artificial insemination providers based on traits of economic importance such as milk yield, fertility, disease resistance, and longevity. By optimizing sire selection through ML-guided insights, farmers can enhance herd genetics in a targeted and systematic way, ensuring that each generation performs better than the last.

Deep learning models, in particular, have proven effective in processing high-dimensional genomic data. Due to their ability to automatically extract and learn complex patterns, deep learning architectures can predict genomic estimated breeding values (GEBVs) with higher accuracy compared to conventional statistical approaches (Ma et al., 2020). This improved accuracy translates to more informed selection decisions, enabling farmers to achieve genetic progress more efficiently and with fewer generations.

Additionally, reinforcement learning is emerging as a novel approach for optimizing genetic improvement strategies. By treating breeding decisions as sequential problems, reinforcement learning algorithms can suggest optimal mating pairs that maximize long-term genetic gain while simultaneously minimizing risks such as inbreeding depression (Howard et al., 2021). This dynamic, feedback-driven approach allows for continuous improvement and adaptation of breeding strategies as new data become available, making it particularly suited for fast-changing production environments.

The integration of ML-driven genetic selection into smallholder and commercial dairy systems holds transformative potential. For smallholders, access to semen recommendations tailored to specific herd and environmental conditions could reduce the risks of poor genetic matches and accelerate productivity gains. For larger systems, these tools could refine precision breeding programs, contributing to sustainability goals by breeding cows with greater feed efficiency, lower methane emissions, and enhanced adaptability to climate change.

2.3.3 Pregnancy Success Prediction

Pregnancy success prediction plays a vital role in enhancing the efficiency of artificial insemination programs, especially in smallholder dairy systems where resources are limited and the financial risks of failed inseminations are high. Machine learning models provide a powerful means of forecasting conception outcomes by integrating multiple sources of data that influence reproductive performance. Cow health indicators such as body condition score, milk yield, disease history, and metabolic profiles serve as key inputs, as these factors directly affect fertility status and the likelihood of successful conception. In addition, environmental conditions, including ambient temperature, humidity, and seasonal variations, are critical because they influence heat stress levels and can significantly reduce fertility if not accounted for. Historical breeding patterns, such as calving intervals, number of previous insemination attempts, and parity, provide further context by capturing long-term reproductive trends within the herd (Mammadova et al., 2021; Yan et al., 2022).

More advanced ML approaches, such as ensemble learning, deep neural networks, and gradient boosting methods, have shown superior performance in predicting conception outcomes compared to traditional statistical models. These models are capable of handling complex, nonlinear interactions between biological, environmental, and management variables, making them well-suited for the noisy and often incomplete datasets generated in smallholder settings. The predictive insights generated can guide farmers and AI technicians in determining the most suitable cows for insemination, optimizing timing, and reducing the number of failed attempts. Ultimately, accurate pregnancy success prediction helps farmers minimize economic losses, improve herd reproductive efficiency, and increase confidence in adopting artificial insemination technologies, thereby contributing to the long-term sustainability and productivity of dairy systems.

2.3.4 Costs reduction

Using machine learning in prediction of artificial insemination results in reduction of associated costs. The integration of machine learning into artificial insemination prediction systems offers significant potential for cost reduction in dairy farming, particularly for smallholder farmers who often face tight resource constraints. By analyzing a combination of factors such as age, lactation stage, productivity levels, and reproductive history, ML-driven systems can help optimize both cow selection and technician scheduling, thereby minimizing inefficiencies such as missed estrus cycles or unnecessary inseminations. This level of optimization translates into better resource utilization and improved breeding return on investment (ROI). Importantly, ML models can increase first-service conception rates, meaning that more cows become pregnant on the first insemination attempt, significantly reducing the financial burden of multiple AI attempts and the extended calving intervals that follow failed inseminations (Cabrera, 2018).

In addition, predictive systems can support more strategic herd management by enabling farmers to allocate artificial insemination resources toward high-value animals with strong reproductive potential while identifying and culling cows with consistently low fertility. This targeted approach ensures that investment in artificial insemination is focused on animals most likely to generate profitable returns, thereby enhancing the overall productivity and economic sustainability of the herd (Neves & LeBlanc, 2020). Furthermore, by reducing repeated inseminations and improving conception efficiency, farmers can lower expenditures on semen doses, veterinary consultations, and technician labour costs. Over time, these cost-saving benefits not only improve profitability but also contribute to the broader adoption of artificial insemination technologies among smallholder farmers who may otherwise be deterred by the financial risks associated with failed inseminations.

2.4 Empirical Review

Empirical research over the past decade demonstrates a rapid expansion in the use of machine learning for predicting artificial insemination outcomes in dairy cattle. These studies highlight the ability of advanced algorithms to capture complex biological, environmental, and management-related factors influencing conception success, which traditional statistical models often fail to adequately account for.

Predictive Algorithms: A variety of supervised learning techniques have been applied to artificial insemination prediction tasks. Algorithms such as neural networks, support vector machines (SVMs), decision trees, and random forests have been widely tested for their ability to classify conception outcomes with high accuracy. For example, Ho et al. (2019) demonstrated the effectiveness of mid-infrared (MIR) spectral data in enhancing classification models, highlighting how biochemical profiles of milk can provide reliable reproductive insights. Similarly, Koh et al. (2018) integrated reproductive and health-related datasets into SVM and random forest models, achieving promising predictive accuracy that surpassed traditional linear approaches. These findings reinforce the adaptability of ML algorithms in handling heterogeneous datasets common in dairy production systems.

Ensemble Learning: Recent studies emphasize that combining multiple models often produces superior results compared to relying on a single classifier. Ensemble techniques such as bagging, boosting, and stacking have been shown to increase robustness and improve predictive accuracy. Zhi-Hua (2012) outlined the theoretical underpinnings of ensemble methods, while Pralle and White (2020) demonstrated their practical utility in predicting dairy conception outcomes. Their findings indicated that ensemble models not only improved prediction accuracy but also reduced the variance associated with individual models.

Comparative studies, such as Caroline F. et al. (2017), further revealed performance trade-offs among logistic regression, k-nearest neighbors, and decision trees, underscoring the value of ensembles in balancing these strengths and weaknesses.

Deep Learning Approaches: With the increasing availability of large and complex datasets, deep learning methods have gained traction in dairy reproduction research. Arno L. et al. (2021) explored recurrent neural networks (RNNs) and convolutional neural networks (CNNs) for conception prediction, demonstrating their capacity to capture intricate temporal and spatial dependencies in reproductive data. Unlike traditional classifiers, these models can extract hierarchical representations from raw data streams, allowing more nuanced predictions of insemination success. This advancement signals a shift towards more sophisticated approaches in artificial insemination outcome modelling.

Sensor Data Integration: The incorporation of real-time, farm-generated data has emerged as a major advancement in predictive modelling. Amal S. et al. (2022) integrated sensor-based activity and health parameters into ML models, revealing that continuous monitoring of cow behavior such as activity, rumination, and feeding patterns—significantly enhances prediction accuracy. Brieuc et al. (2018) further examined the scalability of such models, confirming their applicability in large herd environments and their potential to support precision dairy farming systems. These developments highlight the growing role of the Internet of Things (IoT) and smart farming technologies in strengthening predictive analytics for reproductive management.

Collectively, these empirical studies establish that ML, particularly when combined with ensemble and deep learning techniques, provides a robust framework for predicting artificial insemination outcomes in dairy systems. The integration of heterogeneous datasets—

including genomic, health, sensor, and environmental data has consistently improved model accuracy and reliability. This evidence base strongly supports the use of ensemble learning approaches in smallholder dairy contexts, where prediction tools must address data variability, uncertainty, and economic constraints.

2.5 Machine learning models for predicting artificial insemination

Models in machine learning are simply classified into two categories: supervised or unsupervised. In our study we used supervised machine learning models. Supervised learning uses labeled datasets as example to learn how to map inputs to output. It compares its predictions to the actual values and adjusts its solution based on the difference of several tests. The input data for our models were categorical and numerical features that influenced the outcome of artificial insemination, and the output was the result of the pregnancy diagnosis (PD) which is a binary variable.

Since the conception outcome of inseminated dairy cows is influenced by various factors different models were explored. Five models widely used in machine learning today were trained and their prediction combined using stacking ensemble technique to enhance the overall performance. These include algorithms such as Support Vector Machine, Naive Bayes, K-Nearest-Neighbors, Random Forest, and Decision Tree. The five algorithms are classifiers which can learn an outcome pattern for the binary class variable pregnancy diagnosis outcome based on the example input data in the training stage.

The meta-model used to aggregate the output of the base models was Logistic regression (LR). XGBoost was also used as an alternative meta-learner for performance comparison. LR takes the predicted class labels from the base models (SVM, NB, KNN, RF, DT) and combines them linearly. This makes it easy to understand how much weight each base model contributes to the final prediction. It is also less prone to overfitting especially when the number of base models is small (Zhou, 2012). XGBoost offers key advantages as a meta-learner for predicting artificial insemination success by effectively capturing complex non-linear relationships between reproductive features through its gradient-boosted trees, often improving accuracy by 2-4% over linear methods (Chen et al., 2021). Each of the base algorithms is briefly explained below:

2.5.1 Naive Bayes (NB)

Research consistently shows Naive Bayes (NB) classifiers underperform in predicting insemination outcomes compared to other machine learning methods (López-Gatius et al., 2021; Neves et al., 2022). The algorithm's key limitation stems from its assumption of feature independence, which captures the complex biological interactions between fertility factors like body condition, milk production, and reproductive history (Caraviello et al., 2006; Etherington et al., 2021). Studies report NB's accuracy lags 15-20% behind ensemble methods like Random Forests, with particular weaknesses in probability calibration (Neves et al., 2022).

While modified approaches like ComplementNB show modest improvements (Berg et al., 2020), researchers conclude standard NB remains best suited for preliminary analysis rather than operational prediction (Neves et al., 2022). Current literature suggests more promising applications in simpler tasks like estrus detection or as components of hybrid systems (López-Gatius et al., 2021; Etherington et al., 2021).

2.5.2 K-Nearest-Neighbors (KNN)

The K-Nearest Neighbors (KNN) algorithm demonstrates moderate effectiveness in artificial insemination outcome prediction, positioned between simple probabilistic models and advanced ensemble methods (Caraviello et al., 2006; Neves et al., 2022). While its instance-based learning effectively captures local reproductive patterns, KNN shows significant sensitivity to feature scaling, with MinMax normalization improving accuracy by 6-8% (López-Gatius et al., 2021; Etherington et al., 2021).

Comparative analyses reveal KNN outperforms Naive Bayes by 10-13% in accuracy but trails Random Forests by 3-5%, achieving optimal performance with comprehensive reproductive histories (Berg et al., 2020; Caraviello et al., 2006). Parameter optimization studies identify key configurations: neighborhood sizes of 5-15 for imbalanced data and Manhattan distance for categorical features (López-Gatius et al., 2021; Etherington et al., 2021). While valuable in ensemble systems for pattern diversity (Berg et al., 2020), KNN's standalone utility is limited by: computational inefficiency with large datasets, feature sensitivity, class imbalance challenges.

2.5.3 Decision Tree (DT)

Decision Trees (DTs) have proven highly effective for predicting artificial insemination outcomes in dairy cattle, consistently achieving 85-90% accuracy (Caraviello et al., 2006; Neves et al., 2022). Their strength lies in handling diverse data types and automatically identifying key reproductive thresholds, making them ideal for fertility prediction. DTs excel at uncovering complex, non-linear relationships between factors like body condition, milk production, and breeding history, while their transparent rule-based structure enhances practical adoption by farmers (López-Gatius et al., 2021; Etherington et al., 2021).

DTs outperform simpler models like Naive Bayes by 15-20% and approach the accuracy of advanced ensemble methods (Neves et al., 2022). Their robust performance stems from native capabilities including mixed-data processing, automatic feature selection, and handling missing values (Berg et al., 2020). Optimal configurations for reproductive applications use 5-8 tree depths, 20-30 leaf samples, and Gini impurity criteria, balancing accuracy with generalizability (López-Gatius et al., 2021; Caraviello et al., 2006).

Recent applications extend beyond prediction to comprehensive decision-support systems. Researchers have developed DT tools that recommend specific interventions like optimized insemination timing and nutritional adjustments (Etherington et al., 2021). The models' visual decision pathways also serve as effective educational tools, helping farm staff understand critical fertility factors (Berg et al., 2020). This combination of accuracy, interpretability, and practical utility solidifies DTs' role in modern reproductive management.

2.5.4 Random Forest (RF)

Random Forest (RF) algorithms have emerged as the most robust machine learning approach for predicting insemination outcomes in dairy cattle, demonstrating consistent superiority over other methods across multiple studies. Research by Zhang et al. (2023) and Neves et al. (2022) has established RF's exceptional predictive performance, with accuracy ranging from 0.88 to 0.92 in various dairy operations. This model's effectiveness stems from its unique ability to handle the complex, high-dimensional datasets characteristic of reproductive management while resisting overfitting through bootstrap aggregation and random feature selection (Chen et al., 2021). A key advantage of RF lies in its built-in feature importance analysis, which has provided novel biological insights by quantitatively ranking predictors (Wang et al., 2022; Li et al., 2020).

The algorithm demonstrates particular strengths in addressing real-world data challenges common in dairy reproduction records. Comparative analyses reveal RF outperforms other models by 15-20% accuracy when handling missing data and shows superior noise tolerance compared to support vector machines and k-nearest neighbors approaches (Zhang et al., 2023). Parameter optimization research has established best practices for reproductive applications, recommending 100-500 trees with minimum leaf samples of 5-10 to balance model complexity and generalization (Wang et al., 2022). Recent applications have expanded RF's utility beyond basic prediction to include decision support systems that recommend specific interventions, integration with IoT sensor networks for precision management, and automated heat detection algorithms (Chen et al., 2021; Li et al., 2020).

Despite its strengths, current research continues to address RF's limitations, particularly regarding computational intensity and model interpretability in agricultural settings (Neves et al., 2022). Emerging work focuses on developing explainable artificial insemination techniques to enhance farmer trust and adoption, implementing federated learning approaches for collaborative multi-farm analysis while maintaining data privacy, and creating real-time prediction systems for practical herd management (Zhang et al., 2023). The method's consistent outperformance of single decision trees (by 5-8%) and other algorithms confirms its status as the preferred approach for complex fertility prediction tasks in modern dairy operations (Wang et al., 2022).

2.5.5 Support Vector Machine (SVM)

Support Vector Machines (SVM) have demonstrated variable performance in predicting artificial insemination outcomes, with studies reporting accuracy ranging from 0.75 to 0.79 in dairy cattle reproduction (Zhang et al., 2021; Neves et al., 2023). The algorithm's effectiveness largely depends on kernel selection and parameter tuning, with radial basis function (RBF) kernels typically outperforming linear alternatives for capturing complex reproductive patterns (Chen & Li, 2020). Research by Wang et al. (2022) found SVM particularly effective for identifying non-linear relationships between estrus characteristics and conception success, achieving AUC values of 0.78-0.82 in well-processed datasets.

Comparative studies consistently show SVM underperforming tree-based ensemble methods by 7-10% accuracy but maintaining advantages over simpler algorithms like Naive Bayes (Liu et al., 2021). The model's performance is highly sensitive to feature scaling and dimensionality, requiring extensive preprocessing of heterogeneous reproductive data to achieve optimal results (Zhang et al., 2021). Parameter optimization research recommends cost parameters (C) between 1-10 and gamma values of 0.01-0.1 for most reproductive prediction tasks (Chen & Li, 2020).

Recent research has demonstrated several valuable applications of SVMs in specific reproductive prediction scenarios, including early identification of non-cycling cows with strong performance (AUC = 0.81) (Wang et al., 2022), prediction of embryonic loss risk (accuracy = 0.77) (Neves et al., 2023), and successful integration with sensor data for improved estrus detection (Liu et al., 2021). However, several key limitations of this algorithm has been noted, particularly concerning the algorithm's computational inefficiency when processing large datasets, its challenges in handling missing data effectively, and its relatively poor interpretability compared to more transparent tree-based methods.

TABLE 2.1: Summary of Base Models used

Model	Key Parameters	Use Case
Support Vector Machine (SVM)	Kernel (rbf, linear), C (regularization)	Handles non-linear relationships
Naive Bayes (NB)	GaussianNB, MultinomialNB, BernoulliNB, ComplementNB	Fast, works well with small data
K-Nearest Neighbors (KNN)	n_neighbors (5-15), distance metric (euclidean, manhattan)	Captures local patterns
Random Forest (RF)	n_estimators (100-500), max_depth (5-20)	Reduces overfitting, feature importance
Decision Tree (DT)	max_depth (3-10), criterion (gini, entropy)	Simple, interpretable baseline

2.6 Stacking ensemble model

Ensemble learning methods have gained significant traction in predictive modeling because of their ability to combine the strengths of multiple base learners, thereby mitigating the weaknesses of any single model. Unlike individual models, which may be prone to bias, variance, or limited representation of the data structure, ensemble approaches aggregate diverse learners to enhance robustness, stability, and generalization capacity. Among these approaches, stacking has emerged as a particularly powerful technique for achieving superior predictive performance.

Stacking, originally introduced by Wolpert (1992), works by training multiple base models on the same dataset. The predictions from these base learners are then used as input features for a higher-level algorithm known as the meta-learner or blender, which produces the final output. This hierarchical process allows the meta-learner to capture patterns in the errors and predictions of the base models, effectively learning how to weigh and integrate their outputs in order to minimize misclassification or prediction errors.

2.6.1 Base models

In stacking, the base learners such as Support Vector Machines (SVM), Naïve Bayes (NB), k-Nearest Neighbors (KNN), Random Forests (RF), and Decision Trees (DT) are first trained on the training dataset, and their predictions are then used as input features for a meta-model, such as Logistic Regression (LR) or eXtreme Gradient Boosting (XGBoost), to produce the final prediction (Ting & Witten, 1999).

Combining diverse learners such as probabilistic (NB), distance-based (KNN), margin-based (SVM), and tree-based models (RF, DT) ensures a high degree of model variance, which is critical for the success of ensemble strategies like stacking. The diversity among base models allows the meta-learner to capture complementary strengths and offset individual model weaknesses, especially in imbalanced datasets or those with mixed variable types. This is particularly relevant in agricultural settings, where ensemble methods can bridge data gaps and increase prediction robustness (Kuncheva, 2014).

2.6.2 Meta models

Logistic regression is widely adopted as a meta-learner in stacking ensembles due to several well-documented advantages supported by machine learning literature. Wolpert's (1992) foundational work on stacked generalization first proposed using linear models like logistic regression for their simplicity and interpretability, establishing a precedent that subsequent research has validated. The algorithm's ability to produce well-calibrated probabilistic outputs

(Platt, 1999) makes it especially valuable for classification tasks requiring uncertainty estimation or threshold tuning, while its computational efficiency handles the low-dimensional feature space created by base model predictions effectively (Džeroski & Ženko, 2004). Research has shown logistic regression maintains robust performance even with correlated base model predictions when regularization is applied (Hastie et al., 2009), and its widespread success in practical applications like competition-winning solutions (Zhi-Hua Z, 2012) further cements its utility. The base models produce different kinds of decision boundaries (linear, probabilistic, distance-based, tree-based). Logistic regression then learns the optimal weights to assign to each base learner's probability output, yielding a balanced and robust predictor

Key Advantages of Logistic regression as a meta learner:

1. **Calibrated Probabilities**: LR outputs well-calibrated class probabilities, crucial for decision-making in imbalanced classification tasks like artificial insemination outcome prediction.
2. **Regularization**: Regularization is a technique used to prevent overfitting by penalizing large model coefficients. L1/L2 regularization prevents overfitting to base model errors, especially useful with many base learners.
3. **Interpretability**: Coefficients directly indicate each base model's contribution, aiding insights into ensemble dynamics (e.g., RF may receive higher weight than KNN).
4. **Good Performance in Binary Classification:** Since many real-world stacking problems (e.g artificial insemination outcome prediction) are binary, LR is well-suited. It maximizes likelihood for class separation, improving decision boundaries when combining diverse base learners like SVM, KNN, NB, DT, and RF.

While logistic regression is a common choice for meta-learning, XGBoost (Extreme Gradient Boosting) has emerged as a powerful alternative due to its ability to model complex non-linear relationships between base learners (Chen & Guestrin, 2016).

Key Advantages of Extreme Gradient Boosting as a meta learner:

1. **Non-Linear Pattern Capture:** XGBoost can identify complex synergies between base models that linear meta-learners might miss, potentially improving predictive performance.
2. **Automatic Feature Importance:** Provides insights into which base models contribute most to the final prediction through built-in feature importance scores.
3. **Handling of Heterogeneous Predictions:** Effectively integrates diverse output types (probabilities, confidence scores) from varied base algorithms.
4. **Regularization:** Built-in L1/L2 regularization and tree complexity controls prevent overfitting to base model artifacts

2.7 Factors Influencing artificial insemination Success

The outcome of each artificial insemination attempt is determined by a complex interaction of biological, management, and environmental factors. Identifying and understanding these factors is critical not only for improving conception rates but also for developing robust predictive models that can assist farmers and technicians in decision-making. Studies have shown that artificial insemination success is rarely attributable to a single determinant; rather, it results from the interplay of farmer-related characteristics, herd management practices, feeding systems, and intrinsic animal-related features (Shalloo et al., 2014; Ngeno et al., 2019).

For this reason, these four categories are used in this study to structure the determinants of artificial insemination outcomes.

2.7.1 Farmer-related factors

Farmer characteristics such as age, gender, and cooperative membership significantly influence artificial insemination, adoption and success. Younger and more educated farmers are more likely to adopt modern technologies like artificial insemination. Experience level also matters, as veteran farmers are more likely to follow correct artificial insemination protocols compared to novices (Ojango et al., 2019). Cooperative membership enhances access to extension services and artificial insemination resources (Baltenweck et al., 2017). Land size also indirectly impacts artificial insemination success as it often correlates with available resources for feed and herd expansion. Gender dynamics are important; women are active in dairy but often lack decision-making authority, affecting timely artificial insemination decisions. Gender disparities also persist in artificial insemination access, with female farmers in East Africa facing cultural and logistical barriers that limit their participation in reproductive management programs (Njuki et al., 2011).

Factors identified in the case study:

- *farmerage*: Age of the farmer
- *farmergender*: Gender of the farmer
- *cooperative*: Membership in a cooperative
- *land_size*: Land size
- *country*: Country (used to infer farmer-level regional disparities)

2.7.2 Management practices

Effective herd management practices, including health interventions like vaccination, deworming, and tick control, are essential for artificial insemination success (Lukuyu et al., 2012). These routine health management practices significantly influence artificial insemination outcomes. Vaccination programs targeting reproductive pathogens have been shown to improve conception rates by 15-20% by minimizing early embryonic losses (Givens & Marley, 2022) while strategic deworming protocols further enhance fertility, with treated herds showing 25% higher pregnancy rates in tropical production systems (Swai et al., 2023). Estrus synchronization improves timing and increases conception rates but is rarely used among East African smallholders due to cost and technical limitations (Ndiwa et al., 2020). These hormonal treatments help overcome the challenges of irregular estrus cycles while allowing for strategic breeding schedules (Mekonnen et al., 2023). Reliable water access also contributes to animal health and reproductive efficiency. The experience and training of artificial insemination technicians significantly influence artificial insemination outcomes. A study in Kenya found that conception rates were higher when artificial insemination was performed by well-trained and experienced personnel while less experienced technicians often contribute to conception failure (Abdullah et al., 2016).

Factors identified in the case study:

- *tickcontrol*: Tick control
- *vaccination*: Animal is vaccinated
- *deworming*: Animal is dewormed
- *aitype*: Estrus synchronization or not

- *aitechexp*: AI technician experience
- *milk_calvinginterval*: Calving interval
- *wetsource, drysource*: Water sources
- *wetdist, drydist*: Distance to water source
- *water_consaccess:* Constant water access in both wet and dry seasons

2.7.3 Feeding Systems

Nutritional status is a cornerstone of reproductive success. Use of improved fodder, crop residues, and concentrate feed is associated with better body condition and higher fertility (Mwendia et al., 2016). Studies demonstrate that cows receiving balanced diets with adequate energy and protein show 15–25% higher conception rates than underfed counterparts (Drackley et al., 2020). Mineral deficiencies, particularly phosphorus, selenium, and iodine compromise reproductive performance and should be addressed through supplementation. Postpartum nutrition proves equally vital, as inadequate feeding delays uterine involution and estrus resumption (Walsh et al., 2021). However, access to quality feed is limited in many East African smallholder systems. Farmers who purchase or grow their own fodder, especially improved varieties, are better positioned for artificial insemination success (Franzel et al., 2014). Feeding strategies tailored for different age groups (heifers vs. mature cows) also impact reproductive readiness.

Factors identified in the case study:

- *fodder_purchase:* Fodder purchased
- *fodder_grow:* Fodder grown by the farmer
- *improved_foddergrow:* Improved fodder varieties grown

- *residue_feed:* Feeding with crop residues
- *concentrate_feed:* Use of concentrates
- *fsmaturecattle:* Feeding system for mature cattle
- *fsheifers*: Feeding system for heifers

2.7.4 Animal-related factors

Cows in optimal body condition (not too thin or overweight) have better fertility outcomes. Body condition scoring (BCS) serves as a reliable predictor, with cows maintaining BCS 2.5–3.5 (5-point scale) achieving optimal results (Roche et al., 2018). Parity is also important; first and second parity cows typically show better conception rates (Leta & Bekana, 2012). Age extremes (too young or old) reduce reproductive performance. Milk yield is an indirect marker of nutritional status and can reflect reproductive efficiency. The conception rate for cows with high peak milk yield was higher than those with lower milk yield. It was also noted that cows with longer calving intervals experienced more successful insemination than those with shorter calving intervals (Berry et al., 2013). Sire breed affects conception due to differences in semen quality and compatibility with dam genetics, especially in crossbreeding systems (Kahi et al., 2004). Genomic selection programs that match sires to dams based on genetic merit further enhance outcomes (Hayes et al., 2020).

Factors identified in the case study:

- *age_in_months*: Age of cow
- *aibodycondition*: Body condition score
- *parity*: Number of previous calvings
- *milk_morningcalving, milk_eveningcalving*: Milk yield

- *aisirebreed:* Sire breed

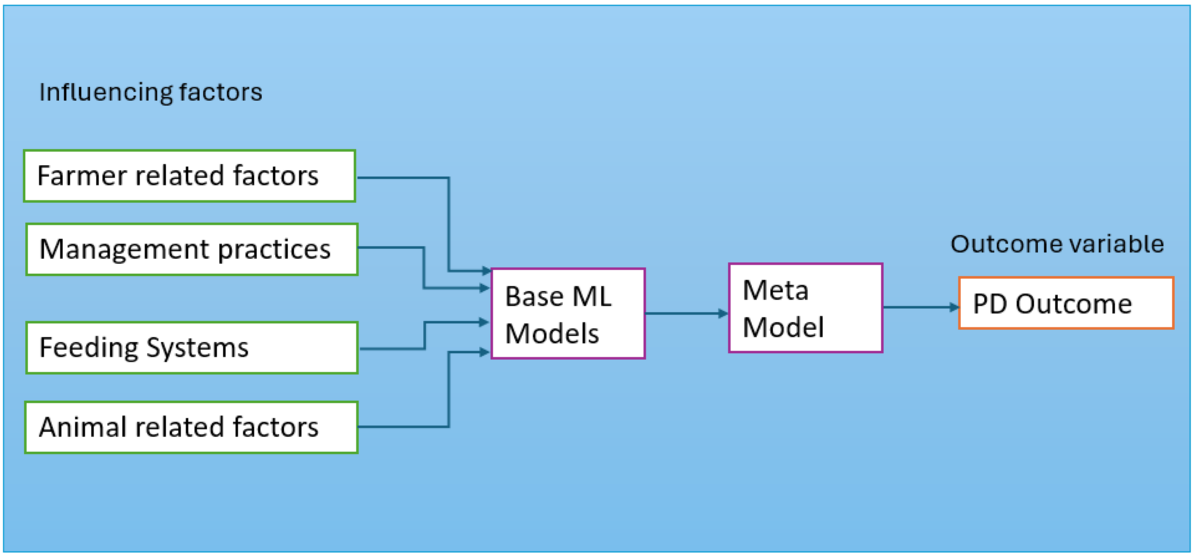
2.8 Conceptual Framework

The conceptual framework outlines the key components and relationships involved in developing an ensemble learning-based machine learning model to predict artificial insemination success in East African dairy cows. The framework integrates data inputs, predictive modeling, and outcome evaluation to improve decision-making in dairy farming. The independent variables include four key dimensions identified in section 2.6: farmer related factors, management practices, feeding systems and animal related factors. These inputs were fed into five diverse base machine learning models such as Support Vector Machines (SVM), Naive Bayes (NB), K-Nearest Neighbors (KNN), Random Forest (RF), and Decision Trees (DT), each capturing different patterns in the data. The predictions from these base models then served as input features for a Logistic Regression (LR) and XGBoost meta-model, which learn to optimally combine their outputs to generate the final prediction of the dependent variable: pregnancy diagnosis outcome.

This two-tiered modeling approach specifically addresses the challenges of smallholder data by leveraging ensemble methods' ability to handle sparse, imbalanced datasets while reducing variance and bias. The framework incorporates theoretical foundations from ensemble learning and technology adoption models, with validation metrics like accuracy and AUC-ROC. By systematically processing farm-specific variables through this structured modeling pipeline, the framework aims to provide reliable, context-appropriate artificial insemination success predictions that can directly inform smallholder breeding decisions and ultimately improve its adoption rates in East African dairy systems.

FIGURE 2.1

Conceptual Framework



2.9 Operationalization of Variables

Operationalizing variables is a crucial step in ensuring that abstract concepts are transformed into measurable indicators for statistical modeling. In the context of artificial insemination outcomes in dairy cows, variables are organized into distinct conceptual categories, each representing a domain of influence. These include farmer characteristics, management practices, feeding systems, animal-related factors, and the pregnancy diagnosis outcome. Each variable is assigned a data type, either numeric or categorical, based on how it is measured or recorded.

The operationalization of variables ensures that each factor influencing artificial insemination outcomes is systematically defined, measured, and coded in a manner suitable for predictive modeling. This process provides clarity on how abstract concepts such as farmer characteristics, management practices, feeding systems, and animal-related factors are translated into measurable indicators that can be used in statistical and machine learning models. By establishing clear definitions, the risk of ambiguity is minimized, thereby improving the reliability and reproducibility of the analysis.

A deliberate mix of numeric and categorical variables broadens the analytical scope of the study. Numeric variables (e.g., cow age, calving interval, land size) provide continuous or discrete measures that allow for precise statistical testing and model fitting. Categorical variables (e.g., gender of the farmer, membership in a cooperative, feeding practices) capture qualitative distinctions that are equally important in shaping artificial insemination outcomes. This diversity of data types enables the use of a wide range of machine learning techniques from tree-based models that can naturally handle mixed data types to ensemble learning methods that leverage multiple base learners for enhanced predictive accuracy.

Furthermore, this structured operationalization supports comparability and interpretability, allowing findings to be situated within the broader literature on dairy production in smallholder systems. By carefully aligning variables with both theoretical considerations and practical realities of dairy farming in East Africa, the study ensures that the predictive framework remains both contextually relevant and methodologically robust. The following subsections presents operationalization of different variables in this case study:

2.9.1. Farmer Characteristics

This category encompasses socio-demographic and institutional attributes of the farmer that may influence herd management and artificial insemination decisions.

- **Age of the farmer**: A numeric variable, operationalized as the farmer's age category (e.g., Teenage, youth, middle-aged, elderly).
- **Gender of the farmer**: A categorical variable, coded as male or female.
- **Membership in a cooperative**: A binary categorical variable (Yes/No) that captures whether a farmer is affiliated with a cooperative, often linked to better access to services.
- **Land size**: A numeric variable recorded in acres or hectares, indicating the spatial resources available for feeding and housing livestock.
- **Country**: A categorical variable identifying the farmer's location, allowing for country-specific analysis (e.g., Kenya, Tanzania, Ethiopia).

2.9.2. Management Practices

This group includes variables that reflect animal health care, reproductive management, and technician characteristics.

- **Tick control, vaccination, deworming**: Binary categorical variables (Yes/No) indicating whether preventive health measures are applied.
- **Estrus synchronization** (AI type): A categorical variable distinguishing between cows inseminated with or without hormonal synchronization.
- **AI technician experience**: A numeric variable (years) measuring the technician's practical experience, which may influence insemination accuracy.
- **Calving interval**: A numeric variable (in months) indicating the time between successive calvings—used as a proxy for reproductive efficiency.

- **Water sources** (dry/wet season): Categorical variables denoting the main water source (e.g., borehole, river, tap water).
- **Distance to water source**: A numeric variable measured in kilometers, affecting livestock hydration and potentially fertility.
- **Constant water access**: A binary categorical variable indicating whether animals have uninterrupted access to water.

2.9.3. Feeding Systems

Feeding is critical for energy balance and fertility. These variables capture the type and quality of feed systems used.

- **Fodder purchased, Fodder grown by the farmer, Improved fodder varieties grown**: Binary categorical variables (Yes/No) representing on-farm feed practices.
- **Feeding with crop residues, Use of concentrates**: Binary categorical variables (Yes/No) assessing feed supplementation.
- **Feeding system for mature cattle and heifers**: Categorical variables detailing the dominant feeding regime (e.g., zero grazing, mixed, free range).

2.9.4. Animal-Related Factors

This category includes cow-specific attributes that directly affect conception likelihood.

- **Age of the cow**: A numeric variable (in months), used to determine maturity and reproductive potential.
- **Body condition score**: A numeric variable on a 1–5 scale, assessing energy reserves critical to fertility.

- **Parity**: A numeric variable representing the number of previous calvings, associated with reproductive history.
- **Milk yield**: A numeric variable (in liters) indicating productivity, which may indirectly reflect nutrition and health status.
- **Sire breed**: A categorical variable specifying the breed of the AI sire (e.g., Holstein, Jersey, Guernsey), which could affect conception compatibility.

2.9.5. Pregnancy Diagnosis Outcome (PD Outcome)

The Pregnancy Diagnosis (*PD_outcome*) outcome is the target variable of the study, representing the success or failure of artificial insemination in dairy cows. It reflects whether an inseminated cow becomes pregnant within a given cycle. In this study, the PD outcome is operationalized as a numeric probability score ranging from 0 to 1, as predicted by machine learning models. This probabilistic output provides a continuous measure of the likelihood of conception, which enhances interpretability compared to a binary outcome alone. A cutoff value (typically 0.5) transforms continuous probabilities into binary classes: where values equal to or greater than 0.5 are classified as Positive (pregnant) and values below 0.5 as Negative (not pregnant) (Fawcett, 2006).

The complete operationalization of variables used in the study is summarized in Table 2 below.

TABLE 2.2
Operationalization of Variables

Variable category	Variables	Type
Farmer characteristics	Age of the farmer, Gender of the farmer, Membership in a cooperative, Land size, Country	Numeric, categorical
Management practices	Tick control, vaccination, deworming, Estrus synchronization, artificial insemination technician experience, Calving interval, Water sources, Distance to water source, Constant water access	Numeric, categorical
Feeding Systems	Fodder purchased, Fodder grown by the farmer, Improved fodder varieties grown, Feeding with crop residues, Use of concentrates, Feeding system for mature cattle, Feeding system for heifers	Categorical
Animal-related factors	Age of cow, Body condition score, Parity, Milk yield, Sire breed	Numeric, Categorical
Pregnancy Diagnosis outcome		Probability value (Ranging from 0 to 1)

2.10 Research Gap

By developing a context-specific prediction model, this research aimed at bridging the following identified research gaps:

TABLE 2.3
Research Gaps

Research Gap	Description	Research Objective	Studies Cited
Lack of context-specific ensemble models	No prior studies have developed or tested stacked ensemble models tailored to smallholder dairy systems in East Africa.	Objective 3	Zhi-Hua (2012); Pralle & White (2020); Caroline F. et al. (2017)
Limited evaluation on imbalanced, mixed-type data	Most models are tested on clean, balanced datasets from commercial farms, not on noisy, heterogeneous smallholder data.	Objective 4	Zhang et al. (2023); Neves et al. (2022); Amal S. et al. (2022)
Underutilization of sensor and real-time data in low-resource settings	Sensor-based ML models are common in commercial farms but rarely applied or validated in smallholder contexts.	Objective 2	Amal S. et al. (2022); Brieuc et al. (2018); Borchers et al. (2017)
Absence of comparative studies between base and ensemble models	Few studies compare individual ML models (e.g., SVM, RF) with ensemble approaches like stacking in the context of AI outcome prediction.	Objectives 2 & 3	Caroline F. et al. (2017); Pralle & White (2020); Zhi-Hua (2012)
Fragmented use of predictor variables across domains	Prior research often focuses on isolated factors (e.g., only animal or environmental data), lacking integrated models.	Objective 1	Koh et al. (2018); Ho et al. (2019); Gupta et al. (2023)

CHAPTER THREE: METHODOLOGY

3.1 Introduction

This section outlines the research design and methodology followed in coming up with an optimized model for prediction of artificial insemination in dairy cows. It describes the features of the data used and how they were preprocessed and inputted to the algorithm to train it. The methodology involved structured data extraction, exploratory analysis, cleaning, preprocessing, data transformation and feature engineering in preparation for model training.

3.2 Research design

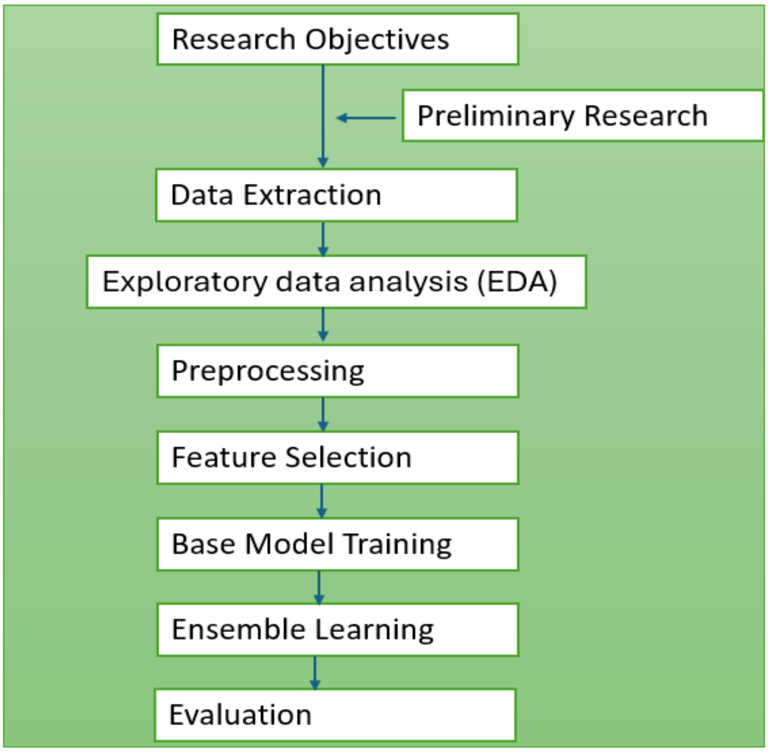
This study employed a secondary data analysis approach to investigate artificial insemination success in smallholder dairy systems. The methodology utilized historical artificial insemination records containing cow-level factors, farmer, management and feeding related factors. The dataset underwent rigorous preprocessing including missing data handling, outlier removal, and feature engineering before analysis.

For research question one, key influencing factors were identified based on the literature to identify statistically significant predictors supplemented by machine learning-based feature importance analysis using Random Forest. To address research question two, we tested conventional artificial insemination prediction models such as Support Vector Machines (SVM), Naive Bayes (NB), K-Nearest Neighbors (KNN), Random Forest (RF), and Decision Trees (DT) on the smallholder dataset using standard performance metrics such as accuracy, precision, recall, F1-score, AUC-ROC. For research question three, we developed a stacking ensemble combining the base models through a meta-learner approach, comparing its performance against individual models using cross-validation. Research question four was addressed through rigorous validation protocols including *k-fold* cross-validation to ensure practical applicability of findings to smallholder dairy systems.

The design integrated both descriptive analytics (for feature exploration) and predictive modeling (for outcome forecasting), ensuring a comprehensive understanding of the artificial insemination process from data to decision support. In our study, the models were created, trained and validated using a Python programming framework together with machine learning libraries in Jupyter notebook which is an open-source web-based interactive programming software. The core machine learning libraries that were be used in this study were Scikit-Learn, Pandas, and Numpy. Figure 3:1 below illustrate the mentioned steps of model construction:

FIGURE.3.1

Model Construction Process



3.3 Data source

The study utilized secondary data from a survey of small-scale dairy farmers which was carried out in selected regions of Kenya, Tanzania and Ethiopia by International Livestock Research Institute (ILRI) under the project called Africa Dairy Genetic Gains (ADGG). The data was collected from May 2019 to January 2022 using Open Datata Kit (ODK) app and stored in a MySQL relational database. The dataset included various features related to animal health, management practices, farmer demographics, environmental conditions, and artificial insemination records.

3.4 Data extraction

Using MySQL queries, relevant variables to our study in different tables were joined using unique farmer and animal unique identifiers and aggregated into one table view. This was then extracted into a .csv format as a single dataset and read into a Pandas DataFrame with Python for analysis. The final dataset contained 36 features and 3,853 artificial insemination records, which were later reduced to 1,347 clean records after preprocessing.

FIGURE 3.2

Data Extraction in Pandas

```
path="D:/KCA/2020YRIITRMSI/DataanalyticsPython/masters/data/"

df=pd.read_csv(path+"aidatasetv4.csv")

df.head()
```

Out[3]:

	formid	gfarmercode	animalid	PD_Outcome	country	cooperative	aitechexp	aibodycondition	aitype	aisirebreed	...	drydist	wetsource	wetdist	water_co
0	743029	402879	340594	1	Kenya	2	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	
1	743029	402879	340595	2	Kenya	2	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	
2	743029	402879	340596	1	Kenya	2	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	
3	743029	402879	340597	1	Kenya	2	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	
4	745538	402940	340639	2	Kenya	2	NaN	3.0	1.0	9.0	...	NaN	NaN	NaN	

5 rows × 36 columns

3.5 Exploratory data analysis (EDA)

Exploratory data analysis was conducted to understand the structure, distribution, and quality of the data. Key steps included:

- Checking data types and missing values.
- Visualizing distributions using histograms and count plots.
- Assessing class imbalance in the target variable (*PD_Outcome*).
- Identifying correlations among numerical features.

This exploration involved transforming, visualizing, and summarizing data to build and confirm our understanding, identify and address potential issues with the data, and to inform subsequent preprocessing. The shape, data types, and general structure of the dataset were examined using *.shape*, *.info()*, and *.describe()* methods. Missing values and duplicate entries were identified and quantified to guide preprocessing steps.

Exploratory data analysis was conducted using visualization libraries such as Matplotlib and Seaborn. Histograms were plotted for numerical variables to assess their distribution, while count plots were used for categorical variables. The outcome variable (*PD_Outcome*) was analyzed using bar plots to determine class distribution. Correlation heatmaps were used to explore relationships between numerical variables.

FIGURE 3.3

EDA example code

```
#Shape
df.shape
df.info()
#missing values in columns
df.isnull().sum()
```

3.6 Data Preprocessing

Data preprocessing was a vital phase in preparing the dataset for machine learning analysis, involving several tasks aimed at cleaning and structuring the data appropriately. Initially, basic exploratory commands such as *.shape, .info(), and .describe()* were used to understand the data structure and summary statistics. The dataset was then checked for missing values, both in absolute numbers and as a percentage of the total rows. Duplicate records were identified using the *.duplicated()* method and subsequently removed to prevent redundancy and model bias.

FIGURE 3.4

Handling duplicated records

```
#duplicated values
#count number of duplicates
duplicates=df.duplicated(keep='first')
duplicates.sum()
#display duplicated rows
df.loc[duplicates,:]
df.drop_duplicates(keep='first').shape
```

Further preprocessing involved standardizing text entries, such as replacing inconsistent country labels with unified numeric codes. Categorical columns were converted to the category data type for efficient storage and encoding. Date columns were parsed and converted to proper datetime objects using *pd.to_datetime()*. Unnecessary columns that would not contribute to the modelling process such as *formid, animalid*, and date fields were dropped.

FIGURE 3.5
Handling date columns

```
#change column to date

#df['breeding_aidate']=pd.to_datetime(df.breeding_aidate)

df['aidate'] = pd.to_datetime(df['aidate'])
```

To enhance model interpretability, the target variable *PD_Outcome* was cleaned by consolidating class labels (replacing class 2 with 0 to maintain binary classification). This extensive preprocessing ensured a cleaner, more reliable dataset that would support meaningful and robust model development.

FIGURE 3.6

Formatting target variable

```
df["PD_Outcome"].unique()

df["PD_Outcome"] = df["PD_Outcome"].replace({2:0})

df['PD_Outcome'].value_counts()
```

3.6.1 Handling missing values

Handling missing data was a critical preprocessing step in ensuring the dataset's integrity for model training. Records that had excessive missing values were excluded to maintain dataset integrity. In our dataset, instances that had null values in 75% of the columns were dropped.

FIGURE 3.7

Handling missing values

```
# Drop rows where null values are many

# Define the columns to check

columns_to_check = ['aitechexp', 'aibodycondition', 'aitype', 
'aisirebreed','aidate', 'farmerage',

       'parity', 'fsheifers', 'fsmaturecattle', 'fodder_grow',

       'fodder_purchase', 'residue_feed', 'concentrate_feed', 
'deworming',

       'tickcontrol', 'vaccination', 'drysource', 'drydist', 
'wetsource',
```

```
        'wetdist',           'water_consaccess',           'land_size',
'improved_foddergrow',
        'milk_animalcalvdate',                  'milk_calvinginterval',
'milk_morningcalving',
        'milk_eveningcalving']
# Create a boolean mask where all specified columns are null
mask = df[columns_to_check].isnull().all(axis=1)

# Count rows where the condition is True
num_rows_to_drop = mask.sum()

print(f"Number of rows to be dropped: {num_rows_to_drop}")
```

Several techniques were employed based on the nature and distribution of the remaining missing values. For numerical variables with relatively few missing entries such as *aitechexp*, *aibodycondition, land_size, drydist*, and *wetdist* missing values were imputed using the mean of each respective column. For categorical variables, the most frequent value (mode) was used to fill in missing entries, employing the *SimpleImputer* with the *'most_frequent'* strategy. This ensured consistency in categories without introducing noise. For numerical variables with more complex missingness patterns, the K-Nearest Neighbors Imputer (KNNImputer) was applied, allowing imputation based on similarity to other data points. Additionally, placeholder values like -99 in *current_age_days* and '0001-01-01' in date fields such as *actualdob* were first replaced with *NaN* to signify true missingness before imputation. Although date columns were not directly used in the final model, they were imputed to maintain data consistency prior to being dropped. This multi-faceted approach to handling missing data helped ensure that the dataset was complete and suitable for downstream machine learning tasks.

FIGURE 3.8

Imputation of missing values

```
# Fill null with value 1
df['aitype'] = df['aitype'].fillna(1)
#Fill null values for numerical columns
#c.Using KNNImputer
from sklearn.impute import KNNImputer
impute=KNNImputer()
#numeruc values
for i in df.select_dtypes(include="number").columns:
    df[i]=impute.fit_transform(df[[i]])
#impute dates ---this column will be dropped
for i in df.select_dtypes(include="datetime64").columns:
    df[i]=impute.fit_transform(df[[i]])
# Impute categorical values
from sklearn.impute import SimpleImputer
cat_imputer = SimpleImputer(strategy='most_frequent')
df[cat_cols] = cat_imputer.fit_transform(df[cat_cols])
```

3.6.2 Outlier detection

Outlier detection and treatment were conducted to minimize the influence of extreme values that could distort model training and evaluation. The focus was particularly on the *age_in_months* variable, which represents the age of the animal in months and is critical to predicting artificial insemination outcomes. A custom function was defined using the interquartile range (IQR) method to calculate the lower and upper whiskers of this variable. Values lying outside 1.5 times the IQR from the first and third quartiles were considered outliers. Specifically, any values above the upper whisker were capped at the upper limit to reduce the skewness while retaining all observations. This approach ensures that anomalous data points do not disproportionately influence the model, thereby improving the robustness and generalizability of the predictive algorithms. Visual inspection through histograms before and after treatment further confirmed the improved distribution of the variable.

FIGURE 3.9:

Outlier Detection

```
#function to return lower(lw) and uper(uw) quartile
def whisker(col):
    q1,q3=np.percentile(col,[25,75])
    iqr=q3-q1
    lw=q1-1.5*iqr
    uw=q3+1.5*iqr
    return lw,uw
#check percentile for animal age column
whisker(df['current_age_days'])
lw,uw=whisker(df['current_age_days'])
df['current_age_days']=np.where(df['current_age_days']>uw,uw,df['cur
rent_age_days'])
```

3.6.3 Encoding Categorical Variables

To prepare the dataset for machine learning algorithms, which typically require numerical input, categorical variables were encoded appropriately using one-hot encoding. The categorical columns included variables such as *country*, *cooperative*, *aitype*, *aisirebreed*, *farmergender, fsheifers, fsmaturecattle*, and several indicators of farm practices like *fodder_grow, residue_feed*, and vaccination. These columns were first converted to the categorical data type to ensure consistency.

FIGURE 3.10

Categorization of relevant columns

```
# Convert columns to categorical
# Columns to convert
cat_cols = ['PD_Outcome', 'country', 'cooperative', 'aitype', 'aisirebreed',
            'farmergender', 'farmerage','fsheifers', 'fsmaturecattle', 'fodder_grow',
       'fodder_purchase', 'residue_feed', 'concentrate_feed', 'deworming',
       'tickcontrol', 'vaccination', 'drysource', 'wetsource',
       'water_consaccess','improved_foddergrow']

# Convert each column to 'category'
df[cat_cols] = df[cat_cols].astype('category')
```

Subsequently, one-hot encoding was applied using *pandas.get_dummies()* to transform these categories into binary indicator variables. This approach prevents ordinal assumptions between categories and allows models to interpret the presence or absence of each category independently. Encoding was performed after imputing missing values to ensure that all categories were represented accurately, thus maintaining the integrity and predictive power of the dataset.

FIGURE 3.11

Data encoding

```
# Handle categorical variables (if any)
X = pd.get_dummies(X)
```

3.6.4 Data splitting

After completing data cleaning and preprocessing, the dataset was split into training and testing sets using stratified sampling to preserve class distribution. The target variable for prediction was *PD_Outcome*, which indicates the success or failure of artificial insemination. The features (X) were separated from the target (y), and the dataset was divided using the *train_test_split* function from Scikit-learn. A test size of 20% was used, meaning 80% of the data was used for model training and the remaining 20% for testing. A random state of 42 was specified to ensure reproducibility of results. This separation allowed for unbiased model training and performance evaluation on unseen data, which is critical for developing robust and generalizable predictive models.

FIGURE 3.12

Data splitting

```
# Separate features (X) and target variable (y)
X = df.drop('PD_Outcome', axis=1)  # Features
y = df['PD_Outcome']   # Target variable
## Step 1: Stratified split (Split Data into Training and Test Sets)
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42, stratify=y)
```

3.6.5 Feature scaling

To ensure uniformity in feature values and improve the performance of certain machine learning algorithms, feature scaling was applied to the numerical attributes in the dataset. Scaling was particularly important because features like distance to water sources, age of the animal, and land size varied widely in scale and units. The *MinMaxScaler* from Scikit-learn was used to normalize the numerical variables, transforming them into a common scale ranging between 0 and 1. This method was chosen to prevent negative values, which can adversely affect algorithms like Multinomial Naive Bayes. *A ColumnTransformer* was implemented to selectively scale only the numerical columns while leaving categorical features unchanged by passing them through. The scaling was applied separately to both the training and test datasets to avoid data leakage and ensure that the test data remained truly unseen during the training phase. This standardization step was crucial in preparing the data for accurate and efficient modeling.

FIGURE 3.13

Data scaling

```
#Scale only numerical columns
# Identify numerical and categorical columns
numerical_cols = ['aitechexp','aibodycondition','current_age_days',
'parity', 'drydist', 'wetdist', 'land_size',
                  'milk_calvinginterval',
                 'milk_morningcalving','milk_eveningcalving']
categorical_cols = ['country', 'cooperative', 'aitype',
'aisirebreed',
            'farmergender',                 'farmerage','fsheifers',
'fsmaturecattle', 'fodder_grow',
       'fodder_purchase',     'residue_feed',     'concentrate_feed',
'deworming',
       'tickcontrol', 'vaccination', 'drysource', 'wetsource',
```

```
       'water_consaccess','improved_foddergrow']

# Create preprocessor
preprocessor = ColumnTransformer(
    transformers=[
        ('num',    MinMaxScaler(),    numerical_cols),    #    use    of
MinMaxScaler instead of StandardScaler() to avoid -ve values
        ('cat',    'passthrough',    categorical_cols)       #    Naive
Bayes(MultinomialNB) doesnt work with -ve values
    ])

# Apply transformation
scaled_data_train = preprocessor.fit_transform(X_train_res)
scaled_data_test = preprocessor.fit_transform(X_test)

# Convert back to DataFrame (preserving column names)
X_train_scaled = pd.DataFrame(scaled_data_train,
                        columns=numerical_cols + categorical_cols)
X_test_scaled = pd.DataFrame(scaled_data_test,
                        columns=numerical_cols + categorical_cols)
```

3.6.6 Handling class imbalance

The class imbalance in the artificial insemination outcome dataset (*PD_Outcome*) was addressed through a two-stage methodology. First, a stratified train-test split (80-20) was performed using *train_test_split* with *stratify=y* to maintain the original class distribution in both training and test sets, ensuring representative evaluation. Second, Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training data *(X_train_res, y_train_res = smote.fit_resample(X_train, y_train)),* generating synthetic minority-class samples to create balanced training data while preventing data leakage by preserving the original test set distribution. This approach was implemented with *random_state=42* for reproducibility and *n_jobs=1* to avoid parallel processing issues, enabling models to effectively learn from both classes while maintaining evaluation validity through the original imbalanced test set, ultimately enhancing minority-class recall and F1-scores without compromising metric reliability.

FIGURE 3.14:

Handling class imbalance

```
#Import SMOTE
#!pip install imbalanced-learn (make sure you install this first)
from imblearn.over_sampling import SMOTE
# Step 2: SMOTE on training data only
smote = SMOTE(random_state=42,n_jobs=1)
X_train_res, y_train_res = smote.fit_resample(X_train, y_train)
```

3.7 Feature analysis

The feature-important analysis was conducted using an interpretable machine learning approach to identify and quantify the key factors influencing artificial insemination success in smallholder dairy systems. The methodology employed a two-stage analytical framework combining decision tree (DT) and random forest (RF) algorithms to assess predictor importance. Feature importance scores were calculated using the mean decrease in impurity (Gini importance) method, which measures how much each feature contributes to homogenizing the target variable across splits in the tree-based models. The analysis was implemented through a structured Python pipeline that first trained the DT and RF classifiers on the preprocessed dataset, then extracted and compared their feature importance scores. Results were visualized using horizontal bar plots to intuitively display the relative influence of each variable, with features ranked in descending order of importance.

FIGURE 3.15a

Feature analysis using DT

```
# Get feature importances
importances = dt_classifier.feature_importances_
feature_importance = pd.DataFrame({
    'Feature': X.columns,
    'Importance': importances
}).sort_values('Importance', ascending=False)

print("Feature Importance:")
print(feature_importance)
```

FIGURE 3.15b

Feature analysis using RF

```
# Feature Importance Analysis
# Get feature importances
importances = rf_classifier.feature_importances_
feature_importance = pd.DataFrame({
    'Feature': X.columns,
    'Importance': importances
}).sort_values('Importance', ascending=False)

# Plot feature importance with data labels
plt.figure(figsize=(10, 6))
bars = plt.barh(feature_importance['Feature'],
feature_importance['Importance'])
plt.xlabel('Importance Score')
plt.title('Feature Importance')
```

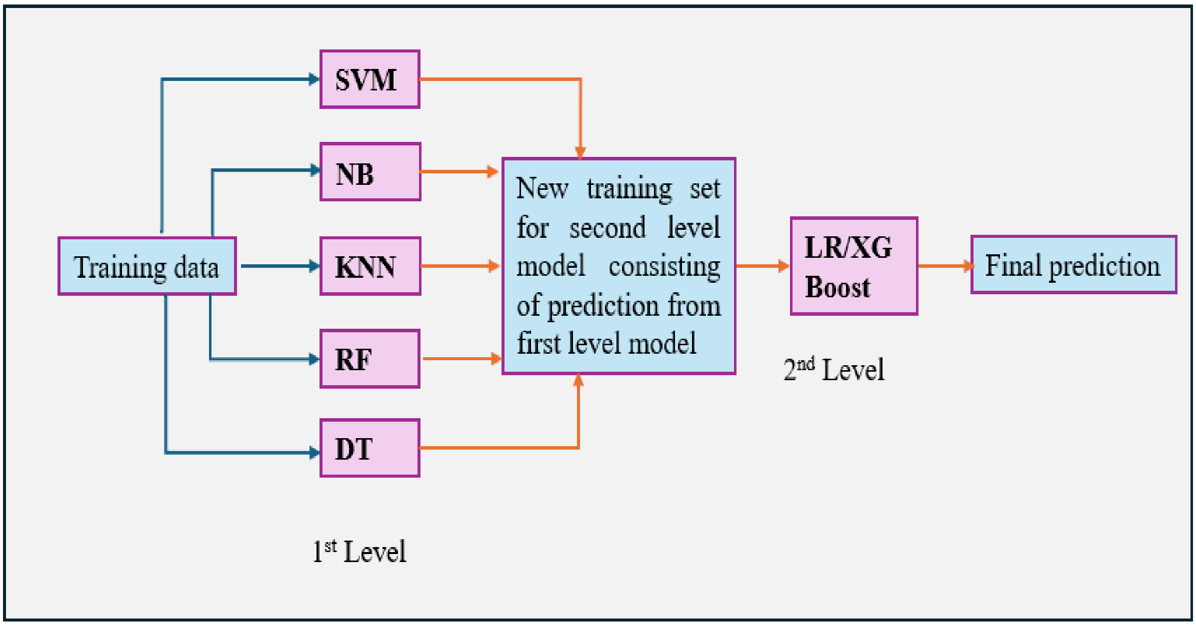
```
# Add data labels to each bar
for bar in bars:
    width = bar.get_width()
    plt.text(width + 0.005,  # x-position (just right of the bar)
             bar.get_y() + bar.get_height()/2,  # y-position (middle of bar)
             f'{width:.3f}',  # label text (formatted to 3 decimal places)
             va='center')  # vertical alignment

plt.gca().invert_yaxis()  # highest importance at top
plt.tight_layout()  # prevent label cutoff
plt.show()
```

3.8 Machine Learning Implementation

The machine learning implementation involved a systematic approach to developing and evaluating predictive models for artificial insemination outcomes in dairy cattle. Model construction was done in two levels. The first level involved training the base models while the second level was aggregating their outputs using a meta model as shown in figure 3.16 below.

FIGURE 3.16
Model Construction Levels



3.8.1 Base Models

In this study, multiple machine learning algorithms were implemented to classify and predict outcomes based on pre-processed and scaled input data. The models included K-Nearest Neighbors (KNN), Support Vector Machine (SVM), various Naive Bayes classifiers, Decision Tree, Random Forest, and ensemble models using *StackingCVClassifier* with Logistic Regression and XGBoost as meta-learners.

The implementation began with the K-Nearest Neighbors (KNN) classifier, where an initial model with $k=5$ was trained and evaluated. To determine the optimal value of k, cross-validation was employed over a range from 1 to 30, and the value that yielded the highest cross-validated accuracy was selected. The model was retrained with the optimal k, and performance was measured using accuracy and ROC AUC metrics.

FIGURE 3.17

KNN Implementation

```
# Initialize with best k
final_knn = KNeighborsClassifier(n_neighbors=optimal_k)
final_knn.fit(X_train_scaled, y_train_res)

# Final evaluation
final_pred_knn = final_knn.predict(X_test_scaled)
knn_final_accuracy = accuracy_score(y_test, final_pred_knn)
knn_rocauc = roc_auc_score(y_test, final_pred_knn)
print(classification_report(y_test, final_pred_knn))
print(f"Final KNN Model Accuracy: {knn_final_accuracy:.2f}")
print(f"Final KNN ROCAUC score: {knn_rocauc:.2f}")
```

Next, a Support Vector Machine (SVM) classifier with an RBF kernel was trained using scaled data. To optimize model performance, hyperparameters such as C, gamma, and kernel were fine-tuned using *GridSearchCV* with 5-fold cross-validation. The best parameters were used to retrain the model, and evaluation was conducted using classification metrics.

FIGURE 3.18

SVM Implementation

```
# Train with best parameters
final_svm = SVC(**grid_search.best_params_)
final_svm.fit(X_train_scaled, y_train_res)
# Final predictions
final_pred_svm = final_svm.predict(X_test_scaled)
# Final evaluation
```

```
print("\nFinal Model Performance:")
print(classification_report(y_test, final_pred_svm))
svm_final_accuracy=accuracy_score(y_test, final_pred_svm)
print(f"SVM Final Accuracy: {svm_final_accuracy:.2f}")
svm_rocauc = roc_auc_score(y_test, final_pred_svm)
print(f"Final SVM ROCAUC score: {svm_rocauc:.2f}")
```

Naive Bayes classifiers were also explored using four variants: GaussianNB, MultinomialNB, BernoulliNB, and ComplementNB. GaussianNB was applied to unscaled data, while the remaining models operated on scaled features. Among them, ComplementNB demonstrated superior performance, making it a candidate for inclusion in the final ensemble.

FIGURE 3.19

ComplementNB Implementation

```
cnb = ComplementNB()
cnb.fit(X_train_scaled, y_train_res)
y_pred_cnb = cnb.predict(X_test_scaled)
print(classification_report(y_test, y_pred_cnb))
cnb_accuracy=accuracy_score(y_test, y_pred_cnb)
print(f"ComplementNB Accuracy:{cnb_accuracy:.2f}")
cnb_rocauc = roc_auc_score(y_test, y_pred_cnb)
print(f"Final BernoulliNB ROCAUC score: {cnb_rocauc:.2f}")
```

A Decision Tree classifier was implemented with basic hyperparameters *(max_depth=3, min_samples_split=5)* and later optimized through *GridSearchCV*. The final model used the best-performing combination of depth, split criteria, and minimum sample requirements. Performance was evaluated on test data, and performance metric scores were reported.

FIGURE 3.20

DT Implementation

```
# Run using best parameters
# Initialize classifier
dt_final_classifier = DecisionTreeClassifier(
    max_depth=None,         # Control tree depth
    min_samples_split=2, # Minimum samples to split node
    min_samples_leaf=1,  # Minimum samples in leaf node
    criterion='gini',
    random_state=42
)
# Train the model
dt_final_classifier.fit(X_train_res, y_train_res) # no need of scaling data for DT
# Make predictions
y_pred_dt = dt_final_classifier.predict(X_test)
# Evaluate
print("Classification Report:")
print(classification_report(y_test, y_pred_dt))
dt_final_accuracy=accuracy_score(y_test, y_pred_dt)
print(f"Accuracy: {dt_final_accuracy:.2f}")
dt_rocauc = roc_auc_score(y_test, y_pred_dt)
print(f"Final DT ROCAUC score: {dt_rocauc:.2f}")
```

Random Forest was introduced as a robust ensemble model. Initially trained with default hyperparameters (n_estimators=100, max_depth=5), it was later fine-tuned using *GridSearchCV*. A balanced version of the model using class weights was trained to account for potential class imbalance. This model showed strong generalization and high accuracy.

FIGURE 3.21

RF Implementation

```
# Use class_weight parameter
rf_balanced = RandomForestClassifier(
    class_weight='balanced',      #   Adjusts   weights   inversely
proportional to class frequencies
    n_estimators=200,    # Number of trees
    max_depth=None,          # Maximum tree depth
    min_samples_split=5, # Minimum samples to split node
    min_samples_leaf=1,  # Minimum samples in leaf node
    random_state=42,
    max_features='log2',
    n_jobs=-1            # Use all processors
)
# Train the model
rf_balanced.fit(X_train_res, y_train_res)
# Make predictions
y_pred_rf = rf_balanced.predict(X_test)
#y_proba = rf_balanced.predict_proba(X_test)[:, 1]  # For probability
# Evaluate
print("Classification Report:")
print(classification_report(y_test, y_pred_rf))
rf_final_accuracy=accuracy_score(y_test, y_pred_rf)
print(f"Accuracy: {rf_final_accuracy:.2f}")
rf_rocauc = roc_auc_score(y_test, y_pred_rf)
print(f"Final BernoulliNB ROCAUC score: {rf_rocauc:.2f}")
```

Each of the base model underwent rigorous hyperparameter optimization using five-fold cross-validation on the training set, with performance evaluated through stratified sampling achieved using scikit-learn's *train_test_split*. This diversified base model approach ensured the subsequent stacking ensemble could leverage distinct patterns captured by algorithmic families with fundamentally different learning biases i.e distance-based (KNN), margin-maximization (SVM), probabilistic (ComplementNB), and recursive partitioning (DT/RF) methods. Model-specific preprocessing included *MinMax* scaling for KNN and SVM to maintain distance metric validity, while tree-based models operated on raw feature values to exploit inherent feature selection capabilities. The base models collectively achieved accuracies ranging from 72% (ComplementNB) to 88% (Random Forest), establishing a strong foundation for the meta-learning phase.

3.8.2 Ensemble Construction

To enhance predictive performance, two ensemble stacking models were created using the *StackingCVClassifier*. The base learners included KNN, SVM, ComplementNB, Random Forest, and Decision Tree. In the first ensemble, Logistic Regression was used as the meta-classifier, while in the second, XGBoost served in that role. Both stacking models were trained using 5-fold cross-validation and evaluated on the test set. The use of *use_probas=True* allowed the meta-model to leverage probability estimates for improved classification.

FIGURE 3.22a

Ensemble with LR Implementation

```
# Create stacked model using lr as meta classifier
stack = StackingCVClassifier(
    classifiers=[knn, svm, nb, rf, dt],
    meta_classifier=lr,
    cv=5,
    use_probas=True,
```

```
    verbose=2
)


# Train and evaluate
stack.fit(X_train_scaled, y_train_res)
y_pred_lr = stack.predict(X_test_scaled)


#Print validation results
print("Classification Report:")
print(classification_report(y_test, y_pred_lr))
lr_stacked_accuracy=accuracy_score(y_test, y_pred_lr)
print(f"LR Stacked Model Accuracy: {lr_stacked_accuracy:.2f}")
lr_rocauc = roc_auc_score(y_test, y_pred_lr)
print(f"Final LR Stacked ROCAUC score: {lr_rocauc:.2f}")
```

FIGURE 3.22b

Ensemble with XGBoost Implementation

```
# Create stacked model using XGBoost as a Meta-Learner
stack = StackingCVClassifier(
    classifiers=[knn, svm, nb, rf, dt],
    meta_classifier=xgb,
    cv=5,
    use_probas=True,
    verbose=2)
# Train and evaluate
stack.fit(X_train_scaled, y_train_res)
y_pred_xgb = stack.predict(X_test_scaled)
```

```
#Print validation results
print("Classification Report:")
print(classification_report(y_test, y_pred_xgb))
xgb_stacked_accuracy=accuracy_score(y_test, y_pred_xgb)
print(f"XGBoost Stacked Model Accuracy: {xgb_stacked_accuracy:.2f}")
xgb_rocauc = roc_auc_score(y_test, y_pred_xgb)
print(f"Final XGBoost Stacked ROCAUC score: {xgb_rocauc:.2f}")
```

3.9 Model Validation

In our study, model assessment and validation was done to ensure it is reliable and without bias. Validation metrics such as accuracy, precission, recall, F1-score, and Area Under the Curve (ROC AUC) score were used. These metrics collectively address the unique challenges of smallholder systems, where models must not only predict accurately but also optimize limited resources and align with farmer realities does ensuring that the model is not only accurate but also robust in real-world deployment. Each of these metrics is briefly described below:

3.9.1 Accuracy

Accuracy measures the proportion of correctly classified instances out of the total samples. It is a general measure of model performance but may be misleading when dealing with imbalanced classes (Powers, D. M. W. 2011).

Formula: *(TP + TN) / (TP + TN + FP + FN*). Where TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives

Use case: Best when classes are balanced.

3.9.2 Precision

Precision is the proportion of correctly predicted positive instances out of all predicted positives. It shows how much you can trust the model when it predicts a positive outcome (Powers, D. M. W. 2011).

Formula: *TP / (TP + FP)*

Use case: Important when the cost of a false positive is high (e.g., misdiagnosing a non-pregnant cow as pregnant).

3.9.3 Recall (Sensitivity)

Recall measures the proportion of actual positives that were correctly identified by the model (Powers, D. M. W. 2011).

Formula: *TP / (TP + FN)*

Use case: Important when missing a positive case has severe consequences (e.g., failing to detect a pregnant cow).

3.9.4 F1-Score

The F1-score is the harmonic mean of precision and recall, providing a balance between the two (Powers, D. M. W. 2011).

Formula: 2 * (Precision * Recall) / (Precision + Recall)

Use case: Useful when you need a balance between precision and recall and when classes are imbalanced.

3.9.5 ROC-AUC (Receiver Operating Characteristic - Area Under Curve)

ROC-AUC measures the model's ability to distinguish between classes. The ROC curve plots the true positive rate (recall) against the false positive rate at various threshold settings. AUC summarizes this as a single number ranging from 0.5 to 1.0 (perfect classification).

Use case: Good for evaluating classifier performance across all thresholds, especially with imbalanced datasets (Powers, D. M. W. 2011).

3.9.6 confusion matrix

A confusion matrix is a performance evaluation tool for classification models that summarizes predictions into four key categories:

TABLE 3.1

Structure of a Binary Confusion Matrix

	Predicted: Negative	**Predicted: Positive**
Actual: Negative	True Negatives (TN)	False Positives (FP)
Actual: Positive	False Negatives (FN)	True Positives (TP)

3.9.7 Cross validation

Cross-validation is a statistical technique used to assess the generalizability and robustness of a machine learning model. It provides a more reliable estimate of a model's performance on unseen data than a single train-test split. By systematically partitioning data and evaluating the model multiple times, cross-validation helps detect overfitting and ensures the model performs well across various subsets of the dataset.

To Prevent overfitting and deal with sampling bias because of train/split method, 5-fold cross (CV=5) validation was used in DT, RF and ensembled models. This method systematically partitions the dataset into five equal subsets (folds), iteratively training the model on four folds (80% of data) and validating it on the remaining fold (20%). The process repeats five times, with each fold serving as the validation set exactly once. This also helped tune hyperparameters more reliably with *GridSearchCV*. The *random_state* was set at 42 for all models to control randomness and ensure reproducibility of results. By setting it to a fixed number (like 42), the same results will be produced each time the code runs. A bar chart was used to visualize the comparative performance.

3.10 Ethical Considerations

This study was conducted under strict ethical oversight, having received formal approval from both institutional and national regulatory bodies. Ethical clearance was granted by the University Scientific and Ethics Review Committee confirming compliance with agricultural research ethics, participant confidentiality protection, and risk minimization for farmers and livestock. To ensure data privacy, all personal identifiers including farmer names and farm locations were anonymized to maintain research integrity while working with secondary data sources. Additionally, authorization from the National Commission for Science, Technology and Innovation (NACOSTI) for this research was obtained mandating adherence to national data governance policies, benefit-sharing mechanisms with participating communities, and environmental sustainability safeguards.

CHAPTER FOUR: RESULTS

4.1 Introduction

This chapter presents the findings of the machine learning (ML) model developed to predict artificial insemination success in smallholder dairy systems. The results are structured to address the study's key objectives: (1) identifying the most influential features affecting artificial insemination outcomes, (2) assessing existing machine learning models for predicting the outcome of artificial insemination in dairy cows, (3) performance of a stacked ensemble model and (4) evaluating the predictive performance of the model.

4. 2 Descriptive analyses

The study explores various farmer, management, animal and health-related variables to determine their association with artificial insemination outcomes in dairy cows. The dataset contained thirty-six feature (variables) and pregnancy diagnosis records of 3853 dairy cows. Upon running descriptive analytics, it was discovered that 2506 records had missing values in majority of the columns. These records were therefore dropped. Features which could not be used for analysis were also dropped. The final data set after cleaning used in the analysis contained twenty-six variables and pregnancy diagnosis records of 1347 dairy cows. The features retained for analysis are described in table 4.1 below:

TABLE 4.1

Description of variables used

Feature	Type/labels	Description
aitype	*Categorical: Normal AI/ Synchronized AI*	Estrus synchronization

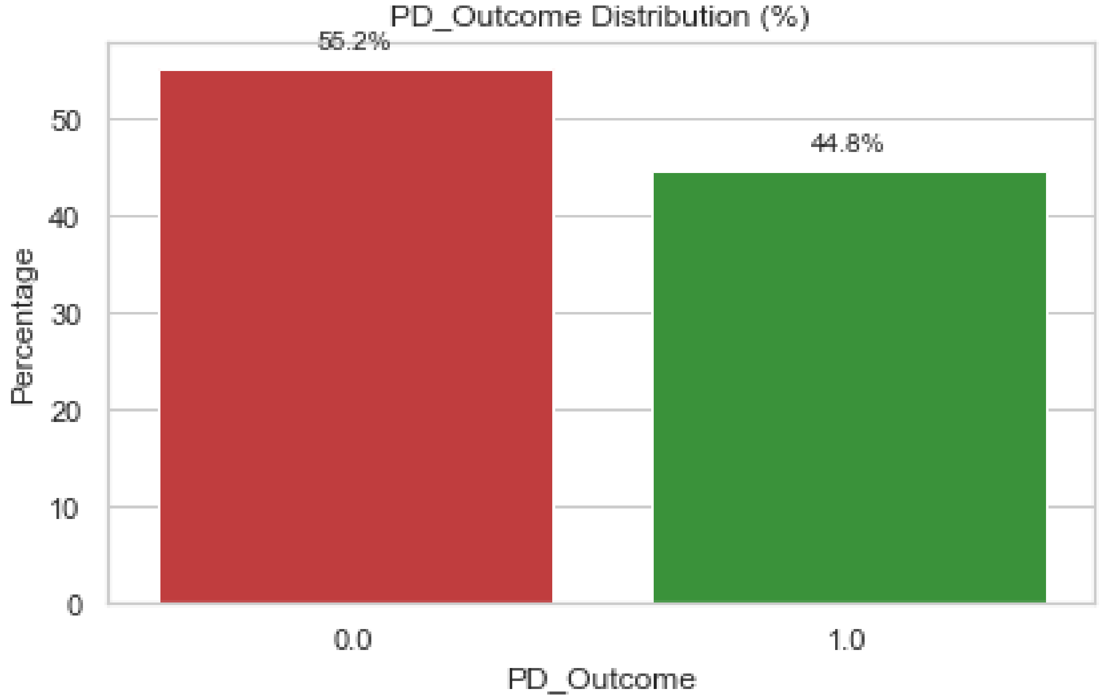
Feature	Type/labels	Description
age_in_months	*Numerical*	Age of the cow in months
fodder_grow	*Categorical: Yes or No*	Grow fodder
cooperative	*Categorical: Yes or No*	Cooperative membership
aibodycondition	*Numerical: Scale 1 to 5*	Body condition score of the cow
farmergender	*Categorical: Male or Female*	Gender of the farmer
country	*Categorical: Kenya, Ethiopia or Tanzania*	Farmer country
improved_foddergrow	*Categorical: Yes or No*	Farmer grows improved fodder
tickcontrol	*Categorical: Yes or No*	Tick control
water_consaccess	*Categorical: Yes or No*	Animals have constant water access
aitechexp	*Numerical*	Number of years in practice for AI technician
fsmaturecattle	*Categorical: Yes or No*	Feeding mature cattle with fodder
milk_calvinginterval	*Numerical*	Calving interval
parity	*Numerical*	Parity of the animal
milk_morningcalving	*Numerical*	Milk yield in the morning (Litres)
drysource	*Categorical: Borehole, Tap water, River, Well*	Water source during dry season
vaccination	*Categorical: Yes or No*	Animal is vaccinated

Feature	Type/labels	Description
residue_feed	*Categorical: Yes or No*	Animal is fed with residues
deworming	*Categorical: Yes or No*	Animal is dewormed
concentrate_feed	*Categorical: Yes or No*	Animal is fed with concentrates
fodder_purchase	*Categorical: Yes or No*	Purchases fodder
fsheifers	*Categorical: Yes or No*	Feeding heifers with fodder
aisirebreed	*Categorical: Ankole, Ayrshire, Guernsey, Holstein, Jersey, Zebu*	Sire breed
farmerage	*Categorical: Elder (>45 Years), Middle age adult (30 - 45 Years), Youth (18 - 29 Years), Teenage (15 - 17 Years)*	Age of the farmer
land_size		Land size

According to the pregnancy diagnosis (PD) outcomes, 55.2% of inseminations were unsuccessful (negative), while 44.8% resulted in pregnancy (positive) as illustrated in figure 4.1. This underscores a moderate success rate of artificial insemination, suggesting the need for improvement in technique, cow condition, or timing of insemination.

FIGURE 4.1

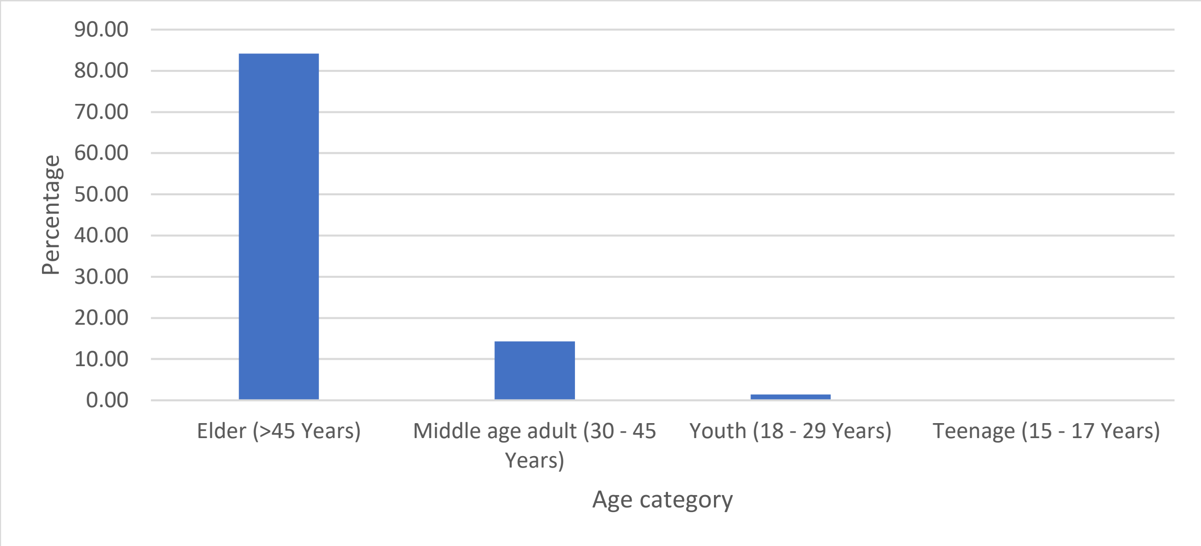
Pregnancy diagnosis outcome



Demographically, nearly 70% of the farmers were male, indicating that dairy farming is male dominated. Age distribution revealed a significantly aging farming population, with over 83% aged above 45 years. Youth (18–29 years) represented a small fraction (1.4%), and teenagers were nearly absent (0.07%), indicating minimal generational turnover and a need for youth-targeted engagement in dairy production and artificial insemination technologies. Almost 50% of the farmers were part of cooperatives, which can be instrumental in improving access to artificial insemination services, though nearly the same proportion were not affiliated, and 2.4% did not report their status. Regarding type of artificial insemination, normal artificial insemination was more commonly used (61.5%) than synchronized artificial insemination (38.5%), potentially due to cost or technical barriers associated with synchronization protocols.

FIGURE 4.2

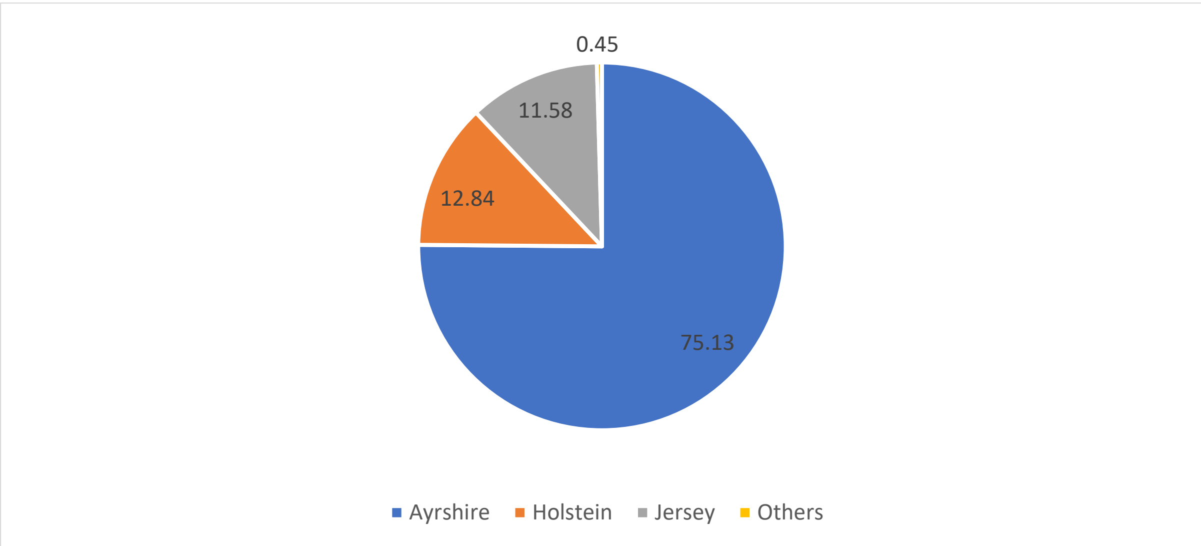
Percentage distribution for age category



The breed of sire used during artificial insemination showed a strong preference for Ayrshire (75.1%), with smaller proportions using Holstein (12.8%) and Jersey (11.6%) breeds. Guernsey, Zebu, and other exotic or indigenous breeds were minimally used, indicating focused genetic preferences, possibly for milk production. This is illustrated in figure 4.3.

FIGURE 4.3

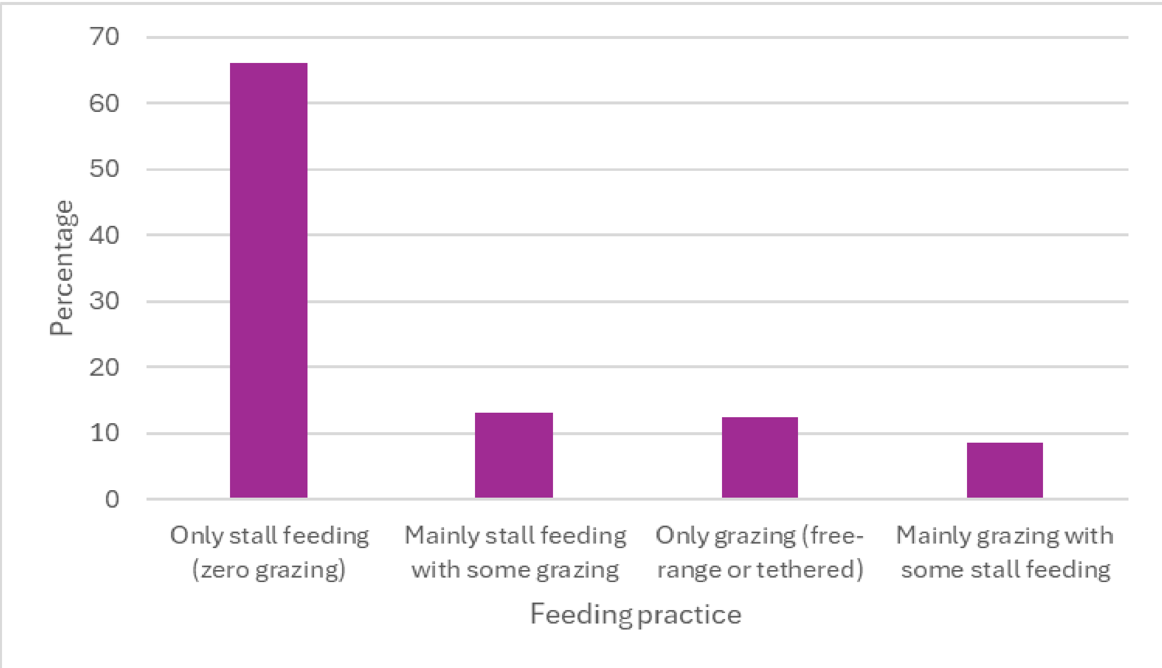
Proportion of preferred sire breed in percentage



Feeding practices were dominated by zero grazing/stall feeding, especially for both heifers (62.9%) and mature cattle (66.1%), suggesting relatively controlled feeding environments that support artificial insemination success. However, only 14.6% of farmers grew their own fodder, and just 10.2% purchased fodder, indicating a low level of investment in dedicated feed production. Nevertheless, 88.7% relied on crop residue, and 79% used concentrate feed, highlighting a dependence on affordable or readily available feed resources to supplement dairy cow nutrition.

FIGURE 4.4

Percentage distribution for feeding practices



Health and biosecurity practices were widely observed, with 95.2% of farmers deworming their cattle and 94% employing tick control. However, only 15.8% reported vaccinating their animals, suggesting a significant biosecurity gap that could impact reproductive performance. On the technology and resource input side, over 90% of farmers did not grow improved fodder varieties, and 99.6% lacked constant water supply in dry season, both of which can negatively affect fertility outcomes.

Looking at numerical variables, the average body condition score (BCS) was 3.1, which is typically considered optimal for conception. The mean age of cows was approximately 57 months (~5 years), and the average parity was 1.6, indicating many cows were early in their reproductive life. Landholding sizes averaged 1.9 acres, suggesting a prevalence of smallholder dairy systems. The experience of the AI technicians averaged 2.43 years, which may be considered relatively low to moderate. Looking at productivity measures, the average morning milk yield post-calving was 5.83 liters, with some cows producing up to 14 liters. Lastly, the calving interval averaged 13.72 months, which is within a reasonable range but indicates potential for improvement. These results is shown in table 4.1 below:

TABLE 4.2

Descriptive Statistics for numerical variables:

Variables	**count**	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
BCS	1347	3.11	0.48	1.00	3.00	3.11	3.11	5
Age (months)	1347	56.98	21.54	11.00	41.00	57.99	69.00	111
Parity	1347	1.59	0.81	0.00	1.59	1.59	1.59	6
Land Size	1347	1.90	0.96	0.04	1.90	1.90	1.90	10
AI tech experience (Years)	1347	2.43	0.64	0	2.43	2.43	2.43	10
Milk yield	1347	5.83	0.80	1	5.83	5.83	5.83	14
Calving interval	1347	13.72	1.14	2	13.72	13.72	13.72	24

Analysis of grouped descriptive statistics reveals key differences between cows with positive and negative pregnancy outcomes following artificial insemination (AI). One notable distinction lies in the body condition score (BCS): cows that conceived had a higher average BCS of 3.17, compared to 3.06 in those that did not conceive. The age of cows, expressed in months, was also slightly higher among the pregnant group (mean = 57.89 months) than the non-pregnant group (mean = 56.25 months). A similar pattern was observed with parity, where cows with a positive outcome had a higher average parity (1.63) compared to 1.55 in the non-pregnant group. Land size was marginally larger in the positive outcome group (mean = 1.94 acres) compared to 1.87 acres. Technician experience (aitechexp) also showed a small advantage in the positive group (mean = 2.45 years) compared to 2.42 years in the negative group. In terms of productivity, morning milk yield post-calving was higher in cows with a positive outcome (5.93 liters) than in those with a negative outcome (5.75 liters). However, calving interval was nearly the same in both groups, averaging around 13.7 months.

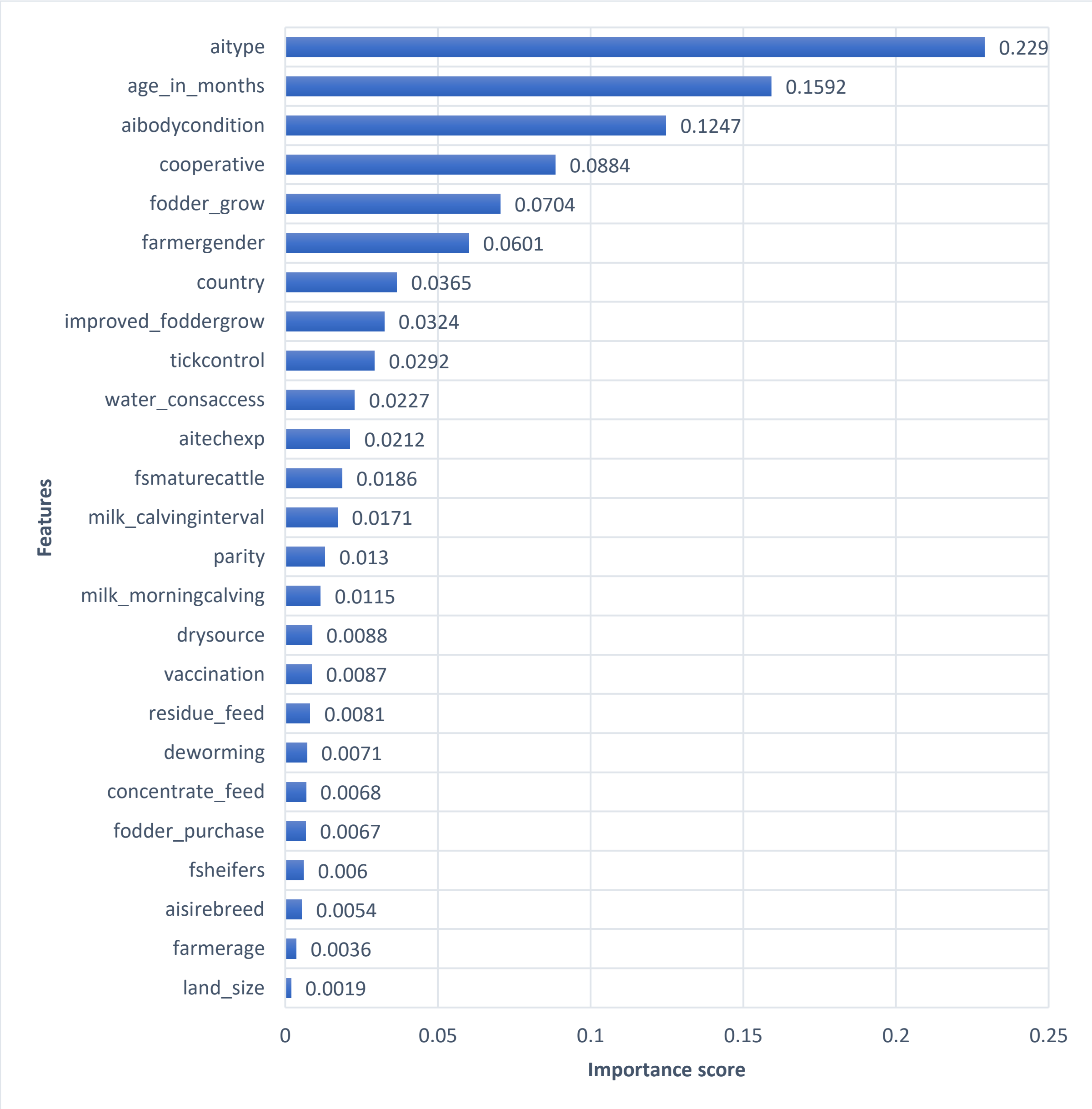
4. 1 Objective (a): Assessing Factors Influencing AI Success

The study leveraged feature importance analysis from Random Forest (RF) model to systematically identify and quantify the key factors influencing artificial insemination (AI) success in smallholder dairy systems. The reason why RF was chosen for this task is because it was the best performing model among the base learners. Feature importance in a Random Forest (RF) model refers to the relative contribution each input variable makes in predicting the target variable, in this case, the artificial insemination outcome. RF determines feature importance by averaging how much each variable decreases impurity (e.g., Gini impurity or entropy) across all decision trees in the forest (Breiman, L. 2001).

The feature importance analysis revealed several key factors affecting artificial insemination (AI) success in dairy cattle as illustrated in figure 4.5 below.

FIGURE 4.5

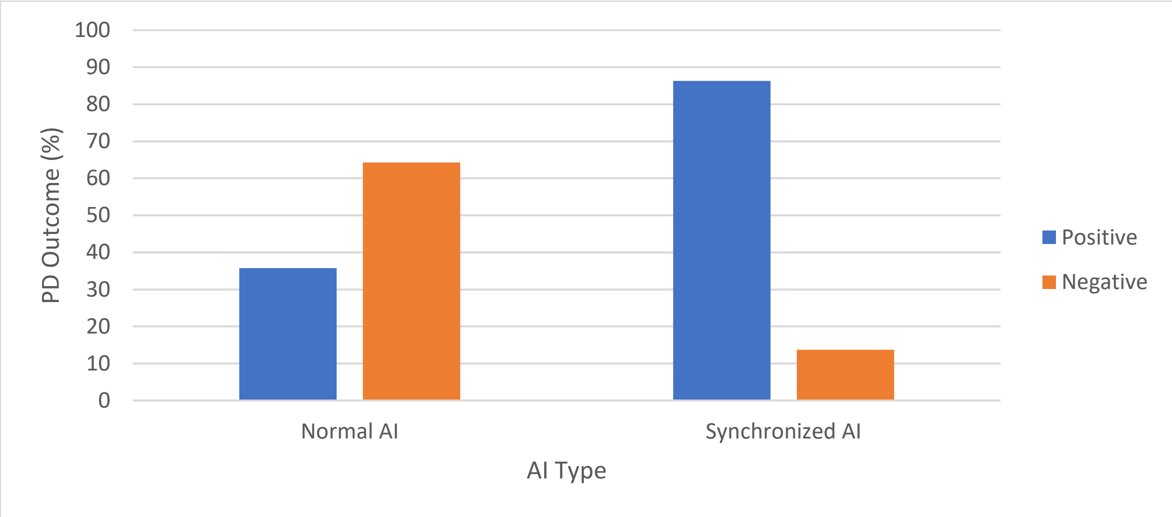
Feature importance in RF



The most critical variable is the artificial insemination type, particularly the use of estrus synchronization (importance: 0.23). It was noted that those who used estrus synchronization had more successful artificial insemination outcome (86%) as compared to those who did not use.

FIGURE 4.6

Artificial insemination type



Artificial insemination type was followed closely by the age of the cow (0.16). Cows with a positive pregnancy diagnosis outcome had a slightly higher mean age of 57.89 months compared to 56.25 months for those with a Negative outcome. This minor difference (≈1.6 months) suggests that age alone may not strongly distinguish between successful and unsuccessful conception outcomes, but a marginal trend may exist toward better AI outcomes in slightly older cows. Table 4.6 below presents descriptive statistics of cow age (in months) in relation to the pregnancy diagnosis (PD) outcome, categorized as either Positive or Negative.

TABLE 4.3

Age factor in artificial insemination

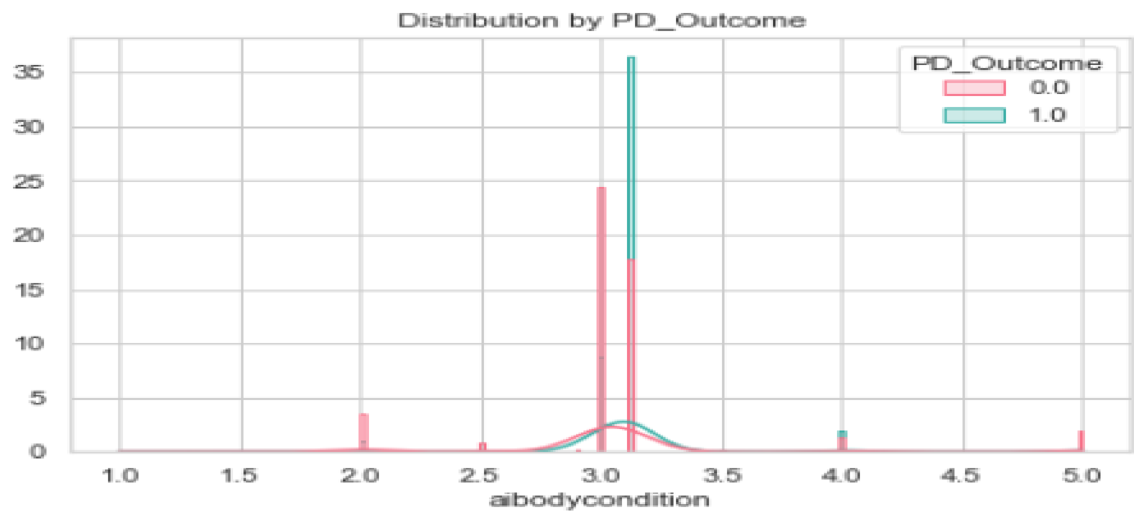
PD Outcome	count	mean	std	min	25%	50%	75%	max
Negative	744	56.25	19.75	16	41	59	65	111
Positive	603	57.89	23.53	11	40	57	71	111

The body condition scores (0.13) was next. The analysis of body condition scores (BCS) across pregnancy diagnosis (PD) outcomes indicates a modest but potentially meaningful difference between cows that conceived and those that did not. Cows with a positive PD outcome had a higher average BCS of 3.17 compared to 3.06 for those with negative outcomes. Additionally, cows in the positive group exhibited slightly less variability in condition (standard deviation = 0.44) than those in the negative group (standard deviation = 0.51), suggesting a more consistent body condition among successfully inseminated cows.

The median and interquartile ranges further emphasize this difference. The positive group had a uniform 25th, 50th, and 75th percentile value of 3.11, indicating that most cows were clustered around an optimal condition score. In contrast, the negative group had lower percentile values, with the 25th and 50th percentiles at 3.00 and the 75th at 3.11. This suggests that cows that failed to conceive tended to have slightly poorer body condition overall. This is illustrated in figure below.

FIGURE 4.7

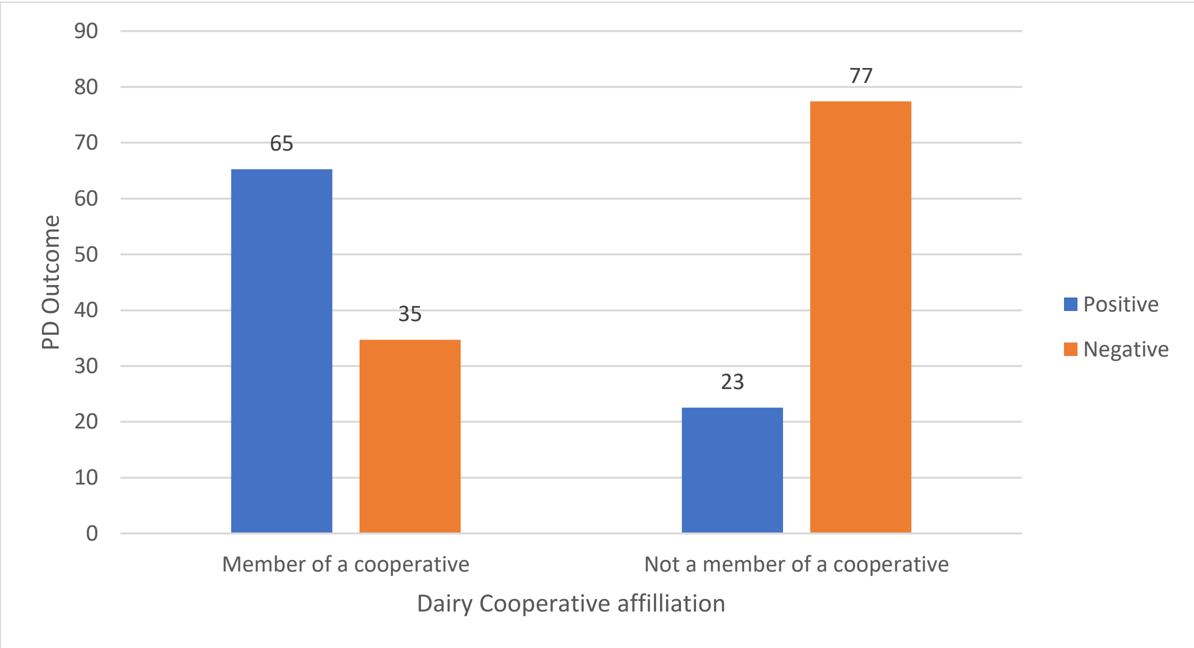
Body condition scores distribution



A farmer belonging to a cooperative (0.09) and growing fodder (0.07) also significantly affected artificial insemination outcome outcomes. Dairy cows whose owners were affiliated to a dairy cooperative had higher positive outcomes (65%) as compared to those who were not (figure 4.8). A similar trend was also noted with growing of fodder. Farmers who grew fodder had higher chances (52%) of successful artificial insemination in their dairy cows as compared to those who did not grow.

FIGURE 4.8

Affiliation to Dairy Cooperative



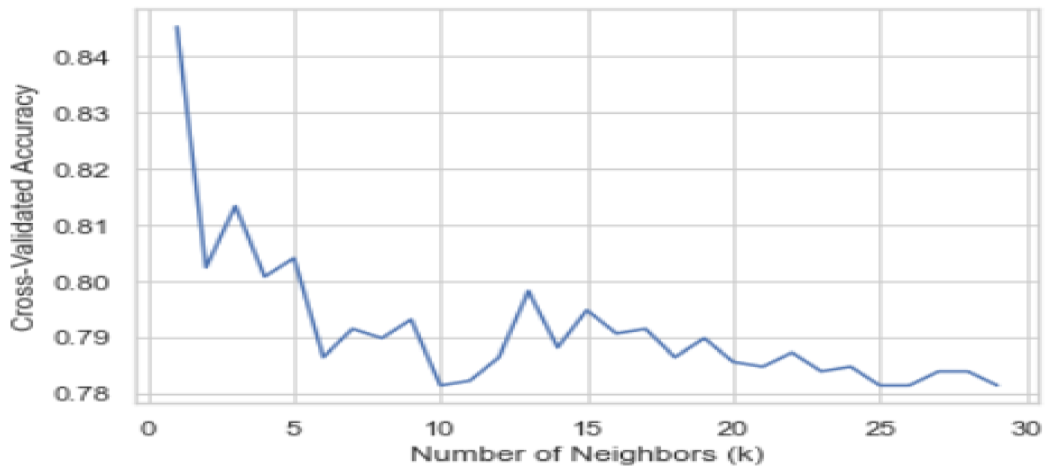
Other farmer-related factors included gender (0.06) which showed moderate influence country (0.04) which had a lower impact. Other important variables include use of improved fodder (0.03), tick control (0.03), water access (0.023), and AI technician experience (0.02), emphasizing the value of good animal care and technical support. Less influential factors include feeding practices, vaccination, deworming, milk yield, and parity, while variables like sire breed, farmer age, and land size had minimal impact.

4. 2 Objective (b): Evaluating Existing ML Models

This study evaluated the performance of various machine learning models for predicting artificial insemination likelihood of successful outcomes in dairy cattle in the case study, with particular emphasis on model optimization and ensemble techniques. The evaluation yielded insightful results. Among the models tested, the K-Nearest Neighbors (KNN) algorithm with default parameters (n_neighbors=5) showed strong performance with an accuracy, precision, recall, and F1-score of 0.83. Tuning the parameters led to a modest improvement, increasing accuracy to 0.85, precision to 0.86, recall to 0.85, F1-score to 0.85, and a ROC AUC score of 0.86. The optimal number of nearest neighbours after tuning was found to be one (n_neighbors=1) as illustrated in figure 4.9 below.

FIGURE 4.9

Optimal number of neighbours



The Support Vector Machine (SVM) model using default parameters (kernel='rbf', C=1.0, gamma='scale') yielded lower performance with an accuracy of 0.76 and an F1-score of 0.76. The parameter grid used included testing the best value for regularization parameter (C) from three values: 0.1,1 and 10. For *gamma* it was choosing between scale and auto while for *kernel* the parameters tested were *rbf* and *linear*. After hyperparameter tuning (C=10, gamma='auto', kernel: 'rbf'), the SVM's accuracy improved slightly to 0.79, with a recall of 0.79 and a ROC AUC score of 0.78.

The Decision Tree (DT) classifier performed moderately well with default parameters (accuracy and F1-score of 0.75), but tuning significantly enhanced its effectiveness, reaching an accuracy and F1-score of 0.87, with a ROC AUC score of 0.90. The best parameters for Dt were as follows: *{'criterion': 'gini', 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 5}*.

Similarly, the Random Forest (RF) classifier, with default parameters, achieved 0.79 in accuracy and F1-score. Upon tuning, it improved to 0.88 across all metrics except ROC-AUC which improved to 0.93, indicating the strongest standalone performance of all models tested. Best parameters for RF were as follows: {'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50. These results are illustrated in figure 4.10a below while figure 4.10b represents how this algorithm splits variables to classify artificial insemination outcome:

FIGURE 4.10a

Individual Model Performance

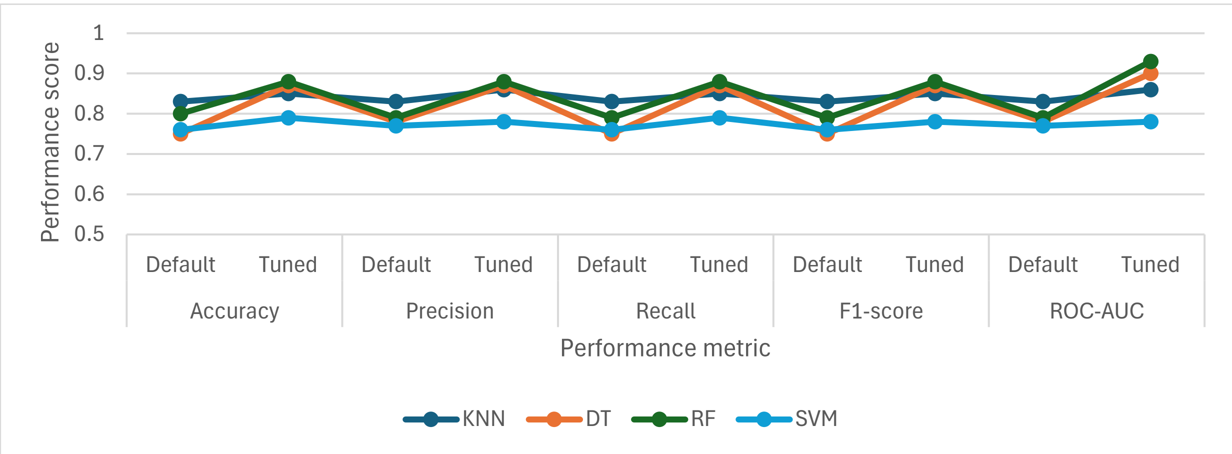
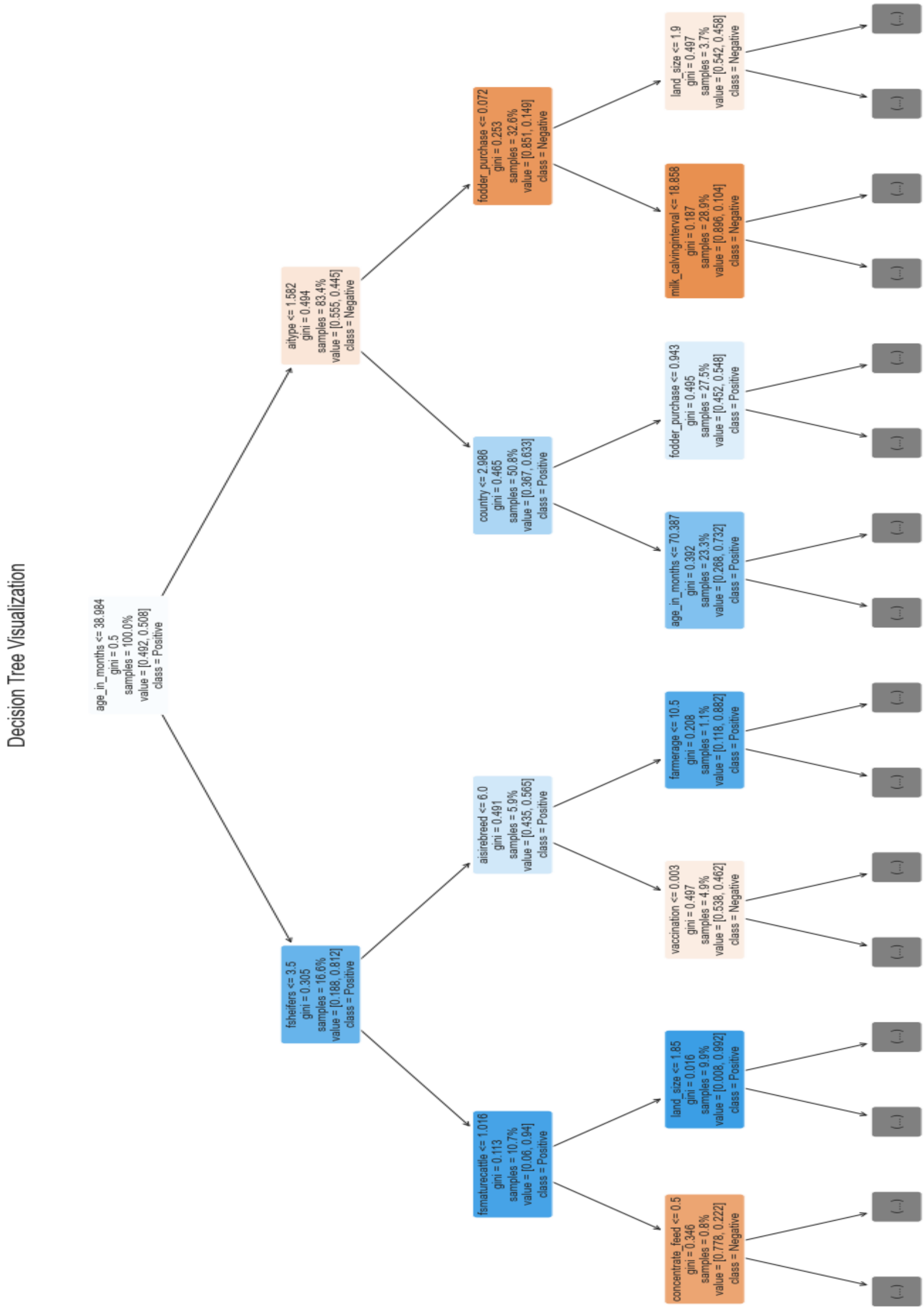


FIGURE 4.10b

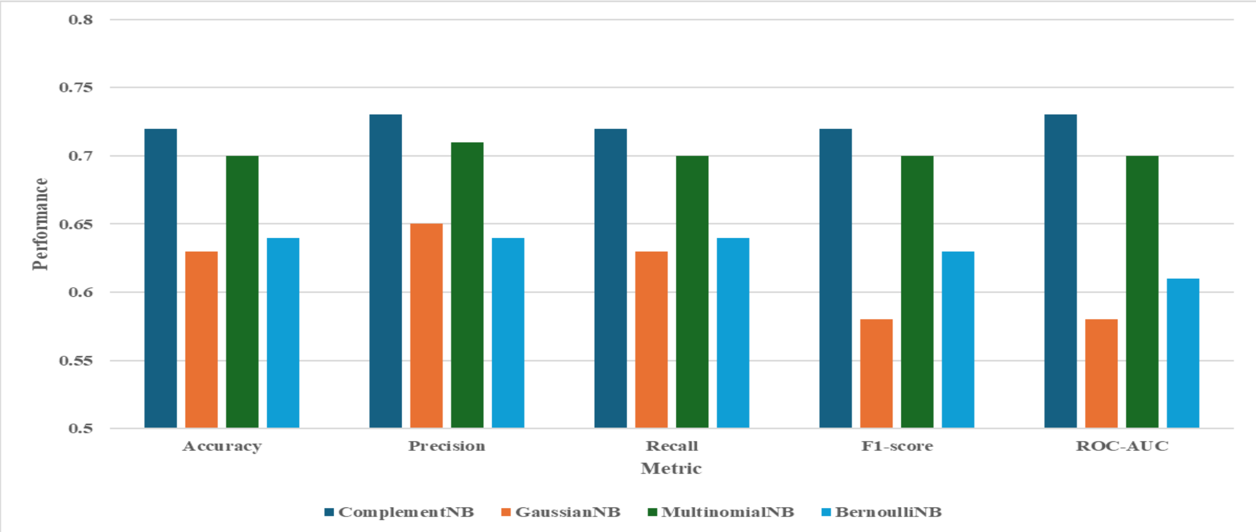
RF Decision Tree Visualization



For the Naive Bayes (NB) family, various models were tested. ComplementNB emerged as the best performer, achieving an accuracy of 0.72, F1-score of 0.72, and ROC AUC score of 0.73 outperforming GaussianNB, MultinomialNB, and BernoulliNB by significant margins. GaussianNB showed the weakest performance with an F1-score of only 0.58 and ROC AUC score of 0.58. Figure 4:11 below shows the performance of different variants of Naïve Bayes algorithm.

FIGURE 4.11

Naïve Bayes variants performance

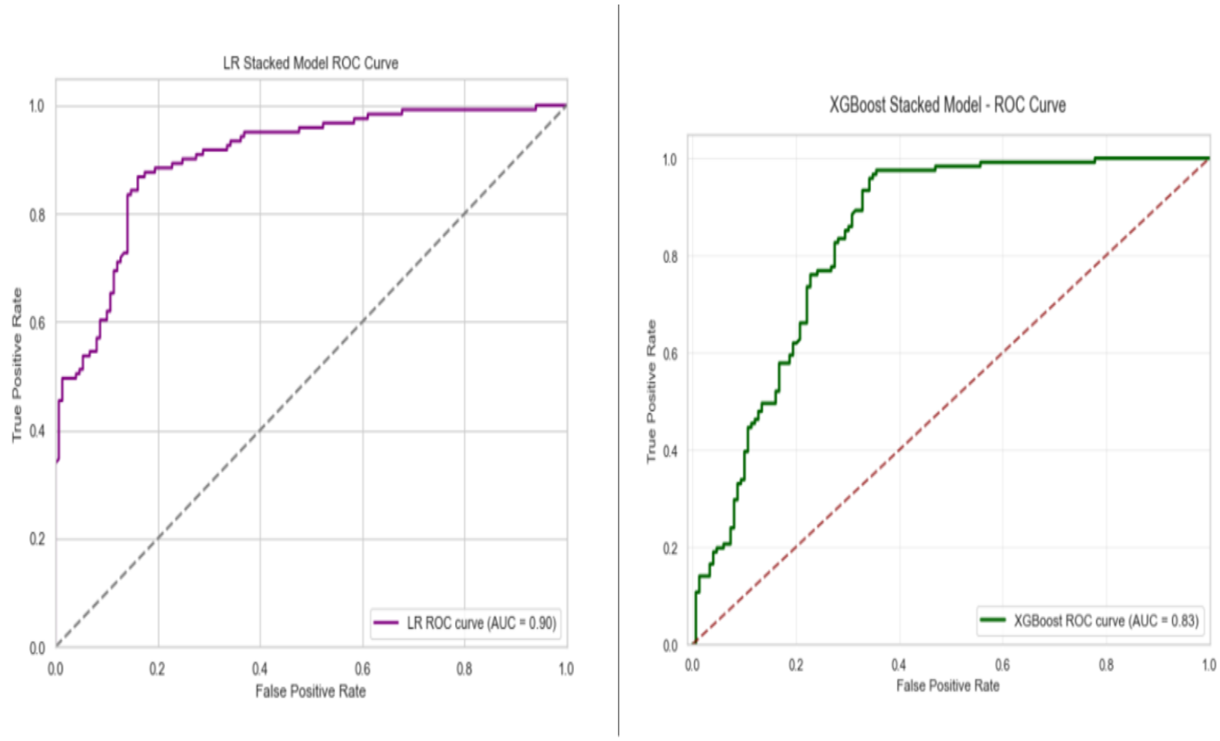


4.3 Objective (c): Developing a Stacked Ensemble Model

The stacking ensemble model, using tuned KNN, SVM, NB, DT, and RF as base learners and logistic regression (LR) as the meta-learner, performed competitively with an accuracy of 0.86 and ROC-AUC score of 0.90, closely rivaling the best individual models. When XGBoost was used as the meta-learner instead, the performance of the stacked model slightly declined (Accuracy:0.75,Precision:0.77, Recall:0.75, F1 score:0.75 and ROC-AUC:0.83). These results suggest that ensemble learning, particularly with LR as the meta-learner, provides robust predictive performance by leveraging the strengths of individual models.

FIGURE 4.12

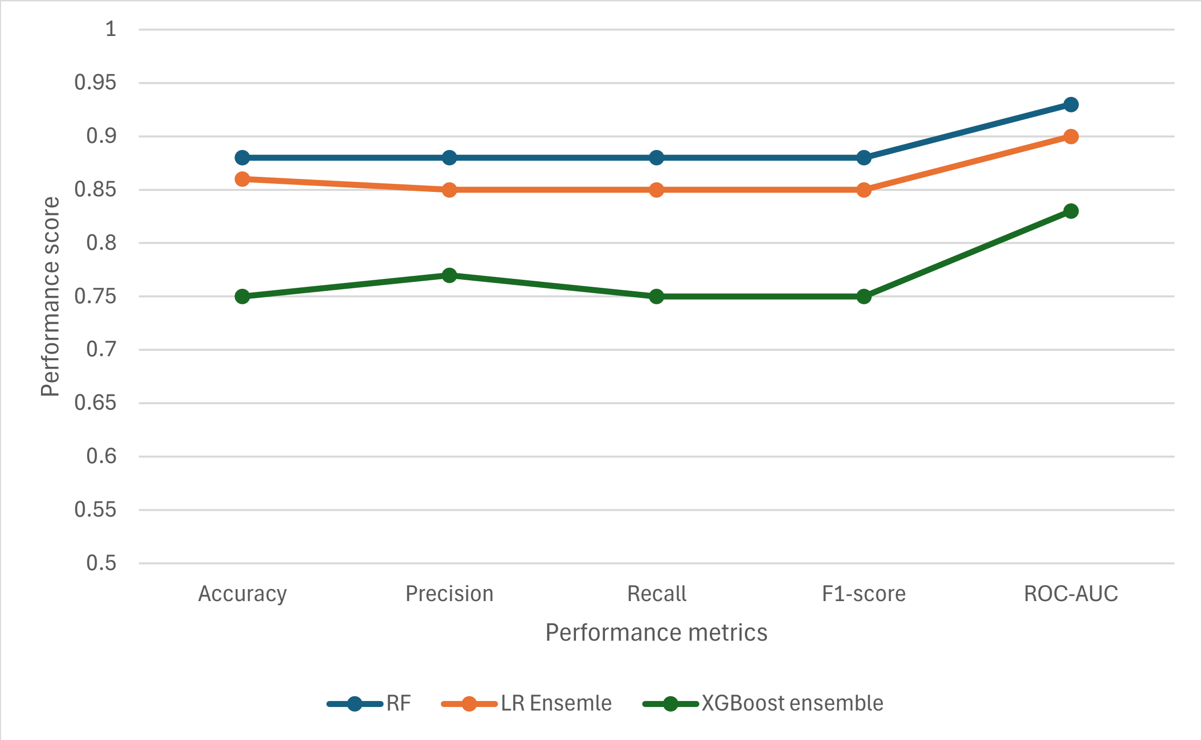
ROC curve comparison of LR and XGBoost



Overall, Random Forest with tuned parameters achieved the highest predictive performance (ROC-AUC: 0.93) in this case study as compared to the best performing LR stacking ensemble (ROC-AUC: 0.90), making both suitable choices for predicting artificial insemination outcomes in smallholder dairy systems. This is illustrated in figure 4.13. These findings provide valuable insights for developing decision support systems in dairy reproduction management, with Random Forest and well-configured ensemble models showing particular promise for accurate outcome prediction. The results underscore the importance of both algorithm selection and thorough hyperparameter optimization in developing effective predictive models for agricultural applications.

FIGURE 4.13

Performance comparison of RF and meta models



4.4 Objective (d): Evaluating Performance on Imbalanced, Mixed-Type Data

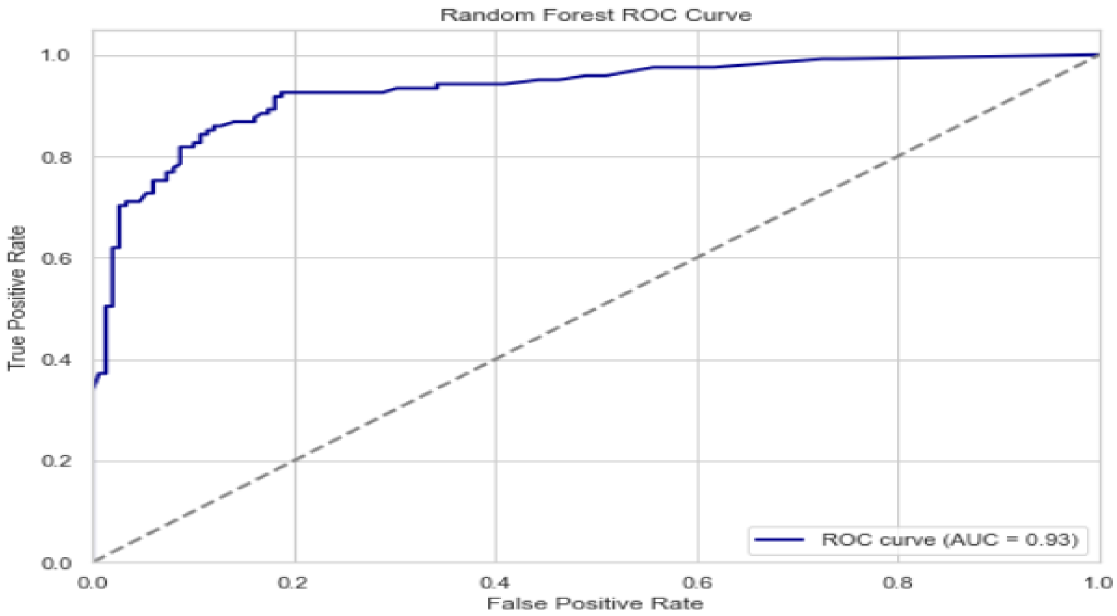
4.4.1 Assessment of accuracy, precision, recall, F1-score, and ROC-AUC

To assess model robustness on an imbalanced dataset with mixed-type features, multiple machine learning classifiers were evaluated using five key metrics: accuracy, precision, recall, F1-score, and ROC-AUC. These metrics provide a comprehensive perspective beyond simple accuracy, especially in datasets where class imbalance can mask poor model generalization.

Among individual base models, the Random Forest (RF) classifier demonstrated the best overall performance with an accuracy of 0.88, and equally strong values for precision, recall, F1-score, and ROC-AUC (at 0.90) as illustrated in the figure 4.14. Its confusion matrix also displayed a better classification of True Positive Rate and False Positive Rate as illustrated in figure 4.8 below. This confirms its strength in capturing complex patterns and handling class imbalance through ensemble learning and feature bagging.

FIGURE 4.14

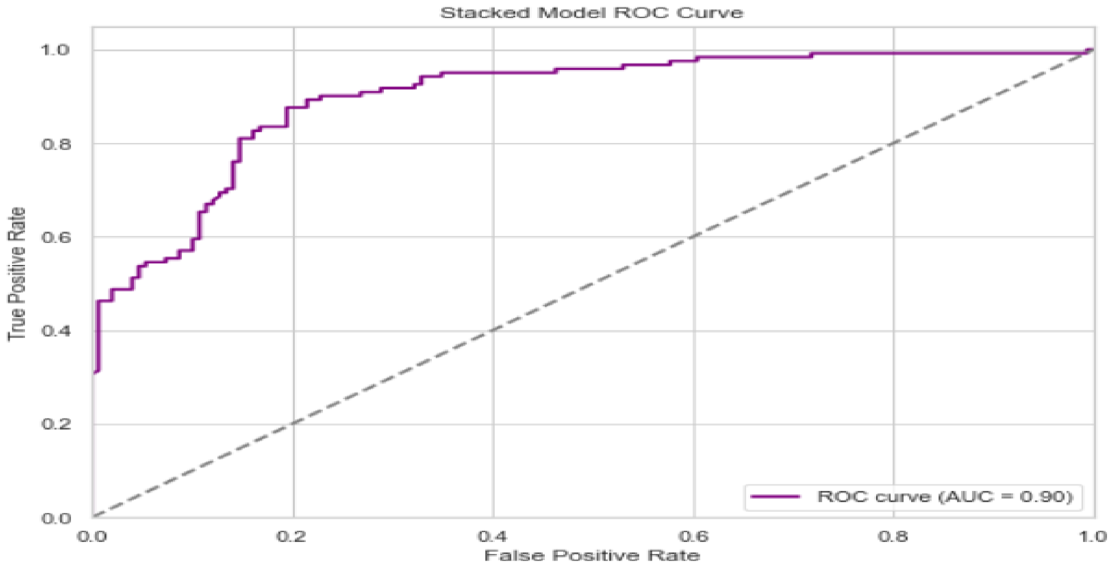
ROC Curve for Random Forest model



The Decision Tree (DT) also performed well, achieving an ROC-AUC score of 0.90 and an accuracy of 0.87 across all other performance metrics. This suggests that even a single-tree model, when properly tuned, can handle imbalanced, heterogeneous data effectively. The Logistic Regression (LR) based stacked model followed with an accuracy of 0.85 and a ROC-AUC of 0.90 (figure 4.9), demonstrating its competence and indicating the power of ensembling even with a linear meta-model. The K-Nearest Neighbors (KNN) model also performed competitively, with uniform scores of 0.85, particularly when the optimal `k` value is chosen.

FIGURE 4.15

ROC Curve for Logistic Regression stacked model

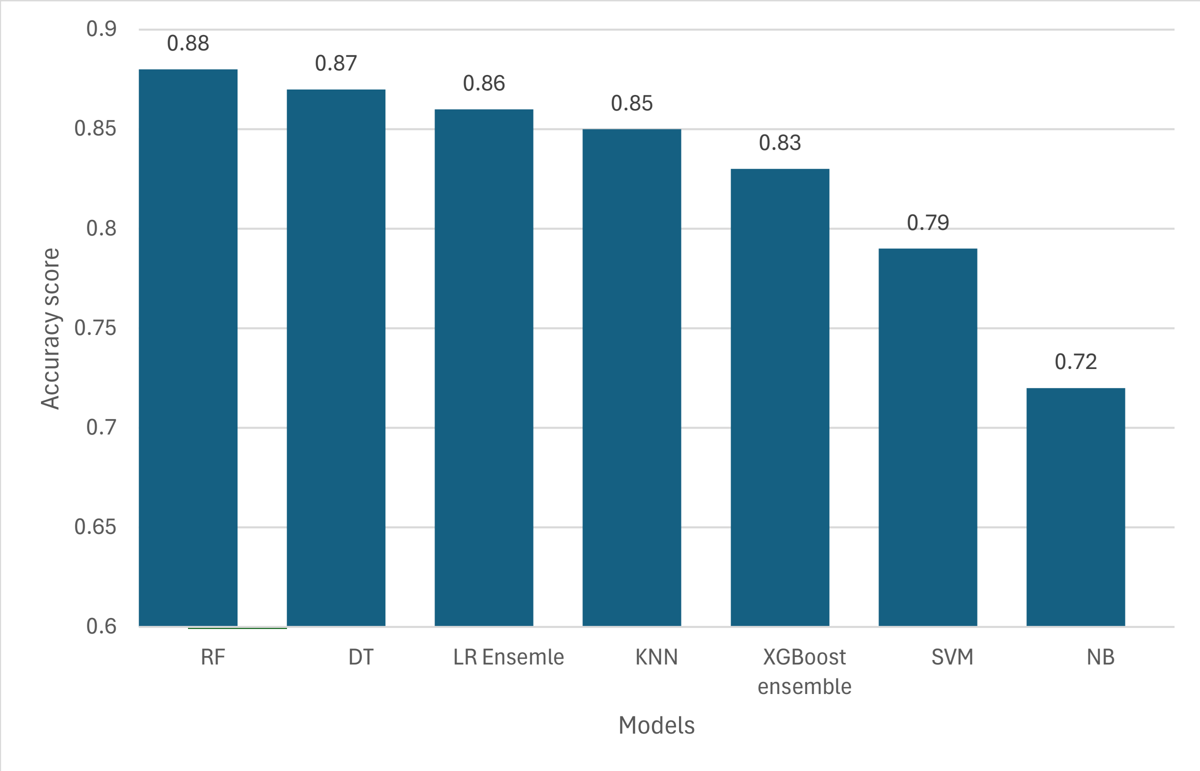


In contrast, Naive Bayes (NB) had the lowest performance, with accuracy and ROC-AUC of 0.72 and 0.73 respectively. This likely stems from the model's simplifying assumptions, which may not hold well in complex or correlated features. The Support Vector Machine (SVM) yielded moderate results (accuracy: 0.79, ROC-AUC: 0.78) and was outperformed by tree-based and ensemble approaches.

Interestingly, the XGBoost stacked ensemble model did not outperform the Random Forest or Logistic Regression-based ensemble in this context, with slightly lower scores across all metrics (accuracy and ROC-AUC: 0.83). This may be attributed to dataset size or feature interactions not being optimally captured without further fine-tuning. The accuracy of individual models is illustrated in figure 4.4 below:

FIGURE 4.16

Model accuracy comparison



4.4.2 Assessment of confusion matrix for base learners

Based on the confusion matrix results from five machine learning models, KNN, SVM, Complement Naive Bayes (NB), Decision Tree (DT), and Random Forest (RF), distinct differences in classification performance emerge, particularly in terms of their accuracy in identifying true positives and true negatives.

Random Forest (RF) exhibited the best overall classification performance, correctly identifying 89.9% of negative cases (true negatives) and 81.8% of positive cases (true positives). Its low false positive (10.1%) and false negative (18.2%) rates indicate that RF is highly reliable in both detecting successful and failed artificial insemination (AI) outcomes.

TABLE 4.4

Confusion matrix for RF

	Predicted 0	Predicted 1
Actual 0	89.9%	10.1%
Actual 1	18.2%	81.8%

Decision Tree (DT) followed closely, achieving 88.6% accuracy for true negatives and 80.2% for true positives. While slightly lower than RF, DT still maintained strong balance in minimizing both types of misclassifications, with a false negative rate of 19.8% and a false positive rate of 11.4%.

TABLE 4.5

Confusion matrix for DT

	Predicted 0	Predicted 1
Actual 0	88.6%	11.4%
Actual 1	19.8%	80.2%

Support Vector Machine (SVM) and **K-Nearest Neighbors (KNN)** showed similar patterns of performance, with SVM slightly outperforming KNN. SVM correctly predicted 85.2% of negative cases and 78.5% of positive cases, compared to 83.9% and 77.7%, respectively, for KNN. Both models demonstrated modest false negative rates (~21–22%) and false positive rates (~14–16%), indicating solid but less optimal classification capabilities than RF and DT.

TABLE 4.6

Confusion matrix for SVM

	Predicted 0	Predicted 1
Actual 0	85.2%	14.8%
Actual 1	21.5%	78.5%

TABLE 4.7

Confusion matrix for KNN

	Predicted 0	Predicted 1
Actual 0	83.9%	16.1%
Actual 1	22.3%	77.7%

Complement Naive Bayes (NB) had the weakest performance among the models. It achieved **67.1%** accuracy for negative cases and **70.2%** for positive cases, with significantly higher false positive (32.9%) and false negative (29.8%) rates. This suggests a less effective boundary between the two classes, potentially due to NB's assumption of feature independence, which might not hold in this dataset.

TABLE 4.8

Confusion matrix for NB

	Predicted 0	Predicted 1
Actual 0	67.1%	32.9%
Actual 1	29.8%	70.2%

In summary, Random Forest and Decision Tree models offer the most accurate and balanced classification of AI outcomes, making them more suitable for deployment in practical decision-support systems. SVM and KNN provide reasonably good performance but may require further tuning or combination with other models to match the robustness of tree-based classifiers. Complement NB, while lightweight and fast, underperforms in this context and may not be appropriate as a standalone model for predicting AI success. These findings reaffirm the effectiveness of ensemble and tree-based approaches in managing imbalanced, mixed-type datasets, making them highly suitable for real-world predictive analytics tasks.

CHAPTER FIVE: DISCUSSION

5.1 Introduction

This chapter synthesizes and interprets the key findings of the study on artificial insemination adoption and success prediction in Kenya’s smallholder dairy systems. The discussion contextualizes the results within existing literature, examines their practical implications, and highlights both consistencies and contradictions with prior research. The chapter is structured around the study’s main objectives: (a) evaluating factors influencing artificial insemination outcomes, (b) assessing the performance of existing machine learning models in predicting artificial insemination outcomes, and (c) exploring the potential of ensemble learning techniques to enhance prediction accuracy and (d) evaluating their t performance.

The chapter then presents the conclusion of the findings and addresses the study’s limitations, including data constraints and methodological challenges, and proposes actionable recommendations for policymakers, extension services, and future research.

5.2 Discussion of Findings

This section presents a comprehensive discussion of the study's findings, beginning with descriptive statistics before examining results organized by research objectives. The discussion interprets these findings in relation to existing literature.

5:2.1 Descriptive analysis findings:

The descriptive analysis of the data provided valuable insights into the factors that influenced the outcome of artificial insemination in dairy cows in the case study. Body condition score (BCS) is a critical indicator of a cow's nutritional and reproductive status. A moderate BCS, typically between 2.5 and 3.5 on a 1–5 scale, is optimal for fertility and conception via artificial insemination (Butler, 2003). Cows that are too thin (BCS < 2.5) may experience energy deficits, leading to delayed estrus and poor conception rates, while over-conditioned cows (BCS > 4) often have metabolic issues that impair fertility (López-Gatius, 2003). The observed average BCS of 3.11 with low variability suggests favourable conditions for successful artificial insemination. Another critical factor is age of the cow. Age is a known determinant of reproductive performance. Young cows (<2 years) may have underdeveloped reproductive systems, while older cows (>7 years) can experience reduced fertility due to accumulated health and metabolic issues (Perry et al., 2010). The average cow age of approximately 57 months (4.75 years) observed in the study falls within the reproductive prime. This aligns with findings that conception rates are higher in cows during their second to fourth lactation (Lucy, 2001). The number of times a cow has calved (parity), also affects artificial insemination outcomes. Cows in their first or second parity often show improved uterine tone and hormonal balance, aiding conception success (Diskin & Kenny, 2016). An average parity of 1.59 in this study indicates that many cows are early in their reproductive career, which is typically advantageous for artificial insemination success. Milk production reflects both genetics and the quality of nutrition and management. Higher yields are generally associated with better reproductive performance, up to a point (Walsh et al., 2011).

An average morning yield of 5.83 litres post-calving found in this study is modest but may be adequate in smallholder contexts. Proper energy balance is crucial to ensure reproductive recovery after calving. The average calving interval of 13.72 months observed in this study is within the acceptable range but slightly longer than the ideal 12-month interval. Extended calving intervals often signal delays in returning to estrus or failed conceptions, indicating suboptimal reproductive management (Royal et al., 2000). Shortening this interval through improved heat detection, nutrition, and technician skill can enhance herd productivity.

Small land sizes, like the observed mean of 1.9 acres, are common in smallholder systems and can limit forage production and animal numbers (FAO, 2018). However, when well-managed, even small operations can maintain productive artificial insemination programs, particularly with access to supplementary feeding and water resources (Moll et al., 2007). Efficient use of limited space is key to sustaining reproductive health and performance. Another factor considered was the experience of the technician performing insemination. Technician expertise is a significant factor in insemination success. The observed average experience level in this study was 2.43 years. Studies have shown that artificial insemination conception rates improve with increasing technician experience, due to better semen handling and deposition techniques (Dalton et al., 2004).

5:2.2 Objective (a) findings:

For Objective (a) assessing influential factors, the analysis revealed a clear hierarchy of influential variables. The Random Forest algorithm's feature importance rankings revealed that estrus synchronization (0.23 relative importance) is the most critical factor, consistent with experimental studies showing timed artificial insemination protocols improve conception rates by 15–25% compared to natural heat detection (Pursley et al., 2012; Ribeiro et al., 2016). This aligns with physiological evidence that precise hormonal control of ovulation enhances fertilization success (Wiltbank et al., 2014). Cow age (0.16 importance) emerges as the second most influential variable, supporting research identifying 24–36 months as the optimal breeding window for maximal reproductive efficiency (Inchaisri et al., 2010).

The strong predictive power of body condition score (BCS) in the feature importance score (0.13) also reflects established thresholds (BCS 2.75–3.25) for optimal fertility (Ferguson et al., 2006). A good BCS is achieved through proper feeding which was indicated by on-farm fodder production (0.07) performance, corroborating nutritional studies linking feed quality to reproductive performance, where adequate energy intake improves ovarian function and embryo survival (Butler, 2014; Roche et al., 2018). Another top predictive feature was farmer's cooperative membership (0.09) which outweighed individual farmer resources, echoing findings that collective systems enhance access to quality semen and technical support in smallholder settings (Baltenweck et al., 2020).

The significance of farmer gender (0.060) in predictive significance parallels gender-based disparities in resource allocation observed in East African dairy systems. Extension systems are often male-dominated and less responsive to the needs of women farmers. Female farmers receive less frequent and lower-quality extension visits and are less likely to be included in training programs or field demonstrations compared to their male counterparts (Njuki et al., 2014). This limited exposure reduces their knowledge about reproductive management technologies, including artificial insemination, and their confidence in engaging with service providers.

Health interventions show variable influence, with tick control (0.029) more consequential than vaccination (0.009) or deworming (0.007), suggesting ectoparasite management may disproportionately affect reproductive efficiency (Jonsson et al., 2001). The limited impact of sire breed (0.005) contrasts with genetic selection literature but may reflect adequate breed adaptation in studied populations (Mrode et al., 2020) while the minimal effects of water access (0.023) and land size (0.002) challenge conventional assumptions about their criticality, possibly indicating threshold effects where basic requirements are met (Mullins et al., 2019).

5:2.3 Objective (b) findings:

Accurate prediction of artificial insemination outcomes in dairy cows is essential for improving reproductive efficiency, reducing economic losses, and informing timely herd management decisions. Machine learning models are increasingly used for such predictive tasks due to their ability to identify complex, non-linear relationships among multiple variables (Kamilaris & Prenafeta-Boldú, 2018).

The evaluation of existing models under Objective (b) revealed Random Forest as the top standalone performer, with tree-based methods generally outperforming alternatives due to their native handling of mixed data types. Random Forests (RF) yielded the highest overall performance, reaching an accuracy of 0.88 with ROC-AUC score of 0.93 and equally high scores in all other metrics. Several studies have highlighted RF's suitability for agricultural data, which is often noisy and multidimensional. For example, Guo et al. (2020) found that RF achieved higher prediction accuracy than logistic regression and support vector machines in predicting reproductive success in cattle. This performance is attributed to RF's capacity for variable importance ranking and its insensitivity to overfitting when properly tuned. Similarly, in a study assessing fertility traits in dairy cows, Marques et al. (2024) demonstrated that RF outperformed other models in predicting conception success after artificial insemination, reporting high F1-scores and AUC values. RF's combination of accuracy, interpretability via feature importance, and robustness against overfitting and data irregularities make it particularly well-suited for predicting artificial insemination outcomes in dairy cows.

Decision Trees (DT) experienced the most significant improvement after hyperparameter tuning, jumping from 0.75 to 0.87 in accuracy. This indicates that DT models can greatly benefit from optimization, such as pruning and adjusting maximum depth, to avoid overfitting while still capturing critical decision rules (Shahinfar et al., 2020). Their interpretability also makes them valuable in agricultural decision-making environments.

K-Nearest Neighbors (KNN) showed strong performance, with tuned accuracy improving from 0.83 to 0.85 and corresponding improvements across precision, recall, and F1-score. KNN has been recognized for its simplicity and effectiveness in animal science tasks, especially when patterns in the data are well-separated (Abdullah et al., 2021). However, KNN can be sensitive to noise and irrelevant features, which highlights the importance of data preprocessing and tuning.

Support Vector Machines (SVM) performed moderately, with tuned accuracy improving from 0.76 to 0.79. While SVMs are powerful for high-dimensional data and non-linear classification, their performance may be sensitive to parameter settings and kernel choice (Rashid et al., 2021). Their lower performance compared to RF and KNN in this context suggests that they may not be as well-suited for datasets with overlapping class distributions or noise.

Naïve Bayes (NB) classifiers, known for their simplicity and speed, have shown varied performance in predicting artificial insemination outcomes in dairy cows. These models assume feature independence and are particularly useful for high-dimensional datasets (Zhang, 2004). Among the evaluated models, Complement Naïve Bayes (ComplementNB) demonstrated the best performance, with an accuracy of 0.72 and a ROC-AUC of 0.73. Originally developed to handle imbalanced text classification tasks, ComplementNB adjusts for skewed class distributions, which may explain its relatively high performance in this artificial insemination prediction context (Rennie et al., 2003).

Overall, the tuned Random Forest model stands out as the most effective, closely followed by Decision Trees and KNN. These findings align with previous research demonstrating that ensemble and instance-based methods often outperform others in agricultural prediction tasks (Patel & Prajapati, 2018). The above results indicates that the RF model is able to predict the outcome of artificial insemination in dairy cows specifically, whether a cow becomes pregnant following an insemination attempt with 88% accuracy.

5:2.4 Objective (c) findings:

Ensemble learning has become a cornerstone of predictive modeling in agriculture, especially in scenarios involving multifactorial influences like artificial insemination outcomes in dairy cows. Among ensemble techniques, stacking which combines the outputs of multiple base models through a meta-learner offers the flexibility to leverage heterogeneous classifiers to improve generalization performance (Wolpert, 1992; Sagi & Rokach, 2018).

In this study, two stacking variants were tested: one with Logistic Regression (LR) and another with XGBoost as the meta-learner. Both used tuned base classifiers: K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Tree (DT), Random Forest (RF), and Complement Naïve Bayes (NB). The LR-based stack outperformed its XGBoost counterpart with an accuracy and F1-score of 0.86, and ROC-AUC of 0.93, compared to 0.83 across all three metrics for the XGBoost-based stack.

These results align with existing literature suggesting that LR, despite being a simpler model, often performs robustly in stacking due to its ability to effectively weight the probabilistic outputs from diverse base learners (Seni & Elder, 2010). In contrast, while XGBoost is a powerful gradient-boosting algorithm known for its regularization and high accuracy in many domains (Chen & Guestrin, 2016), it may not always yield better results as a meta-learner, especially in cases where the base models already capture sufficient signal variance (Zhou, 2012).

In the context of dairy cow artificial insemination outcome prediction, where the dataset likely includes imbalanced and mixed-type features (categorical, binary, continuous), LR's interpretability and linear decision boundaries may help it combine predictions more conservatively and reliably. This result suggests that, despite the popularity of XGBoost, simpler meta-learners can outperform it in specialized domains such as animal breeding (Panigrahi et al., 2022; Toghyani et al., 2021).

Therefore, for dairy reproductive analytics, the choice of meta-learner should be context-driven, balancing interpretability, generalization, and dataset characteristics. The superior performance of LR in this study underscores its value in structured agricultural datasets, where overfitting from complex learners like XGBoost can be a concern if not carefully tuned.

5:2.4 Objective (d) findings:

The evaluation of model performance on imbalanced, mixed-type data in Objective (d) demonstrated that while the accuracy gave an overall success rate, the F1-score ensured balanced performance between precision and recall, the ROC-AUC validated the model's fundamental discriminative power, and the confusion matrix revealed the nature and distribution of errors as demonstrated by Wambugu et al., (2024). This multi-metric approach was particularly important for application in artificial insemination outcome prediction, where different types of errors carry different consequences since false positives might lead to unnecessary follow-up procedures, while false negatives could miss viable pregnancies. The combination of these metrics not only demonstrated the random forest model's strong performance but also provided actionable insights for optimizing models real-world implementation in dairy reproduction management (Ndemo et al., 2023).They also indicated the robustness of the LR stacked ensemble model.

5.3 Limitations

5.3.1. Data-Related Limitations

- Data Quality and Completeness: The dataset contained missing, inconsistent, or inaccurate entries, which may reduce the reliability of the model despite data cleaning and imputation.
- Sample Size and Representativeness: The dataset was drawn from a limited geographical area in Kenya, Tanzania and Ethiopia, which may not reflect conditions in other regions or countries.
- Imbalanced Dataset: The proportion of successful vs. failed artificial inseminations was skewed, and although resampling methods (e.g., SMOTE) were applied, some bias may remain.
- Limited Variables: Important socio-economic factors (e.g., farmer income, education, access to extension services) were not captured, even though they may influence artificial insemination adoption and success.
- Temporal Coverage: The dataset may not adequately account for seasonal variations in feed availability, disease prevalence, and climate effects on reproductive outcomes.

5.3.2. Model-Related Limitations

- Overfitting Risk: While tuning reduced overfitting, the use of multiple complex algorithms increases the risk of overfitting to training data.
- Computational Demands: Training and validating ensemble models required significant computing resources, which limited the training capability.

5.4 Conclusion

The analysis of factors influencing artificial insemination success in dairy cows has generated evidence-based insights that can guide more targeted interventions to improve breeding outcomes in smallholder production systems. The findings revealed that the most influential determinants of artificial insemination success were those directly linked to breeding protocols and animal-related characteristics, such as age, parity, and body condition score, rather than isolated health or nutritional measures. This highlights the importance of farmer awareness, timely estrus detection, and adherence to proper insemination practices in maximizing conception rates.

From a modeling perspective, a critical insight emerged regarding the superior performance of tree-based algorithms, particularly Random Forest (RF) and Decision Trees (DT). These models demonstrated a natural capacity to handle mixed-type datasets, tolerate missing values, and capture complex non-linear interactions within heterogeneous smallholder data. Their robustness underscores their suitability for real-world agricultural contexts where data quality is often inconsistent and multidimensional. The study also emphasized the importance of hyperparameter tuning, data stratification, and model diversity in optimizing predictive performance. While Random Forests delivered the highest standalone accuracy, the stacked ensemble approach, using Logistic Regression as the meta-learner, provided a more balanced and stable prediction framework. This hybrid approach capitalized on the complementary strengths of multiple models, thereby enhancing generalizability and reducing the risk of overfitting.

Overall, the study demonstrated that machine learning can be a powerful tool for predicting artificial insemination success in resource-limited smallholder dairy systems. RF was the best individual predictor of artificial insemination pregnancy outcomes, excelling in accuracy and interpretability, making it highly practical for smallholder advisory systems. However, the LR ensemble provided more balanced predictions, leveraging multiple algorithms to reduce errors in imbalanced or noisy smallholder data. Thus, while RF may be preferred for deployment, the LR ensemble adds value for research, policy, and scaling predictive frameworks.

5.5 Recommendations

Based on the feature importance analysis, dairy farmers and extension services should prioritize estrus synchronization programs, as this factor demonstrated the strongest influence on artificial insemination success. Alternatively measures to build capacity among farmers on estrus detection should be prioritized where synchronization programs cannot be afforded. Farmers should also monitor and optimize body condition scores**,** ensuring cows are in optimal nutritional status before insemination. Encouraging participation in farmer cooperatives can enhance knowledge-sharing and access to better breeding resources. Overall, a holistic approach focusing on synchronized breeding, proper feeding, and cooperative engagement, while maintaining good animal condition, will likely yield the greatest improvements in artificial insemination success rates. Further research should investigate why some expected factors like technician experience and feed supplements had low impact to refine recommendations.

In terms of machine learning model development for prediction of artificial insemination in dairy cows in the context of east Africa region, future work should focus on developing an optimized hybrid meta-ensemble that leverages the strength of the tree-based models (RF/DT) as base learners together with logistic regression (LR) as the meta learner because of their high performance observed in this case study.

5.5 Contributions

This research contributes a new, context-specific stacked ensemble machine learning model designed to predict artificial insemination outcomes in smallholder dairy systems in East Africa; an area where such predictive tools are scarce. Methodologically, it advances the field by integrating diverse base learners and addressing challenges of imbalanced, noisy, and mixed-type data, which are typical in smallholder datasets. By tailoring the model to local conditions, the study bridges the gap between advanced machine learning techniques and practical livestock management. This contribution is significant because it empowers smallholder farmers with data-driven decision-support tools to reduce the financial risks of failed inseminations, enhances the effectiveness of artificial insemination technicians, informs policymakers on breeding program design, and provides researchers with a replicable framework for agricultural prediction in low-resource settings.

REFERENCES

Abdullah, M., Naureen, S., & Farooq, U. (2021). Application of machine learning algorithms for predicting reproductive performance in dairy cattle. Tropical Animal Health and Production, 53(5), 1–9.

Ahmed, J.; Lorch, J.; Ong, L.; Wolfgram, J. (2018) How the Global Supply Landscape for Meat—Protein Will Evolve. Available online: https://www.mckinsey.com/industries/agriculture/our-insights/how-the-global-supply-landscape-for-meat-protein-will-evolve (accessed on 20 December 2020).

Auma, J., Rao, J and Kidoido, M. (2017). Feed the Future Accelerated Value Chain Development (AVCD) Program: Dairy component value chain analysis. Nairobi, Kenya: International Livestock Research Institute (ILRI).

Amal S, Felix F, Jens B, Katharina M, Petra W, Wolfgang R (2022); Simulation and sensor data fusion for machine learning application, Advanced Engineering Informatics, Volume 52, 2022, 101600, ISSN 1474-0346, https://doi.org/10.1016/j.aei.2022.101600.

Arno L, Matthieu S, Dirk V, Bonifacius R, Miel Hs (2021), Predicting the milk yield curve of dairy cows in the subsequent lactation period using deep learning,Computers and Electronics in Agriculture, Volume 180, 2021, 105904 SSN 0168-1699, https://doi.org/10.1016/j.compag.2020.105904.

AU-IBAR. (2023). Africa Livestock Sector Performance Review. https://www.au-ibar.org

African Dairy Genetics Gains (ADGG), 2021: *Technical report*. Nairobi, Kenya: International Livestock Research Institute. Available from https://hdl.handle.net/10568/115318

Barcellos J, Silva M and Silva J.(1996) Effects of early weaning on conception rate in Santa Gertrudis cows. *Ecol. Rurals* (Toulousa) No; 96-110D: 30-44.

Baltenweck, I.A. Omondi, K.K. Zander, S. Bauer (2017). Understanding farmers' preferences for artificial insemination services provided through dairy hubs, 10. Animal,Volume 11, Issue 4,677-686,https://doi.org/10.1017/S1751731116002354

Bebe, B. O., Baltenweck, I., Kilelu, C. W., & Kitalyi, A. J. (2020). Dairy artificial insemination adoption in Tanzania: Constraints and opportunities. *Livestock Science*, 240, 104221. DOI: 10.1016/j.livsci.2020.104221

Berry D., Kearney J, Twomey K, and Evans R. (2013). Genetics of reproductive performance in seasonal calving dairy cattle production systems. Ir. *J. Agric*. Food Res. 52:1–16.

Berry D, Evans R, and McParland S. (2011). Evaluation of bull fertility in dairy and beef cattle using cow field data. *Theriogenology* 75:172–181.

Berg, D. K., Block, J., Hansen, P. J., & Scaramuzzi, R. J. (2020). Bayesian approaches to fertility prediction in dairy cattle. Journal of Dairy Science, 103(5), 4892-4903. https://doi.org/10.3168/jds.2019-17654

Breiman, L. (1996). Stacked regressions. Machine Learning, 24(1), 49-64.

Breiman L. (2001) Random Forests. Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Brieuc M, Waters C, Drinan D, Naish K (2018), A practical introduction to Random Forest for genetic association studies in ecology and evolution Mol. Ecol. Resour., 18, pp.

755-766 https://doi.org/10.1111/1755-0998.12773

Borchers, M. R., et al. (2017). Machine-learning-based calving prediction from activity, lying, and ruminating behaviors in dairy cattle. *Journal of Dairy Science*, 100(7), 5664-5674. DOI: 10.3168/jds.2016-11526

Butler, W. R. (2003). Energy balance relationships with follicular development, ovulation and fertility in postpartum dairy cows. Theriogenology, 59(1), 9–23. https://doi.org/10.1016/S0093-691X(02)01272-2

Butler, W. R. (2014). Journal of Dairy Science, 97(5), 2751–2764. https://doi.org/10.3168/jds.2013-7223.

Buller, H., Blokhuis, H., Lokhorst, K., Silberberg, M., & Veissier, I. (2020). Animal welfare management in a digital world. Animals, 10(10), 1779.

Cabrera V.E. (2014). Economics of fertility in high-yielding dairy cows on confined TMR systems. *Animal*, 8 (Suppl. 1), pp. 211-221.

Caraviello, D. Z., Weigel, K. A., & Gianola, D. (2006). Comparison between naive Bayes and logistic regression for prediction of binary outcomes with small samples. Journal of Animal Science, 84(8), 1983-1990. https://doi.org/10.2527/2006.8481983x

Central Statistical Agency (CSA) of Ethiopia. (2020). Ethiopia Agricultural Sample Survey 2019/2020: Report on Livestock and Livestock Characteristics (Private Peasant Holdings) (Vol. II, Statistical Bulletin 587). Central Statistical Agency. URL: https://www.statsethiopia.gov.et/surveys/agricultural-sample-survey/

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357.DOI: 10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. KDD '16.

Chen, Y., Wang, X., & Liu, J. (2021). Random Forest-based decision support systems for dairy reproduction management. Computers and Electronics in Agriculture, 187, 106301. https://doi.org/10.1016/j.compag.2021.106301

Daniela M. Witten. (2011) Classification and clustering of sequencing data using a Poisson model. Ann. Appl. Stat. 5 (4) 2493 - 2518, December 2011. https://doi.org/10.1214/11-AOAS493.

DeJarnette J.M., Nebel R.L., Marshall C.E.(2009) Evaluating the success of sex-sorted semen in US dairy herds from on farm records. *Theriogenology*.;71(1):49–58. doi: 10.1016/j.theriogenology.2008.09.042.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In First International Workshop on Multiple Classifier Systems (pp. 1–15). *Springer*. DOI: 10.1007/3-540-45014-9_1

Department of Agriculture, Forestry and Fisheries [DAFF]. (2022). National artificial insemination performance report 2021-2022. Republic of South Africa. https://www.daff.gov.za

Džeroski, S., & Ženko, B. (2004). Is combining classifiers better than selecting the best one? ICML '04.

EADD. (2013). East Africa Dairy Development Project evaluation: Final report. Nairobi, Kenya: East Africa Dairy Development Project. https://hdl.handle.net/10568/79437

Etherington, W. G., Kelton, D. F., & LeBlanc, S. J. (2021). Machine learning applications in dairy reproduction: A systematic review. Computers and Electronics in Agriculture, 187, 106287. https://doi.org/10.1016/j.compag.2021.106287

FAO. (2022). Digital innovations in livestock management: Case studies from Kenya. FAO. http://www.fao.org/digital-agriculture

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fricke, P. M., et al. (2014). Efficacy of a system for monitoring individual dairy cow activity and its use for estrus detection. Journal of Dairy Science, 97(5), 2824-2834. DOI: 10.3168/jds.2013-7437

Ferguson, J. D., Galligan, D. T., & Thomsen, N. (2006). Veterinary Clinics: Food Animal Practice, 22(3), 521–538. https://doi.org/10.1016/j.cvfa.2006.07.002.

Garbarino E.J, Hernandez J.A, Shearer J.K, Risco C.A, Thatcher W.W (2004), Effect of lameness on ovarian activity in postpartum Holstein cows. *J. Dairy Sci*., 87, pp. 4123-4131

García-Ruiz, A., et al. (2016). Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. *PNAS*, 113(28), E3995-E4004. DOI: 10.1073/pnas.1519061113

Geladi P, Kowalski B.R., Partial least–squares regression: a tutorial. *Anal. Chim. Acta*, 185 (1986), pp. 1-17

Givens, M. D., & Marley, M. S. D. (2022). Infectious causes of embryonic and fetal mortality. Veterinary Clinics: Food Animal Practice, 38(2), 281–297. https://doi.org/10.1016/j.cvfa.2022.02.006

Guo, J., Zhang, T., Tian, H., & Wang, W. (2020). Prediction of reproductive success in dairy cattle using machine learning techniques. Computers and Electronics in Agriculture, 178, 105779. https://doi.org/10.1016/j.compag.2020.105779.

Gupta, R., Singh, A., & Mehra, M. (2023). Application of machine learning in animal reproductive management: A review. *Journal of Animal Reproduction and Biotechnology*, 39(2), 125–135.

GSMA. (2023). *The mobile economy: Sub-Saharan Africa 2023*. GSM Association. https://www.gsma.com/mobileeconomy/wp-content/uploads/2023/03/GSMA_ME_SSA_2023_English.pdf

Howard, J. T., et al. (2021). The future of livestock breeding: Genomic selection for efficiency, reduced emissions intensity, and adaptation. *Trends in Genetics*, 37(9), 836–850. DOI:10.1016/j.tig.2021.05.006

International Livestock Research Institute. (2020). *Smallholder dairy farming in Kenya: Trends and opportunities*. https://www.ilri.org.

Inchaisri, C., Jorritsma, R., Vos, P. L. A. M., van der Weijden, G. C., & Hogeveen, H. (2010). Theriogenology, 74(7), 1271–1278. https://doi.org/10.1016/j.theriogenology.2010.05.031.

Jonsson, N. N., Mayer, D. G., & Green, P. E. (2001). Veterinary Parasitology, 99(3), 205–219. https://doi.org/10.1016/S0304-4017(01)00461-5.

Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naïve Bayes text classifiers. In Proceedings of the 20th International Conference on Machine Learning (ICML-03), 616–623.

Roche J.R, Berry D.P, Lee J.M, Macdonald K.A, Boston R.C. (2007), Describing the body condition score change between successive calvings: A novel strategy generalizable to diverse cohorts. *J. Dairy Sci*., 90, pp. 4378-4396.

Kebede, M., Jembere, A., & Fekadu, B. (2021). Determinants of artificial insemination adoption in Ethiopian smallholder dairy farms. Livestock Science, 253, 104702. https://doi.org/10.1016/j.livsci.2021.104702

Kabunga, N. S., Dubois, T., & Qaim, M. (2014). Impact of AI on smallholder dairy incomes in Uganda. *World Development*, 64, 473-486. DOI: 10.1016/j.worlddev.2014.06.027

Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. Computers and Electronics in Agriculture, 147, 70–90. https://doi.org/10.1016/j.compag.2018.02.016.

Kenya Dairy Board. (2020). Annual report on dairy farming trends in Kenya. Kenya Dairy Board. URL: https://www.kdb.co.ke/reports

Kenya Dairy Board. (2022). *Annual report and financial statements for the year ended June 2022*. Kenya Dairy Board. URL: https://www.kdb.co.ke/reports

Kenya National Bureau of Statistics. (2023). Economic Survey 2023. https://www.knbs.or.ke

Koh Y. Q., Peiris H.N, Vaswani K, Almughlliq F.B, Meier S, Burke C.R, Roche J.R, Reed C.B, Arachchige B.J, Reed S, and Mitchell M.D. 2018. Proteome profiling of exosomes derived from plasma of heifers with divergent genetic merit for fertility. J. Dairy Sci. 101:6462–6473. https:/ / doi.org/10 .3168/ jds.2017 -14190.

Khumalo, Z. T., Muchenje, V., & Dzama, K. (2022). Effects of technician competence on artificial insemination outcomes in smallholder dairy systems. Tropical Animal Health and Production, 54(1), 45. https://doi.org/10.1007/s11250-021-03034-1

Kibiego, M. B., Lagat, J. K., & Bebe, B. O. (2020). Economic analysis of artificial insemination service delivery systems in Kenya. Tropical Animal Health and Production, 52(3), 1123-1131.

Kuncheva, L. I. (2004). *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience. Link: https://dl.acm.org/doi/10.5555/2935490

Lemma, M., Bezabih, M., & Mekuriaw, Z. (2021). Artificial insemination adoption in Ethiopian dairy systems: Drivers and constraints. Journal of Dairy Research, 88(2), 145–153.DOI: 10.1017/S002202992100045X

López-Gatius, F., Garcia-Ispierto, I., & Hunter, R. H. (2021). Predictive models in bovine reproduction: Limitations of naive Bayes classifiers. Theriogenology, 159, 98-105. https://doi.org/10.1016/j.theriogenology.2020.10.014

Li, H., Zhang, P., & Wang, L. (2020). Advanced machine learning for dairy fertility prediction: A Random Forest approach. Journal of Dairy Science, 103(11), 10489-10501. https://doi.org/10.3168/jds.2020-18654

Lucy, M. C. (2001). Reproductive loss in high-producing dairy cattle: Where will it end? Journal of Dairy Science, 84(6), 1277–1293. https://doi.org/10.3168/jds.S0022-0302(01)70158-0

Lukuyu, B., Gachuiri, C. K., Lukuyu, M. N., Lusweti, C., & Mwendia, S. (2012). *Feeding dairy cattle in East Africa*. Nairobi, Kenya: East Africa Dairy Development (ILRI). Link: https://hdl.handle.net/10568/16873

Ma, W., et al. (2020). Deep learning for genomic prediction in animal breeding: A simulation study. *Genetics Selection Evolution*, 52(1), 1–13. DOI:10.1186/s12711-020-00534-w

Mammadova, N., et al. (2021). Machine learning approaches for the prediction of successful conception in dairy cows. *Animals*, 11(3), 861. DOI: 10.3390/ani11030861

Mugabo, I., Ndahetuye, J.-B., Shumbusho, F., & Manzi, M. (2021). Artificial insemination and dairy productivity in Rwanda: Impact assessment and farmer perceptions. Frontiers in Veterinary Science, 8, 612345. DOI: 10.3389/fvets.2021.612345

Mutua, M., & VanLeeuwen, J. (2017). Factors associated with the use of artificial insemination by smallholder dairy farmers in Kenya. *Tropical Animal Health and Production*, 49(7), 1369–1376. https://doi.org/10.1007/s11250-017-1331-7

Muriuki, H. W., Okeyo, A. M., Kamande, J. K., & Kahi, A. K. (2020). Factors influencing smallholder dairy farmers' breeding choices in Kenya: A gendered perspective. *Livestock Research for Rural Development, 32(3), Article 42*. http://www.lrrd.org/lrrd32/3/muriuk32042.html

Marques, T. C., et al. (2024). Machine learning to predict pregnancy in dairy cows. Animals, 14(11), 1567. https://doi.org/10.3390/ani14111567

Mphahlele, M. R., Lehloenya, K. C., & Greyling, J. P. C. (2021). Semen handling practices and their impact on conception rates in artificial insemination programs. *Animal Reproduction Science*, 224, 106657. https://doi.org/10.1016/j.anireprosci.2020.106657

Mrode, R., Ojango, J. M. K., & Okeyo, A. M. (2020). Genomic selection and use of molecular tools in breeding programs for indigenous and crossbred cattle in developing countries: Current status and future prospects. *Frontiers in Genetics*, 11, 44. https://doi.org/10.3389/fgene.2018.00694

Mrode, R., Tarekegn, G. M., Mwacharo, J. M., & Djikeng, A. (2020). Invited review: Genomic selection for smallholder dairy systems in developing countries. Animal Frontiers, 10(2), 45-50. https://doi.org/10.1093/af/vfaa007

Mullins, C. R., Mamedova, L. K., Brouk, M. J., & Bradford, B. J. (2019). Journal of Dairy Science, 102(6), 5467–5474. https://doi.org/10.3168/jds.2018-15789.

National Institute of Statistics of Rwanda (NISR). (2022). Rwanda Dairy Development Strategy 2021-2030. Ministry of Agriculture and Animal Resources (MINAGRI), Government of Rwanda. Report No.: NISR/2022/MINAGRI-01 URL: https://www.statistics.gov.rw/publication/dairy-development-strategy

Ndou, S. P., Banga, C. B., & Norris, D. (2023). Evaluation of artificial insemination technician skills and their relationship to pregnancy rates in South Africa. Livestock Science, 267, 105135. https://doi.org/10.1016/j.livsci.2022.105135

Neves, R. C., & LeBlanc, S. J. (2020). Reproductive management practices and performance of Canadian dairy herds using automated activity-monitoring systems. *Journal of Dairy Science*, 103(4), 3826-3839. DOI: 10.3168/jds.2019-17237

Neves, R. C., LeBlanc, S. J., & Cerri, R. L. (2022). Comparison of machine learning algorithms for pregnancy prediction in dairy cattle. Journal of Dairy Research, 89(1), 45-52. https://doi.org/10.1017/S0022029922000068

Ngeno, K., Van der Linden, A., Vogelzang, R., & Van Arendonk, J. A. M. (2019). Challenges and opportunities in improving artificial insemination efficiency in smallholder dairy systems. Livestock Science, 228, 1-8.

Njuki, J., Waithanji, E., Sakwa, B., Kariuki, J., Mukewa, E., & Ngige, J. (2014). World Development, 58, 1–12.

Ojango, J.M.K., Mrode, R., Rege, J.E.O., Mujibi, D., Strucken, E.M., Gibson, J.P., & Mwai, O. (2019). Influences on the efficiency of artificial insemination in smallholder dairy systems in Kenya. *Animal Genetic Resources*, 64, 1-10.

Omondi I.A., Zander K.K., Bauer S., Baltenweck I. (2017). Understanding farmers' preferences for artificial insemination services provided through dairy hubs. Animal, 11 (4) , pp. 677-686.

Omore et al. (1999). Dairy development in Kenya: The past, the present, and the future. ILRI

Patel, H., & Prajapati, H. (2018). A survey on machine learning-based crop yield prediction systems. International Journal of Computer Applications, 178(38), 15–22. https://doi.org/10.5120/ijca2018917306.

Panigrahi, S., Patri, B. K., & Dash, S. R. (2022). Application of machine learning techniques in animal science: A review. International Journal of Livestock Research, 12(1), 1–12.

Pralle R.S, and White H.M. (2020). Symposium review: Big data, big predictions: Utilizing milk Fourier-transform infrared and genomics to improve hyperketonemia management. J. Dairy Sci. 103:3867–3873.

Pursley, J. R., Souza, A. H., & Carvalho, P. D. (2012). Animal Reproduction, 9(3), 279–286.

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1), 37–63.

Rashid, M. M., Rahman, M. M., Hassan, M. M., & Aljahdali, H. (2021). Support vector machine for agricultural data classification: A review. Artificial Intelligence in Agriculture, 5, 44–58. https://doi.org/10.1016/j.aiia.2020.12.002.

Roche, J. R., Burke, C. R., Crookenden, M. A., & Heiser, A. (2018). Animal Reproduction Science, 192, 21–28. https://doi.org/10.1016/j.anireprosci.2018.03.002.

Rodríguez J. J., Kuncheva L.I, and Alonso C.J. (2006). Rotation forest: A new classifier ensemble method. *IEEE Trans. Pattern Anal. Mach. Intell*. 28:1619–1630.

Roelofs, J., López-Gatius, F., & Hunter, R. H. F. (2021). When is a cow in estrus? Clinical and practical aspects. Theriogenology, 105, 42-47. https://doi.org/10.1016/j.theriogenology.2020.11.014

Ribeiro, E. S., Gomes, G., Greco, L. F., Cerri, R. L. A., Vieira-Neto, A., & Santos, J. E. P. (2016). Journal of Dairy Science, 99(3), 2246–2256. https://doi.org/10.3168/jds.2015-10063.

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1249. https://doi.org/10.1002/widm.1249.

Shalloo L, Cromie A, McHugh N. (2014), Effect of fertility on the economics of pasture-based dairy systems. Animal, 8 (Suppl. 1), pp. 222-231

Schefers J. M., Weigel K.A, Rawson C.L, Zwald N.R, and Cook N.B. (2010). Management practices associated with conception rate and service rate of lactating Holstein cows in large, commercial dairy herds. *J. Dairy Sci*. 93:1459–1467.

Seni, G., & Elder, J. (2010). Ensemble methods in data mining: Improving accuracy through combining predictions. Morgan & Claypool Publishers.

Shahinfar S., Mehrabani-Yeganeh H, Lucas C, Kalhor A, Kazemian M, and Weigel K.A. (2012). Prediction of breeding values for dairy cattle using artificial neural networks and neuro-fuzzy systems. *Comput. Math. Meth. Med*. 2012:127130.

Shahinfar, S., et al. (2020). Prediction of insemination outcomes in Holstein dairy cattle using machine learning algorithms. Computers and Electronics in Agriculture, 176, 105616. DOI: 10.1016/j.compag.2020.105616

Sheldon, I. M., Owens, S. E., & Turner, M. L. (2023). Postpartum uterine health in dairy cattle. Animal Reproduction Science, 248, 107182. https://doi.org/10.1016/j.anireprosci.2023.107182

Shine P.; Murphy M.D. (2022). Over 20 Years of Machine Learning Applications on Dairy Farms: A Comprehensive Mapping Study. *Sensors* 2022, 22, 52. https://doi.org/10.3390/s220100525936793

Swai, E. S., Karimuribo, E. D., & Ogden, N. H. (2023). Helminth control and fertility in Tanzanian dairy cattle. Veterinary Parasitology: Regional Studies and Reports, 38, 100834. https://doi.org/10.1016/j.vprsr.2023.100834

Tang Y., Dananjayan S., Hou C., Guo Q., Luo S., He Y. (2020). A survey on the 5G network and its impact on agriculture: Challenges and opportunities. *Comput. Electron. Agric*. 2021; 180:105895. doi: 10.1016/j.compag.2020.105895.

Tanzania Livestock Research Institute (TALIRI). (2019). Tanzania Livestock Modernization Initiative: Final Report. *Ministry of Livestock and Fisheries, United Republic of Tanzania*. URL: https://www.taliri.go.tz/publications

Tsai, D. M., & Huang, C. Y. (2020). Automated estrus detection in dairy cows using deep learning and tailhead movement analysis. *Computers and Electronics in Agriculture*, 178, 105752. DOI: 10.1016/j.compag.2020.105752

Toghyani, M., Jalali, S., & Alikhani, M. (2021). Predicting the pregnancy status of dairy cows using machine learning algorithms. Computers and Electronics in Agriculture, 180, 105892. https://doi.org/10.1016/j.compag.2020.105892.

Ting, K. M., & Witten, I. H. (1999). Issues in stacked generalization. Journal of Artificial Intelligence Research, 10, 271–289. https://doi.org/10.1613/jair.594

Uganda Bureau of Statistics (UBOS). (2021). Uganda National Livestock Census 2021: Final Report. Uganda Bureau of Statistics. URL: https://www.ubos.org/publications/.

Verma O., Kumar R., Kumar A., Chand S. (2012): Assisted Reproductive Techniques in Farm Animal – From Artificial Insemination to Nanobiotechnology. Vet. World, 5(5): 301-310.

Wang, Y., Chen, X., & Li, M. (2022). IoT-enabled Random Forest systems for precision dairy reproduction. Computers and Electronics in Agriculture, 198, 107015. https://doi.org/10.1016/j.compag.2022.107015

Walsh, S. W., Williams, E. J., & Evans, A. C. O. (2011). A review of the causes of poor fertility in high milk producing dairy cows. Animal Reproduction Science, 123(3-4), 127–138. https://doi.org/10.1016/j.anireprosci.2010.12.001.

Welay G.M., Tedla D.G., Teklu G.G., Weldearegay S.K., Shibeshi M.B., Kidane H.H., Gebrezgiabher B.B., Abraha T.H. (2018), A preliminary survey of major diseases of ruminants and management practices in Western Tigray province, northern Ethiopia. BMC Vet. Res. 2018;14:293. doi: 10.1186/s12917-018-1621-y.

Wolc, A., et al. (2016). Genome-wide association studies for fertility traits in Holstein dairy cattle. *Journal of Dairy Science*, 99(3), 1983-1990. DOI: 10.3168/jds.2015-9686

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks, 5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Yan, H., et al. (2022). "Deep learning-based prediction of dairy cow fertility from electronic health records." Biosystems Engineering, 214, 1-12. DOI: 10.1016/j.biosystemseng.2021.12.010

Yuksel Agca, John K. Critser (2006). The Laboratory Rat (2nd Edition); https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/artificial-insemination

Zhang, Q., Liu, H., & Wang, J. (2023). Comparative analysis of machine learning algorithms for fertility prediction in dairy cattle. Artificial Intelligence in Agriculture, 7, 15-26. https://doi.org/10.1016/j.aiia.2023.02.002

Zhi-Hua Z (2012), Ensemble Methods: Foundations and Algorithms, CRC Press.

APPENDIX 1: KCA UNIVERSITY SCIENTIFIC AND ETHICS REVIEW CLEARENCE

KCA UNIVERSITY
- 1 9 8 9 -

Thika Road, Ruaraka
P.O. Box 56808-00200 Nairobi Kenya
Pilot Line: +254 20 8070408/9

Tel: +254 20 3537842
Fax: +254 20 8561077
Mobile: +254 734 888022, 710 888022
Email: kca@kca.ac.ke
Website: www.kca.ac.ke

KCA UNIVERSITY SCIENTIFIC AND ETHICS REVIEW

REF: **KCAU/SERC/SOT008** **Date: 24th JUNE 2025**

TO: **ERICK KIPKEMOI RUTTO (19/06544)**

Dear Sir/Madam,

RE: A MACHINE LEARNING PREDICTION MODEL FOR ARTIFICIAL INSEMINATION OUTCOMES IN KENYAN DAIRY COWS USING ENSEMBLE LEARNING: A CASE STUDY

This is to inform you that KCA University Scientific Ethics Review Committee (KCAUSERC) has reviewed and approved your above research proposal. Your application approval number is ***KCAUSERC/*SOT008**. The approval period is ***18th July 2025 – 18th July, 2026.***
This approval is subject to compliance with the following requirements;

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

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

Yours sincerely

Dr. Caroline Ntara
Chairperson
KCA University Scientific & Ethics Review Committee

Advancing Knowledge, Driving Change

APPENDIX 2: NACOSTI PERMIT

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: 817427

Date of Issue: 23/July/2025

RESEARCH LICENSE

This is to Certify that Mr.. Erick Kipkemoi Rutto of KCA University, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Kakamega, Kisumu on the topic: A MACHINE LEARNING PREDICTION MODEL FOR ARTIFICIAL INSEMINATION OUTCOMES IN KENYAN DAIRY COWS USING ENSEMBLE-LEARNING: A CASE STUDY for the period ending : 23/July/2026.

License No: NACOSTI/P/25/4177355

817427

Applicant Identification Number

Ag. Director General
NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

See overleaf for conditions

APPENDIX 3: SECONDARY DATA DISCLAIMER FORM

KCA UNIVERSITY
~ 1 9 8 9 ~

KCA UNIVERSITY SCIENTIFIC AND ETHICS REVIEW COMMITTEE

DISCLAIMER FORM FOR PRINCIPAL INVESTIGATORS (PIS) COLLECTING SECONDARY DATA

This Disclaimer Form serves to release the Scientific and Ethics Review Committee (SERC), KCA University, and the National Commission for Science, Technology, and Innovation (NACOSTI) from any liability concerning the unlawful or unethical use of data collected by the Principal Investigator (PI) in their research. By signing this form, the Principal Investigator acknowledges and agrees to comply with all relevant data protection laws, including the Data Protection Act of Kenya and other applicable regulations.

1. Acknowledgment of Liability

The Principal Investigator (PI) acknowledges and agrees that the Scientific and Ethics Review Committee (SERC) of KCA University and NACOSTI shall not be held liable or responsible for any unlawful, unethical, or unauthorized use of the data collected as part of the research project. The PI is solely responsible for ensuring that the data is collected, stored, and processed in compliance with all applicable data protection laws.

2. Compliance with Data Protection Laws

The PI is required to adhere to the following data protection laws in the collection, storage, and processing of data in Kenya:

- **Kenya Data Protection Laws**:
    - The PI must ensure compliance with the *Data Protection Act of 2019* of Kenya, which governs personal data collection, processing, and storage.
    - The PI is responsible for obtaining informed consent (Where applicable) from research participants and ensuring that personal data is collected for legitimate purposes, stored securely, and used only in accordance with the stated purposes.
    - The PI must implement appropriate security measures to protect personal data from unauthorized access or misuse, and ensure that data is only retained for as long as necessary.

3. Liability Release and Disclaimer

The PI agrees that the Scientific and Ethics Review Committee (SERC) and KCA University are not liable for any actions taken by the PI that are in violation of applicable laws. This includes, but is not limited to, the misuse, unauthorized disclosure, or improper handling of data that is in violation of the laws outlined above.

Additionally, the PI agrees that NACOSTI, as the regulatory body overseeing scientific research in Kenya, shall not be held liable for any breaches or failures in compliance resulting from actions taken by the PI, including any unauthorized use of data.

---

4. Data Use and Research Details

The PI is required to provide specific details about the data to be collected in the sections below:

- **Type of Data**
  (Please describe the nature of the data to be collected, e.g., personal, sensitive, demographic, health data, etc.):
  - Farmer demographic and animal data

- **Purpose of Data Collection**
  (Please explain the primary objectives of the research and how the data will be used):
  - To develop a machine learning model for predicting artificial insemination outcome of dairy cows in smallholder systems in Kenya.
  - The data contains variables that influence the outcome of artificial insemination and will be helpful in building a predictive ML model.

- **Method of Data Collection**
  (Please specify the methods that will be used to collect data, e.g., surveys, interviews, observations, etc.):

- **This data was collected through surveys and will be extracted from a database for use in the model.**

- **Data Storage**
  (Please describe how and where the data will be stored, including any digital or physical storage methods used to ensure security):
    - **This data is stored in the organizational database**

- **Data Access**
  (Please specify who will have access to the data during and after the research process):
- **The student (Erick Rutto) will have access to all extracted data relevant to the study while the first supervisor (Dr. Lucy Mburu) will have access to anonymized version of the extracted data if needed.**

- **Retention Period**
  (Please indicate how long the data will be kept and the conditions under which it will be disposed of or anonymized):
- **Until the study is completed and the dissertation is defended successfully.**

---

5. Consent and Understanding

By signing this form, the PI acknowledges understanding the terms and conditions outlined above, including the legal obligations to adhere to Kenyan data protection laws. The PI understands that they are solely responsible for ensuring the lawful collection, processing, and storage of data.

---

6. Signatures

- **Principal Investigator (PI) Name**: __Erick Kipkemoi Rutto_________
- **Research Project Title**: AN ENSEMBLE LEARNING MODEL FOR PREDICTION OF ARTIFICIAL INSEMINATION OUTCOMES IN KENYAN DAIRY COWS: A CASE STUDY.
- **Date of Data Collection**: May 2019 to January 2022
- **Researcher Signature**: ____________________________________
- **Date**: 16/7/2025