Optimizing Insurance Time for Claim Resolution Through Machine Learning

By

Gladys Jepkoech Kiprono

169112

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

Institute of Mathematical Sciences and iLab Africa

Strathmore University

Nairobi, Kenya

June, 2025

Declaration and Approval

Declaration

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

Student's Name: **Gladys Jepkoech Kiprono**

Sign: _____ Date: 26th May 2025

Approval

The dissertation of **Gladys Jepkoech Kiprono** was reviewed and approved by the following:

Dr. Kennedy Senagi,
Lecturer, Insititute of Mathematical Sciences,
Strathmore Univeristy.

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

Prof. Bernard Shibwabo,
Director of Graduate Studies,
Strathmore University.

Abstract

As insurance plays a key role in financial protection, the ability to predict claim resolution time has become essential for improving operational efficiency and customer satisfaction. This study applied machine learning and explainable AI techniques to estimate the time taken to resolve insurance claims using a dataset from the Association of Kenya Insurers (AKI), comprising 13,000 records with 33 features.

After preprocessing and feature engineering, several regression models were trained and evaluated, with XGBoost emerging as the best-performing model—achieving an R^2 of 51.8%, outperforming Random Forest (45.7%) and Gradient Boosting (46.6%), and recording the lowest RMSE and MAE values.

SHAP analysis highlighted key predictors, including PolicyUpgradesLastYear, PolicyStartYear, and PolicyDurationMonths. The final model was deployed via a Streamlit interface, offering a practical and interpretable solution for real-time prediction of claim resolution time, with potential to enhance decision-making and service delivery in the insurance sector.

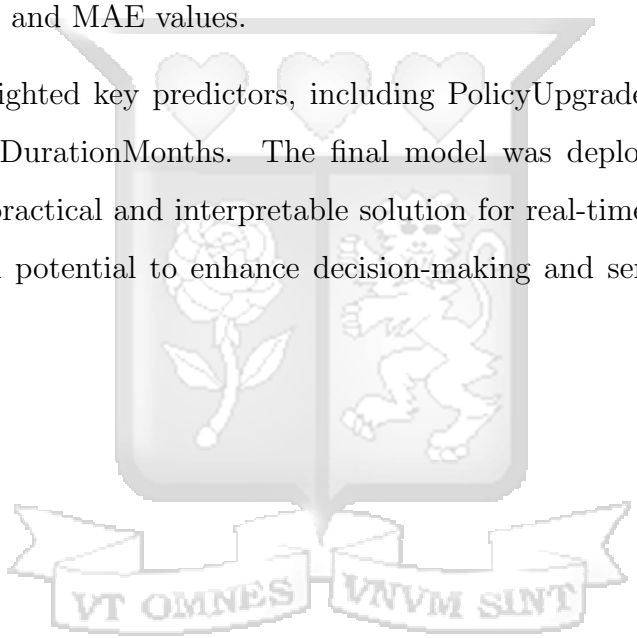


Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
AKI	Association of Kenya Insurers
CSV	Comma-Separated Values
CRM	Customer Relationship Management
MAE	Mean Absolute Error
MAR	Missing At Random
MCAR	Missing Completely At Random
MNAR	Missing Not At Random
ML	Machine Learning
RMSE	Root Mean Square Error
R^2	Coefficient of Determination
SHAP	SHapley Additive exPlanations
SVR	Support Vector Regression
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting

Acknowledgments

I would like to express my deepest gratitude to everyone who supported me throughout this project.

First, I thank Almighty God for the strength, good health, and resilience that allowed me to complete this work.

My sincere appreciation goes to my supervisor, **Dr.Kennedy Senagi**, for his invaluable guidance, encouragement, and constructive feedback throughout the course of this project.

I also extend my heartfelt thanks to the faculty and staff at Strathmore University for providing a conducive learning environment and the resources needed for my research.

Special thanks to my family and friends for their unwavering support, patience, and motivation during this journey.

Lastly, I acknowledge all the individuals and organizations who contributed data, insights, or tools that made this work possible.

Dedication

I dedicate this project to my lovely family whose endless love and support are always energizing and a source of inspiration.

Special thanks to my parents for instilling in me the values of hard work and perseverance and to my friends and mentors for always believing in me even in moments where I doubted myself.

To all those committed to advancing data-driven solutions for financial inclusion in Kenya, this work is for you.

Chapter 1: Introduction

1.1 Background

Insurance serves as a vital financial mechanism for mitigating business risks and protecting individuals from significant losses caused by unforeseen events. Operating on the principle of risk pooling, insurers collect premiums from policyholders and, in return, provide compensation when covered losses occur. This system not only helps manage individual financial burdens but also contributes to broader economic stability by enabling effective planning and resource allocation (Investopedia, 2023).

The insurance industry in Kenya has experienced notable growth since the establishment of its first insurance company in 1901. Currently, 58 registered insurers offer a diverse range of life and general insurance products (Association of Kenya Insurers, 2022). Despite this expansion, insurance penetration remains low, only approximately 2.4% of Kenya's population uses insurance services, significantly below the global average of 6.8% (Investments, 2023). This gap reflects persistent challenges, such as the perception of insurance as a luxury rather than a necessity.

Central to insurance operations is the claims process, the formal request by policyholders for compensation when losses occur. A key determinant of the effectiveness of this process is the time taken to settle claims. Delays can erode trust in insurers and discourage uptake. In the first quarter of 2022, the Insurance Regulatory Authority (IRA) reported a 40% increase in complaints, totaling 506, with delayed settlements cited as a major contributor (Guguyu, 2022). Some insurers, such as Monarch Insurance, recorded 73 unresolved complaints in Q4 2022 alone (University, 2023).

Current claim processing approaches in Kenya are often unpredictable. Policyholders are commonly given vague or inconsistent timelines, leading to frustration when expectations are not met. On the insurer's side, the lack of accurate forecasts complicates resource planning and contributes to workflow inefficiencies (Cytonn Investments, 2023).

This project explores the application of machine learning to predict insurance claim settlement times more accurately. By analyzing historical claim data, policy attributes, claim complexity, and processing workflows, predictive models can provide more reliable estimates. These models not only highlight factors contributing to delays but also enable

insurers to intervene early and manage processing expectations more effectively (Cytonn Investments, 2023).

Implementing predictive analytics in claim resolution offers several benefits to the Kenyan insurance sector:

(a) Transparent and consistent communication of settlement timelines.

(b) Improved operational efficiency through better resource allocation.

(c) Early identification of claims likely to experience delays.

(d) More equitable handling across different claim types.

(e) Continuous process improvement based on data-driven insights.

Through the development and deployment of these models, this research aims to enhance the efficiency and reliability of Kenya's insurance claims process. By fostering transparency and operational consistency, it seeks to strengthen trust between insurers and policyholders and ultimately promote broader insurance adoption. The approach also holds potential as a replicable solution for other emerging markets facing similar challenges in claims processing (PwC, 2024).

1.2 Problem Statement

The insurance claim resolution process in Kenya remains slow and inconsistent. While insurers are required to settle claims within 90 days of establishing liability (Insurance Regulatory Authority (IRA), 2023), many struggle to meet this standard. On average, the process takes between 7 to 10 days, assuming timely submission of accurate documentation, but this timeline is frequently exceeded. This inefficiency is largely due to the continued reliance on traditional, manual assessment methods that lack scalability and speed. The increasing complexity and volume of claims further strain insurers' capacity, resulting in delayed settlements, poor assessments, and inadequate resource planning. These inefficiencies contribute to rising customer dissatisfaction and reputational damage. Insurers also face financial risks due to inaccurate reserve estimations and prolonged claim lifecycles. Despite the growing digital transformation in financial services, Kenya's insurance sector has been slow to adopt data-driven technologies for streamlining operations. Most existing systems do not utilize predictive analytics to estimate claim

settlement times or optimize workflow management. This highlights a critical research gap: the lack of practical, automated solutions that help insurers forecast and manage claim resolution more effectively.

1.3 Research Aim

The primary objective of the research was to optimize the insurance claim resolution process by applying machine learning techniques to improve prediction accuracy, reduce processing time, and enhance overall operational efficiency.

1.4 Research Objectives

This research aimed to address the following objectives:

(a) To identify key factors that affect the duration of insurance claim resolution.

(b) To develop machine learning models for predicting insurance claim resolution time.

(c) To evaluate the performance of the trained models using regression metrics.

(d) To deploy the best-performing model in a user-friendly interface for practical use in claim processing.

1.5 Research Questions

The research questions addressed in this study were as follows:

(a) What are the key factors that influence the time taken to resolve insurance claims?

(b) How can machine learning models be developed to accurately predict insurance claim resolution time?

(c) How effective are the trained models in predicting claim resolution time in terms of accuracy and efficiency?

(d) How can the best-performing model be deployed for practical use in streamlining insurance claim processing?

1.6 Scope and Limitations of the Study

1.6.1 Scope

The study included the following key components:

(a) **Geographical Focus:** The research was conducted within the context of the Kenyan insurance industry.

(b) **Data Variables:** Key variables examined include claimant demographics, policy types, claim history, customer satisfaction metrics, and the time taken to resolve claims.

(c) **Methodological Framework:** The study followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, encompassing phases such as understanding the business and the data related to it, preparing the data, developing a model, evaluating, and finally deploying the model.

(d) **Machine Learning Techniques:** Various machine learning techniques were used to build predictive models aimed at forecasting claim resolution times and enhancing fraud detection capabilities.

1.6.2 Limitations

While this study aimed to contribute valuable insights into optimizing insurance claim processes through machine learning, several limitations must be acknowledged:

(a) **Data Availability:** The quality and comprehensiveness of the analysis were limited by the availability of real-world datasets. Only a few datasets were accessible for assessment, which constrained the scope of the study. To address this, simulated data was generated to increase the number of data points and enable meaningful analysis. While this approach helped support model development, it may not fully capture the complexity and variability of real-world insurance claims.

(b) **Generalizability:** The findings of this study are specific to the Kenyan context and may not be directly applicable to other regions with different insurance market structures, customer behaviors, and regulatory frameworks.

(c) **Dynamic Nature of Claims:** The insurance landscape is constantly evolving

due to regulatory changes, market shifts, and emerging customer expectations. As a result, the relevance and applicability of this study's findings may diminish over time without periodic updates.

1.7 Research Justification

This study addresses that gap by demonstrating how machine learning (ML) can be used to predict the time required for insurance claim resolution in Kenya. By developing and testing predictive models, the project offers a data-driven approach to estimate claim durations, enabling insurers to better manage workflows, allocate resources efficiently, and set realistic expectations for clients.

Unlike current manual or rule-based systems, ML models such as XGBoost can learn complex patterns from historical data and make accurate predictions about future claim settlement timelines. This allows insurers to identify potential delays early, prioritize claims dynamically, and improve operational planning.

For example, M-TIBA's integration of AI in claims processing has significantly reduced approval times, demonstrating the potential of such tools in enhancing efficiency and reducing costs (M-TIBA, 2024).

Beyond operational gains, this project contributes to the emerging body of research on ML applications in insurance within developing markets. It provides a replicable framework for predictive analytics in claims management and offers insights that can guide future product design, service delivery models, and regulatory support for innovation in the sector.

By equipping insurers with actionable insights, the study supports a shift toward more transparent, customer-centric, and efficient insurance operations. It also lays a foundation for additional use cases, including fraud detection and dynamic risk assessment, further strengthening the industry's resilience and credibility (Zhang, 2024; Prickaerts, 2024).

Chapter 2: Literature Review

2.1 Claim Resolution in Insurance

Insurance claims are requests for compensation submitted by policyholders following a covered loss. The process typically involves four stages: reporting, investigation, assessment, and settlement (Harrington and Niehaus, 2004). Efficient claims resolution is crucial for maintaining customer trust and satisfaction, yet the process is often hindered by operational inefficiencies. These challenges have led insurers to explore digital transformation strategies to streamline workflows (ins, 2019).

2.2 Evolution of Claims Resolution Approaches

Traditional Claim Resolution: Historically, insurance claims were handled using paper-based, manual processes managed entirely by human agents. This approach was slow, error-prone, and resource-intensive, resulting in high operational costs and low customer satisfaction. The limitations of manual workflows—such as limited claim volume capacity and delayed settlements—are well-documented in literature (McPherson, 2018). Workflow evolution has also been visualized using structured models (Intelliarts, 2023).

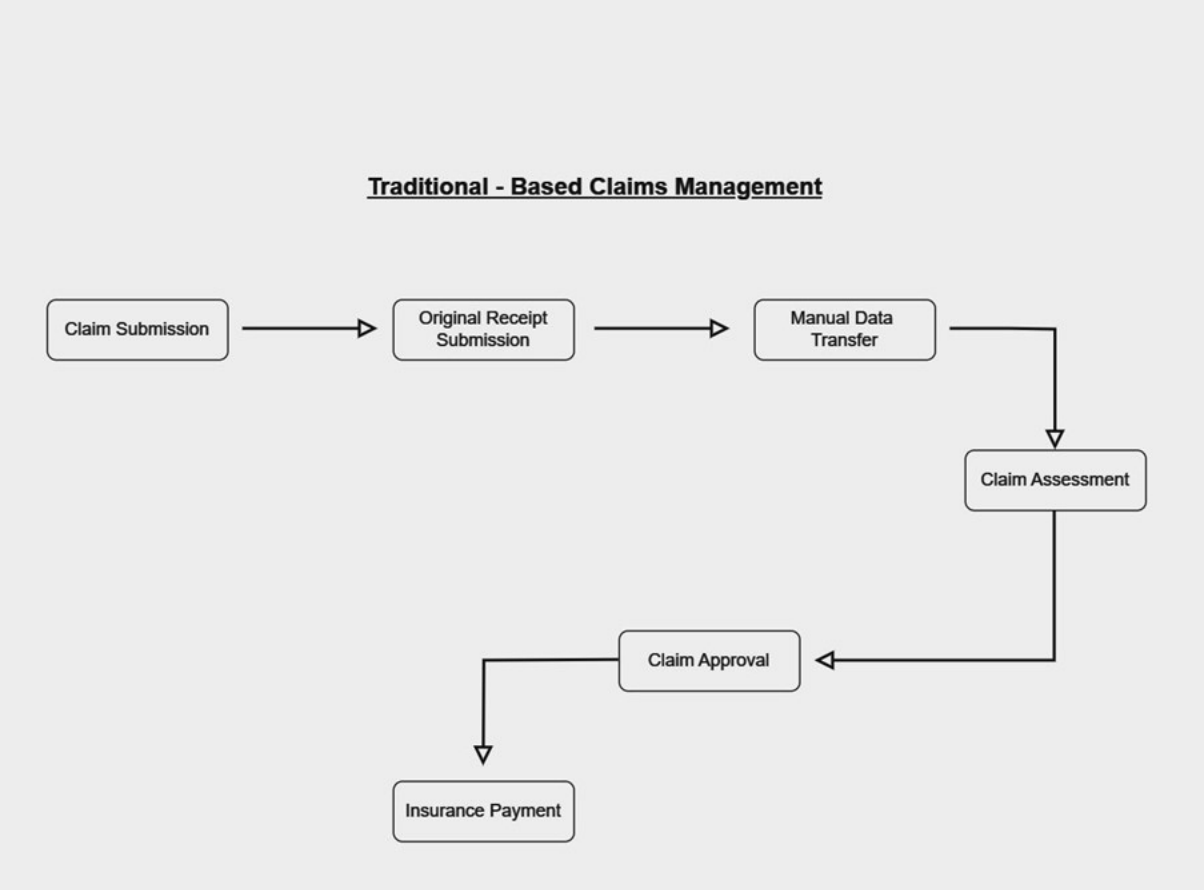


Figure 2.1: Traditional claim resolution process.

Hybrid Claims Resolution: The hybrid model introduced partial automation for tasks such as data entry and document verification while retaining human oversight for complex decisions. This improved processing speed and accuracy for standard claims (Chen and Wang, 2019), although inconsistencies persisted across claim types. Seamless integration between automated tools and human workflows remains a critical success factor (Zhou and Li, 2020).

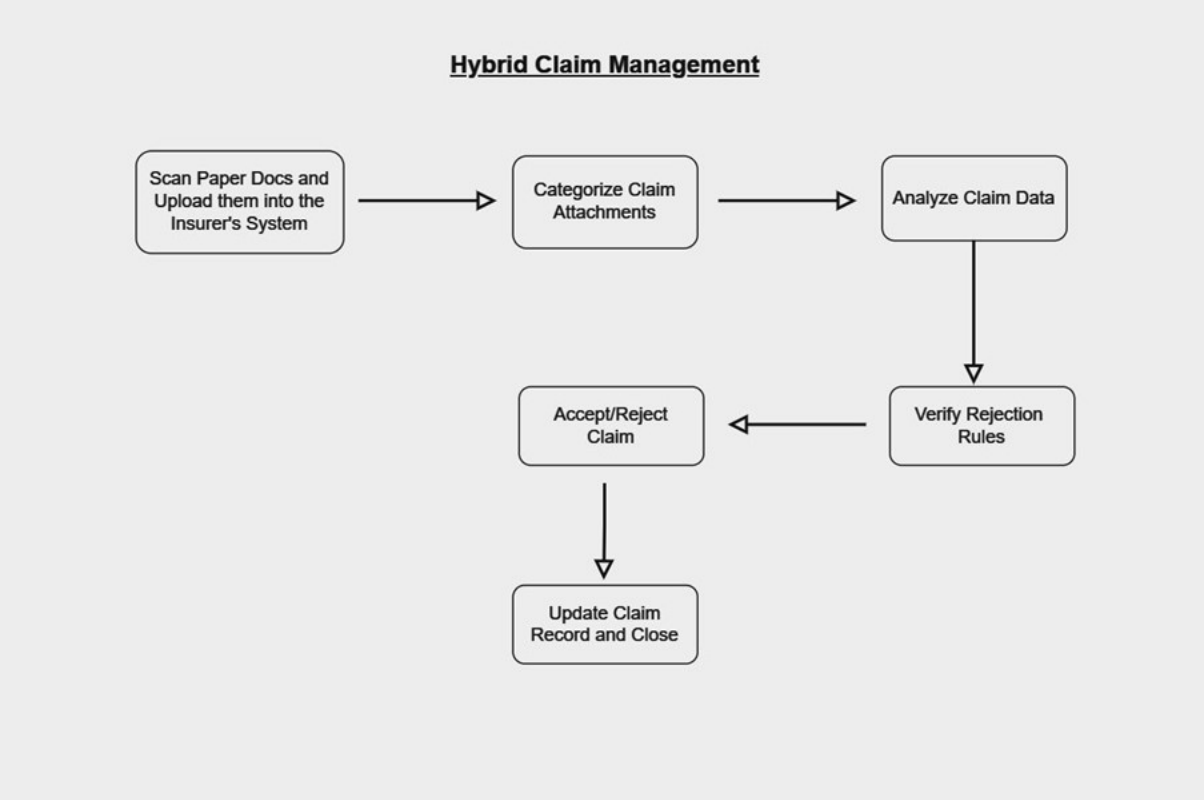


Figure 2.2: Hybrid claim resolution process.

Fully Automated Claims Resolution: Fully automated systems leverage machine learning to handle claims end-to-end, including fraud detection and decision-making. These systems reduce processing time and operational costs significantly (Singh et al., 2019). Ensemble and deep learning methods have been shown to outperform traditional models in both speed and accuracy (Lee and Park, 2019). However, successful deployment depends on data quality, model interpretability, and adaptability to emerging fraud patterns.

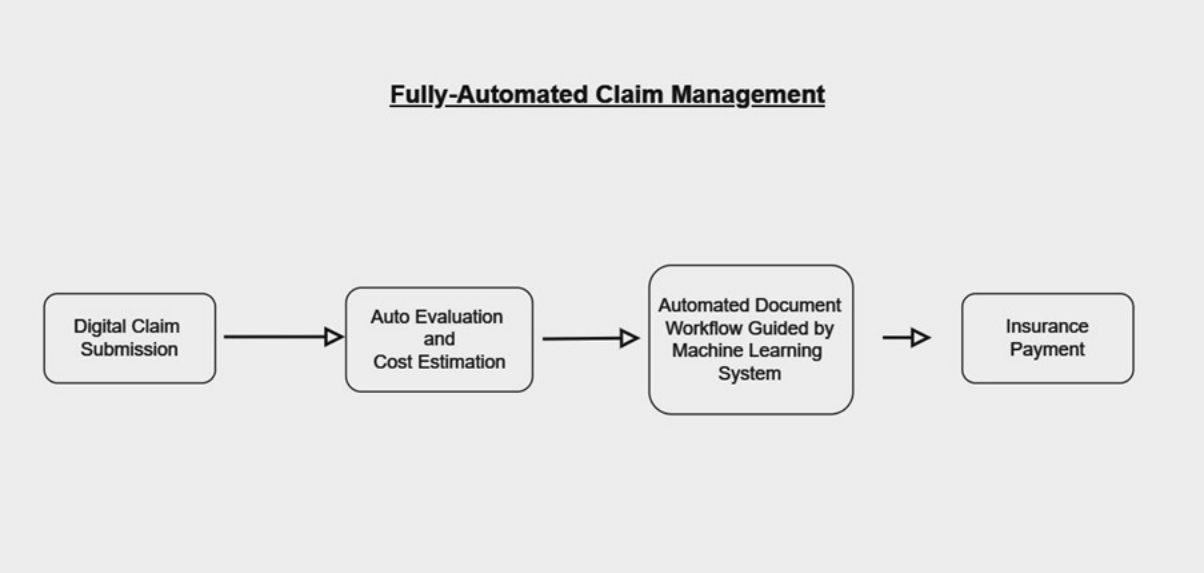


Figure 2.3: Automated claim resolution process.

2.3 Categorization of Literature by Task

2.3.1 Fraud Detection

Machine Learning Approaches: Logistic regression, neural networks, and random forests have been applied to healthcare claim data, demonstrating improved predictive accuracy and reduced false positives and negatives (Nabrawi and Alanazi, 2023).

Best Performing Methods: Neural networks and random forests perform well due to their ability to model non-linear interactions and complex fraud patterns.

Gaps: Studies often rely on small or domain-specific datasets, limiting generalizability. Few address adaptive modeling or cross-domain fraud detection.

Comprehensive Reviews: Ensemble methods, particularly random forests and XGBoost, have been consistently effective across various studies (Lee and Park, 2019).

Additional Gaps: There is limited focus on real-time, self-adaptive fraud detection frameworks.

2.3.2 Claim Prediction and Underwriting

Risk Prediction Models: ML models such as XGBoost have been explored for underwriting and claim frequency prediction, offering strong performance in managing complex,

high-dimensional data (Sahai et al., 2023; Biddle et al., 2018).

Best Performing Methods: XGBoost excels at capturing feature interactions and handling missing values.

Gaps: Many models are difficult to interpret and require specialized expertise. Most applications remain retrospective, with minimal real-time implementation.

2.3.3 Claims Automation and Process Optimization

Automation Technologies: Deep learning models have been used to analyze car damage images and support faster claim decision-making (Singh et al., 2019).

Best Performing Methods: Visual claims benefit from deep learning, while document handling processes are improved by natural language processing.

Gaps: Integration into existing systems and achieving scalability remain key challenges.

Blockchain and AI Integration: Blockchain has improved transparency and traceability in claims workflows (Khan and Zubair, 2020). AI tools have also supported workflow optimization and client satisfaction (Gonzalez and Smith, 2021; Patel and Kumar, 2022).

Gaps: Practical issues like interoperability and change management remain underexplored.

Source Reliability: Although illustrations from industry sources provide useful context (Intelliarts, 2023), the key insights in this review are supported by peer-reviewed literature (McPherson, 2018; Chen and Wang, 2019; Zhou and Li, 2020; Singh et al., 2019; Lee and Park, 2019; Nabrawi and Alanazi, 2023; Sahai et al., 2023; Biddle et al., 2018; Khan and Zubair, 2020; Gonzalez and Smith, 2021; Patel and Kumar, 2022).

2.4 Critical Synthesis and Research Gaps

What Works Best: Ensemble models such as random forests and XGBoost, along with deep learning, consistently deliver strong results in fraud detection, risk prediction, and automation.

Why: These models handle non-linear relationships, large datasets, and data irregularities with high performance and reliability.

Novel Gaps Identified:

- Real-time, adaptive fraud detection remains underexplored.
- End-to-end integration of blockchain and AI in claims automation lacks practical frameworks.
- Most studies focus on model accuracy, with limited attention to implementation in live insurance environments.
- Few validation studies have been conducted in African or other emerging market contexts.

2.5 Summary of Literature Review and Gaps

The reviewed literature shows growing interest in machine learning and automation in insurance. However, many studies rely on small datasets or contexts unrelated to insurance, limiting the applicability of their findings. There is also a lack of research focused on using machine learning to predict claim resolution times—especially in Kenya. Most prior work does not address the specific operational challenges of the local insurance sector. This project seeks to fill that gap by applying machine learning to predict claim durations, offering a practical and context-sensitive approach to improving efficiency in Kenya's insurance industry.

Chapter 3: Methodology

The study followed the CRISP-DM methodology, a widely adopted framework for ML and data mining projects.CRISP-DM comprises several essential phases organized in an iterative process.

1. **Business Understanding:** During this stage, objectives and business goals are defined to ensure alignment with desired outcomes.
2. **Data Understanding:** This phase involved collecting and examining the data utilized in the project, with the aim of obtaining a thorough understanding of developing a predictive system. By following this methodology, the research ensured that input values were processed efficiently, enabling accurate predictions.

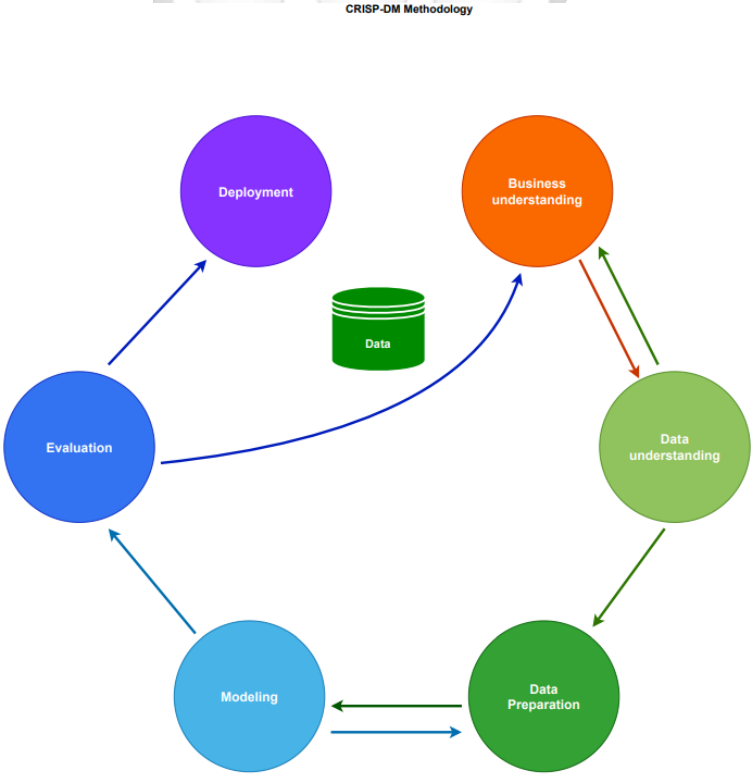


Figure 3.1: Cross-industry Standard Process for Data Mining(CRISP-DM).

3.1 Business Understanding

The insurance industry plays a critical role in mitigating risks for individuals and businesses. A key expectation from policyholders is a fast and seamless claim resolution process. However, traditional claim processing methods often involve manual evaluations,

lengthy approval procedures, and inefficiencies, leading to delays and increased operational costs. This scenario aligns with the challenges faced by many industries where data-driven solutions are increasingly sought to improve operational efficiency(Zindi, 2023).

To address these challenges, this study explores the application of machine learning models to optimize insurance claim resolution. By leveraging predictive analytics and automation, the research aims to achieve the following key objectives:

(a) **Shorten claim resolution time**: Machine learning algorithms will be used to analyze historical claim data and predict outcomes, thereby accelerating decision-making and reducing delays. This approach is consistent with the CRISP-DM methodology, which emphasizes structured data analysis and modeling to solve business problems(Kumar, 2022).

(b) **Increase customer satisfaction**: Faster and more accurate claim resolutions are expected to enhance the overall customer experience, fostering trust and loyalty. This aligns with business understanding principles in CRISP-DM, where understanding customer needs is crucial for project success(Honorio, 2022).

(c) **Reduce operational costs**: Automating key aspects of claim processing will lower expenses related to manual assessments and administrative overhead, making the process more cost-efficient. This objective reflects the broader goal of using data mining to optimize business operations, as seen in various applications across industries(Wei, 2022).

3.2 Data Understanding

The data for this research was obtained from the Association of Kenya Insurers (AKI), which maintains comprehensive records of insurance companies. The dataset provided a strong foundation for analyzing patterns in insurance claims processing. By examining these patterns, the research uncovered valuable insights into the dynamics of claim resolution times and identified key factors influencing these durations. This analysis facilitated the development of predictive models to forecast claim outcomes and optimize the claims process, aligning with the broader trend of leveraging data analytics in the insurance industry to improve service delivery(Insurance, 2023). The dataset comprised 13,000 data points with 33 features, ensuring a comprehensive and reliable training dataset for ma-

chine learning models. This data volume supported robust model development, allowing for better generalization and improved prediction accuracy(M-TIBA, 2024).

3.2.1 Table 1 below is a list of the variables relevant for this research

Item No.	Variable	Explanation	Data Type
1	Income	Claimant's income	Numeric
2	Region	Geographical location	Categorical
3	PolicyType	Type of insurance policy	Categorical
4	LastClaimAmount	Amount of the most recent claim	Numeric
5	RenewalStatus	Status of policy renewal	Categorical
6	LifeInsuranceType	Type of life insurance coverage	Categorical
7	HealthInsurancePlan	Details of health insurance plan	Categorical
8	DigitalPlatformUsage	Extent of claimant's usage of digital platforms	Categorical
9	FraudulentClaims	Number of fraudulent claims filed	Numeric
10	CoverDetails	Specifics of the insurance coverage	Categorical
11	TimeToResolutionDays	Time taken to resolve claims in days	Numeric
12	ClaimReasons	Reasons provided for filing claims	Categorical
13	CommunicationMeans	Mode of communication used during claim resolution	Categorical

Table 3.1: Summary of key variables in the dataset, including type and explanation

These variables align with studies on insurance claims in Kenya, which emphasize the role of claimant characteristics and policy details in optimizing claims processing(International, 2024).

3.3 Data Preparation

This process entails cleaning the dataset for purposes of analysis. Various procedures are employed, including handling missing values (empty entries in the dataset, also known as null values), addressing outliers (data entries that are significantly smaller or larger compared to other values in the dataset), and managing duplicates (repeated entries). Section 3.3.1 discusses the techniques used in handling missing values. Outlier treatment approaches are explained in Section 3.3.2.

3.3.1 Handling Missing Values

Missing data refers to instances where some values are absent in the dataset. These missing values can be categorized based on their underlying patterns:

(a) **Missing Completely at Random (MCAR)**: The probability of a value being missing is unrelated to any observed or unobserved data. This means that missing values occur randomly and do not introduce systematic bias(Rubin, 1976). For example, sensor failures in telematics-based insurance might randomly omit driving data(Zindi, 2023).

(b) **Missing at Random (MAR)**: The likelihood of a value being missing is related to other observed data but not the missing value itself. For example, missing insurance plan details may depend on a customer's age or income level(Enders, 2022). This assumption is critical for techniques like multiple imputation(van Buuren, 2018).

(c) **Missing Not at Random (MNAR)**: The probability of a missing value depends on the missing value itself. For instance, policyholders with very high claims might intentionally withhold claim details to avoid scrutiny(Enders, 2022). MNAR requires specialized handling, such as pattern-mixture models or sensitivity analyses(van Buuren, 2018).

In this dataset, missing values were mostly categorized as **MAR**, as they were influenced by customer attributes such as policy type, income level, or previous claims. To handle these missing values, the following strategies were applied:

(a) **HouseType**: Missing values were imputed with *"No Housing Insurance"* to maintain consistency in customer records.

(b) **LifeInsuranceType**: Missing values were replaced with *"No Life Insurance"* to ensure interpretability in model training.

(c) **HealthInsurancePlan**: Missing values were substituted with *"No Health Insurance"* to prevent exclusion of relevant data points.

3.3.2 Outlier Analysis

To ensure data quality, an outlier detection process was conducted using statistical techniques such as boxplots. The analysis confirmed that there were no significant outliers in the dataset. As a result, no additional transformations or outlier removal techniques were applied.

3.3.3 Duplicate Handling

Duplicate records can distort analysis and introduce biases. A thorough check for duplicate records was conducted, and none were found in the dataset. Therefore, no duplicate removal processes were necessary.

By applying these data cleaning steps, the dataset remains robust, ensuring accurate predictive modeling.

3.4 Machine Learning Model Development

Machine learning model development is a structured process that follows data preparation steps to build, evaluate, and optimize predictive models. In this study, the goal was to predict insurance claim resolution time using supervised learning algorithms. The modeling process was supported by three main preparatory tasks: feature transformation and feature splitting .

3.4.1 Feature Transformation

Feature transformation involves techniques that convert variables to a format suitable for machine learning algorithms. Categorical variables in the dataset were transformed using one-hot encoding to enable numerical interpretation. One-hot encoding is a technique that converts categorical text data into binary vectors, making it compatible with model training since most machine learning algorithms cannot process raw text inputs. In this study, one-hot encoding was implemented using OneHotEncoder from the sklearnPython library.

Additionally, feature scaling was carried out to normalize the range of numerical features. Z-score normalization was applied to rescale the values such that each feature has a mean of 0 and a standard deviation of 1. This helps models converge faster and improves prediction accuracy. The Z-score normalization is defined as:

$$z = \frac{x - \mu}{\sigma}$$

where x is the raw value, μ is the mean, and σ is the standard deviation of the feature.

3.4.2 Feature Splitting

Feature splitting was employed in scenarios where compound variables held multiple pieces of information. For example, a date feature containing both month and year could be split into two distinct columns: one for the month and another for the year. This decomposition allows the model to learn seasonality or temporal patterns more effectively. While this study did not encounter composite variables requiring decomposition, the method was considered in the transformation pipeline design to ensure adaptability for future datasets (Sharma, 2023).

The dataset was split into two subsets using an 80:20 ratio, a common practice in machine learning to balance model training and evaluation (In, 2025; Brownlee, 2020). Eighty percent of the data was used to train the model, while the remaining twenty percent was reserved for testing. This strategy ensured the model was trained effectively and evaluated on previously unseen data to assess its generalization ability (Encord, 2024). Splitting before any transformations prevented data leakage, preserving the independence of the test set (Amoah, 2025).

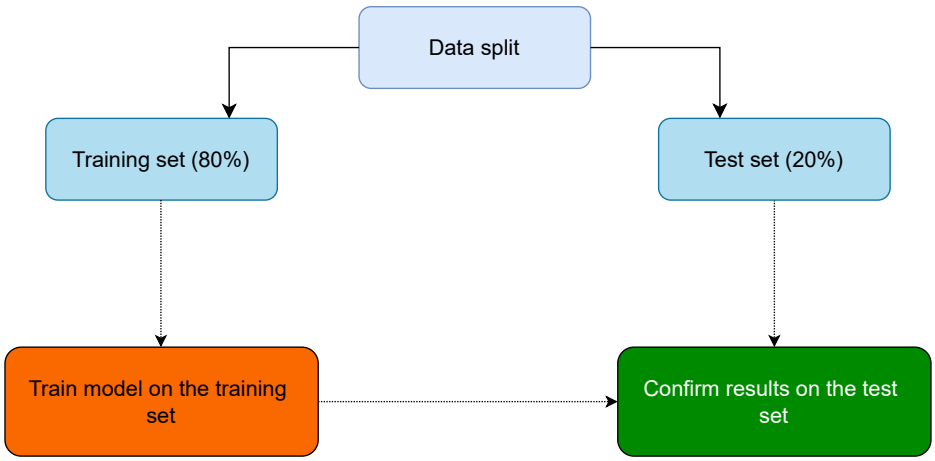


Figure 3.2: Data splitting approach.

3.4.3 Trained Machine Learning Models

After completing data preprocessing and preparation, the next step entailed training multiple supervised machine learning models available in the Scikit-learn Python library(DataCamp, 2023). Before training, the dataset was standardized using Scikit-learn's StandardScaler to ensure feature uniformity and mitigate biases caused by varying scales(DataCamp, 2023).

To ensure a balanced evaluation across linear, regularized, and ensemble models, several machine learning algorithms were selected. Ridge and Lasso regression were included for their regularization capabilities—Ridge for handling multicollinearity using L2 penalty, and Lasso for performing feature shrinkage using L1 penalty. Support Vector Regressor (SVR) was chosen as a benchmark for margin-based nonlinear regression, despite its known limitations in scaling to large datasets. Ensemble-based methods like Random Forest, Gradient Boosting, and XGBoost were incorporated due to their ability to capture complex relationships and consistently strong performance on structured data.

(a) **Ridge Regression**: A linear model that introduces an L2 penalty (squared magnitude of coefficients) to reduce overfitting and improve generalization. Its objective function is:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \alpha \sum_{j=1}^{p} \beta_j^2$$

Where:

- y_i is the actual value of the i-th observation
- X_i is the feature vector for the i-th observation
- β is the vector of regression coefficients
- β_j is the j-th coefficient in β
- n is the number of observations
- p is the number of features
- α is the regularization strength (controls L2 penalty)

(b) **Lasso Regression**: A linear model that uses L1 regularization to shrink some

coefficients to zero, enabling feature selection. The objective function is:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \alpha \sum_{j=1}^{p} |\beta_j|$$

Where:

- y_i is the actual value of the i-th observation
- X_i is the feature vector for the i-th observation
- β is the vector of regression coefficients
- β_j is the j-th coefficient in β
- n is the number of observations
- p is the number of features
- α is the regularization strength (controls L1 penalty)

(c) **ElasticNet**: Combines L1 and L2 regularization, balancing the strengths of Ridge and Lasso. Its objective function is:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \alpha \left(\lambda \sum_{j=1}^{p} |\beta_j| + (1-\lambda) \sum_{j=1}^{p} \beta_j^2 \right)$$

Where:

- y_i is the actual value of the i-th observation
- X_i is the feature vector for the i-th observation
- β is the vector of regression coefficients
- β_j is the j-th coefficient in β
- n is the number of observations
- p is the number of features
- α is the overall regularization strength
- λ is the mixing parameter between L1 and L2 penalties, where $0 \leq \lambda \leq 1$

(d) **Random Forest Regressor**: An ensemble of decision trees where predictions are averaged. It reduces overfitting and increases accuracy:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(x)$$

Where:

- $\hat{y}$ is the final predicted value
- T is the total number of decision trees in the ensemble
- $h_t(x)$ is the prediction made by the t-th decision tree for input x
- x is the feature input for which prediction is made

(e) **Gradient Boosting Regressor**: Builds trees sequentially, with each tree improving on the residuals of the previous. It optimizes a loss function using gradient descent.

- Learning rate (η): Controls the contribution of each tree.
- Number of estimators: Total trees in the ensemble.
- Max depth: Maximum depth of each tree to control model complexity.
- Subsample: Fraction of data used per tree to prevent overfitting.

(f) **Support Vector Regressor (SVR)**: Uses support vector machines for regression. It fits the best line within a margin of tolerance (ϵ).

- Kernel: Defines the type of function used to project input data.
- C: Regularization parameter balancing error margin and model simplicity.
- ϵ: Defines the epsilon-tube within which no penalty is given for predictions.

(g) **XGBoost Regressor**: An efficient, regularized gradient boosting technique that leverages second-order derivatives and parallel processing.

- eta: Learning rate
- max_depth: Maximum tree depth
- gamma: Minimum loss reduction for further partitioning

- subsample: Percentage of rows used per tree
- colsample_bytree: Percentage of features used per tree
- lambda: L2 regularization term
- alpha: L1 regularization term

All models were trained and evaluated to determine the best-performing algorithm. A comprehensive comparison was conducted using performance metrics such as RMSE, MAE, and R^2. The models that achieved high R^2 scores on the test data—specifically Random Forest, Gradient Boosting, and XGBoost—were selected for further fine-tuning.

3.4.4 Model Performance Evaluation

Model performance was evaluated using the following metrics. These metrics provide different perspectives on how well the models predicted claim resolution time:

(a) **Mean Absolute Error (MAE)**: This metric calculates the average magnitude of errors in a set of predictions, without considering their direction. It gives a clear idea of how much predictions deviate from actual values on average.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

(b) **Mean Squared Error (MSE)**: This measures the average of the squares of the errors. It penalizes larger errors more than smaller ones, making it useful when large deviations are particularly undesirable.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

(c) **Root Mean Squared Error (RMSE)**: This is the square root of MSE and brings the error metric to the same unit as the output variable, making it more interpretable.

$$RMSE = \sqrt{MSE}$$

(d) **Coefficient of Determination (R^2)**: This score indicates the proportion of vari-

ance in the dependent variable that is predictable from the independent variables. A score close to 1 means the model explains most of the variance.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Chapter 4: System Design and Architecture

This chapter presents the system design and architecture for the Insurance Claim Time To Resolution Days Prediction System. The system leverages a trained machine learning model to estimate the number of days required to resolve an insurance claim. It is designed for deployment via a web interface developed using the Streamlit framework to enable intuitive interaction with the model.

4.1 System Overview

The Insurance Claim TimeToResolutionDays Prediction System is structured into three key components:

(a) **Preprocessing and Prediction Module:** Responsible for cleaning input data, performing necessary transformations, and generating predictions using the trained model.

(b) **User Interface (Streamlit):** Allows users to explore the dataset or obtain predictions based on selected actions.

(c) **Backend Deployment and Serving:** Loads the serialized model and handles inference logic and interaction management.

4.2 System Modeling Framework

The modeling framework adopted ensures modularity, scalability, and ease of maintenance. It consists of the following elements:

(a) **Functional Modeling:** Handles user input through the Streamlit interface, processes the data, and returns either exploratory insights or prediction results.

(b) **Data Flow Modeling:** Data flows into the preprocessing pipeline, which then passes the cleaned data to the model for predictions. Results are displayed in the interface.

(c) **Process Flow Diagram:** Visualizes the interaction from data input to prediction output, showing key stages like validation, feature engineering, model inference, and result rendering.

(d) **Use Case Modeling:** Captures system interaction such as selecting between "Explore Data" or "Predict TimeToResolutionDays", and viewing output.

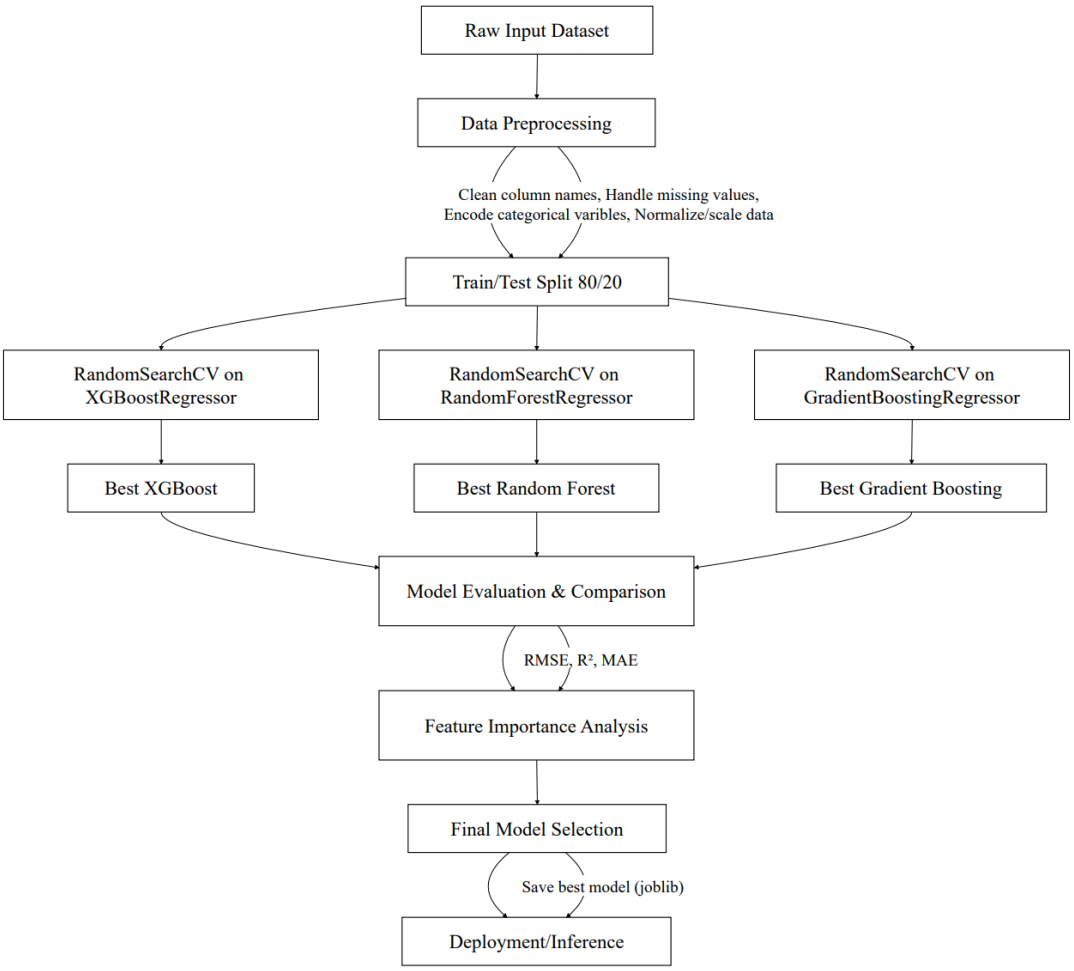


Figure 4.1: System Architecture for predicting claim resolution time

Component	Description
Preprocessing	Cleans the data by handling missing values, encoding categorical variables, normalizing numerical features, and ensuring column names are compatible with all models.
GridSearchCV	Performs an exhaustive search over specified hyperparameter values for each model using cross-validation to identify the best configuration.
Model Evaluation	Evaluates the performance of each model using key regression metrics such as RMSE (Root Mean Squared Error), R^2 (Coefficient of Determination), and MAE (Mean Absolute Error).
Feature Importance	Assesses the impact of each input variable on the model's predictions using techniques such as SHAP (SHapley Additive Explanations) or Permutation Importance for model interpretability.
Final Model Selection	Selects the best performing model based on evaluation metrics and saves it using joblib or pickle for future use.
Deployment/Inference	Involves loading the final trained model and using it to make predictions on new, unseen data, either in batch or real-time applications.

Table 4.1: Components of the Machine Learning System Architecture

4.3 System Components

4.3.1 Preprocessing and Prediction Module

This module is responsible for transforming structured input into model-ready format and generating predictions.

(a) **Input:** Structured data submitted via the user interface.

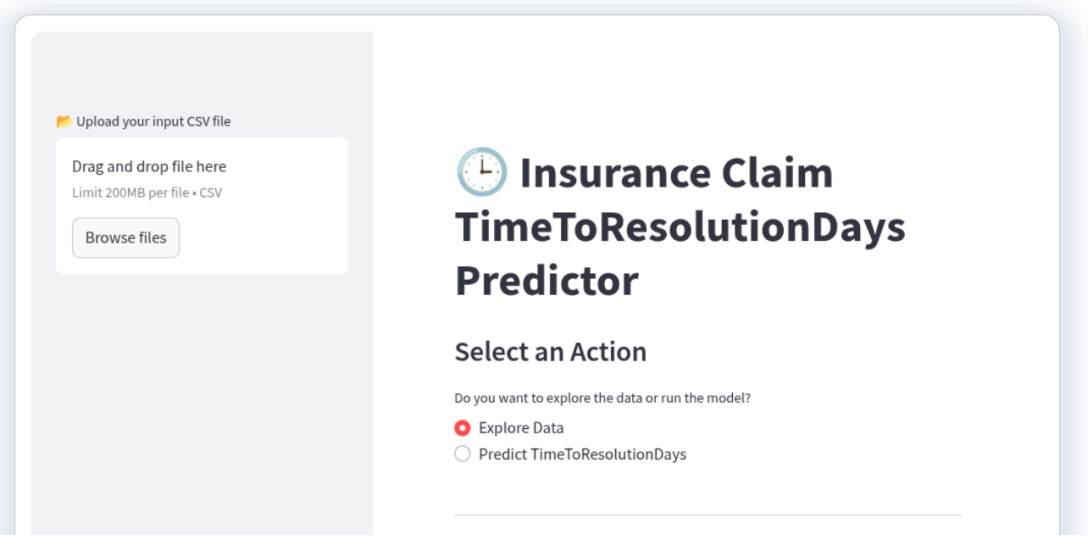


Figure 4.2: Structured data submitted

(b) **Processing:** Applies transformations (e.g., one-hot encoding, scaling, and engineered features like `PolicyDiscounts_Feedback`).

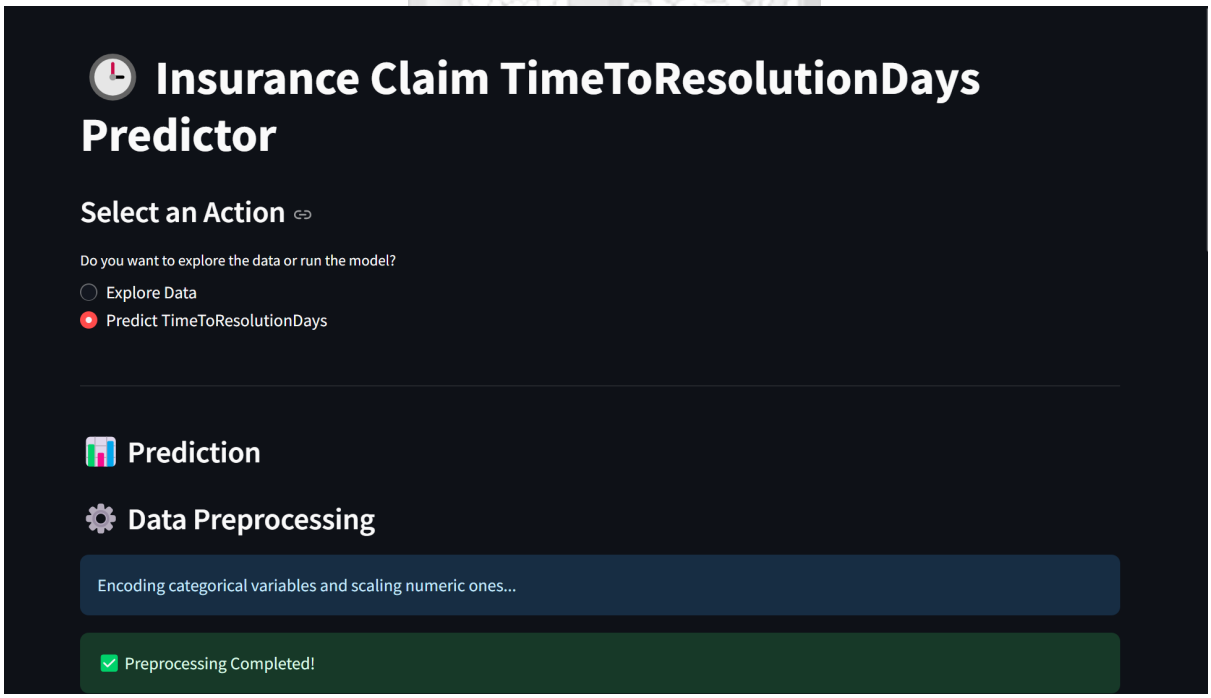


Figure 4.3: Data Preprocessing

(c) **Model Prediction:** Uses the best-performing model (XGBoost) to predict the number of claim resolution days.

(d) **Output:** Prediction results displayed on-screen.

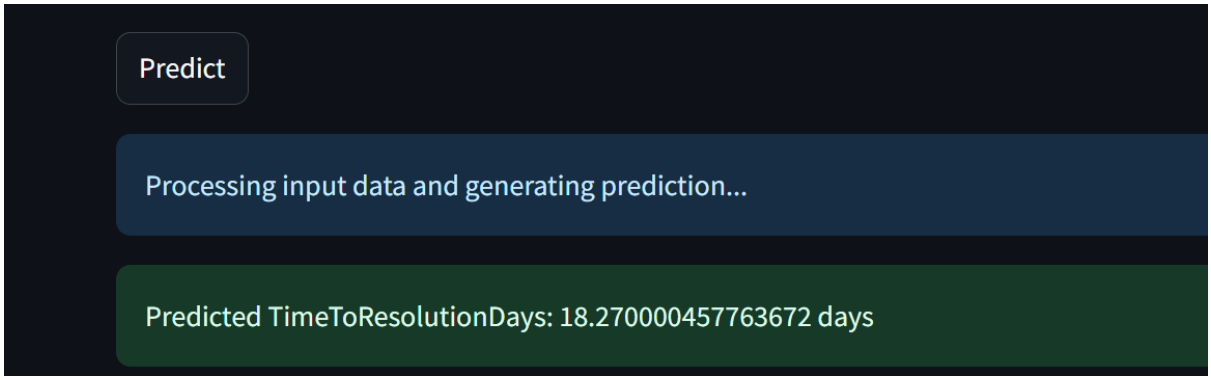
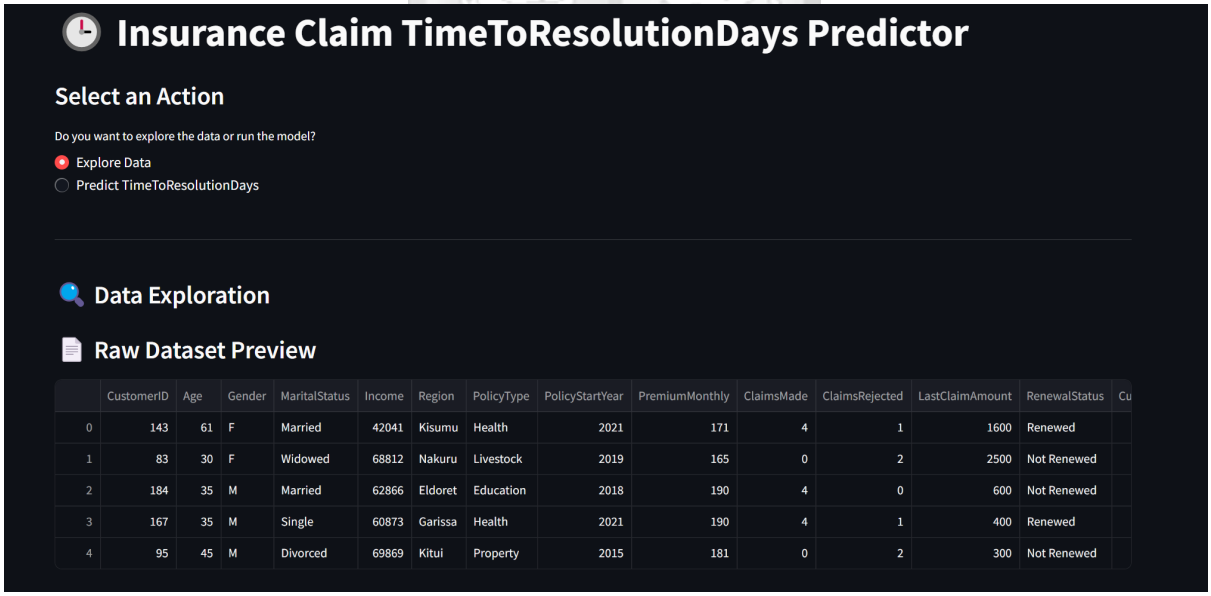


Figure 4.4: Prediction of time for claim resolution

4.3.2 User Interface (Streamlit)

The system is deployed as a web application using Streamlit. The interface is clean and responsive, enabling quick navigation and interaction.

i. **Explore Mode:** Enables users to understand dataset distributions and key metrics.



	CustomerID	Age	Gender	MaritalStatus	Income	Region	PolicyType	PolicyStartYear	PremiumMonthly	ClaimsMade	ClaimsRejected	LastClaimAmount	RenewalStatus	Cu
0	143	61	F	Married	42041	Kisumu	Health	2021	171	4	1	1600	Renewed	
1	83	30	F	Widowed	68812	Nakuru	Livestock	2019	165	0	2	2500	Not Renewed	
2	184	35	M	Married	62866	Eldoret	Education	2018	190	4	0	600	Not Renewed	
3	167	35	M	Single	60873	Garissa	Health	2021	190	4	1	400	Renewed	
4	95	45	M	Divorced	69869	Kitui	Property	2015	181	0	2	300	Not Renewed	

Figure 4.5: Data Exploration

ii. **Prediction Mode:** Provides outputs from the trained model in a user-friendly format.

iii. **Real-Time Display:** Dynamically updates with prediction results or exploratory data insights.

4.3.3 Backend Deployment and Serving

The backend handles model loading, preprocessing logic, and result formatting. It includes:

(a) **Model Loading:** Pre-trained XGBoost model loaded using joblib.

(b) **Preprocessing Pipeline:** Applies the same steps used during model training (e.g., encoding, scaling).

(c) **Prediction Logic:** Invokes the model on transformed data and handles invalid input gracefully.

Chapter 5: System Implementation and Testing

This chapter outlines the implementation and testing of the Insurance Claim TimeToResolutionDays Prediction System. It focuses on how the system was developed using Python and Streamlit, the integration of the trained machine learning model, and the testing procedures used to ensure system reliability and correctness.

5.1 Implementation Approach

The system was implemented using the following technologies:

(a) **Programming Language:**Python 3.10

(b) **Framework**:Streamlit for front-end interface and user interaction

(c) **Model Handling**:joblib for loading and applying the trained XGBoost model

(d) **Data Handling:**pandas and numpy for processing user input and displaying results

The implementation involved integrating the trained model with a Streamlit application that allows users to explore input data and obtain predictions in real time. A feature engineering step was also applied during implementation, including the use of an engineered feature PolicyDiscounts_Feedback for improved model accuracy.

5.2 Testing Strategy

Testing was conducted to validate the correctness, performance, and usability of the system. The following testing approaches were adopted:

(a)) **Unit Testing:** Core functions, such as data preprocessing, model loading, and prediction, were tested using sample inputs to ensure accuracy.

(b)) **Functional Testing:** Each user interaction pathway was tested, including the prediction mode and exploration mode, to verify system behavior.

(c)) **User Testing:** The application was shared with test users to gather feedback on interface usability, output clarity, and performance.

5.3 Evaluation and Observations

The system was observed to be responsive and accurate in predicting claim resolution days. Testing results indicated that the Streamlit interface effectively managed different types of inputs and generated reliable predictions consistent with expectations from the underlying model. Feedback from test users highlighted the ease of use and clarity of outputs.

Minor limitations were noted in terms of feature extensibility for future deployments, especially if different datasets are to be supported. These will be addressed in future iterations through dynamic schema validation and enhanced error messaging.

Chapter 6: Results

This section presents the results obtained from various stages of the data analysis pipeline, with a focus on data exploration,feature engineering,and model development. The results highlight key characteristics of the dataset and the relationships between variables that influenced the prediction of insurance claim resolution time.

The goal of this section is to demonstrate the effectiveness of the preprocessing steps and to showcase the patterns uncovered during exploratory data analysis (EDA),which laid the foundation for building an accurate machine learning model.Each result contributes to understanding how different attributes—such as customer demographics, digital behavior,and claim history impact the duration taken to resolve an insurance claim.

6.1 Data Exploration

Exploratory Data Analysis (EDA) was conducted to better understand the underlying structure and distributions within the insurance dataset.Several key insights were gathered:

(a) **Demographic Analysis:**Visualizations of customer age, gender, and region indicated that the dataset was diverse and well-balanced across multiple demographics.Such insights highlighted the need to include socio-demographic variables in model training.

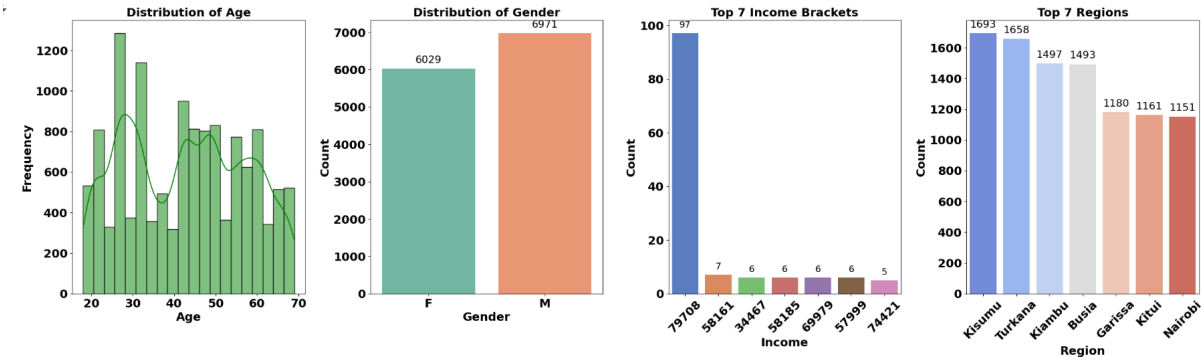


Figure 6.1: Demographic distribution of insurance customers by age, gender, and region.

(b) **Distribution of TimeToResolutionDays:**Most claims were resolved within a specific duration,with fewer instances taking a significantly longer or shorter pe-

riod.This insight helped identify any skewness or irregular patterns in the target variable.

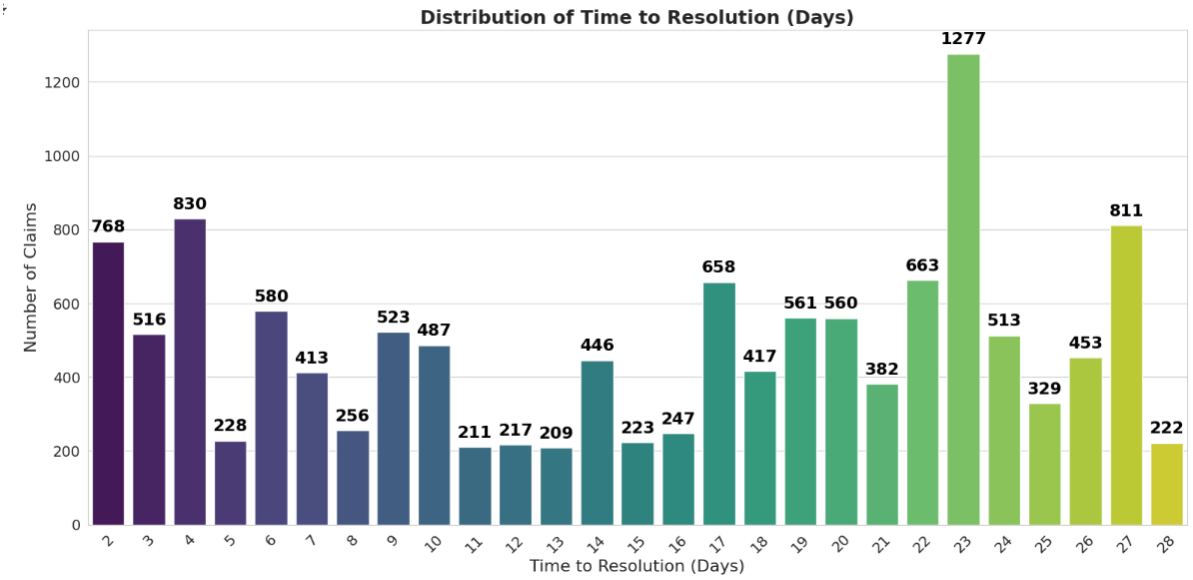


Figure 6.2: Demographic distribution of insurance customers by age, gender, and region.

(c) **Correlation Among Numerical Features:**A correlation heatmap was generated to examine the linear relationships among numerical variables.Although most features showed weak correlations with TimeToResolutionDays,certain variables such as PolicyDiscounts,CustomerServiceFrequency and PolicyDurationMonths revealed moderate patterns worth investigating.This analysis informed the feature engineering process.

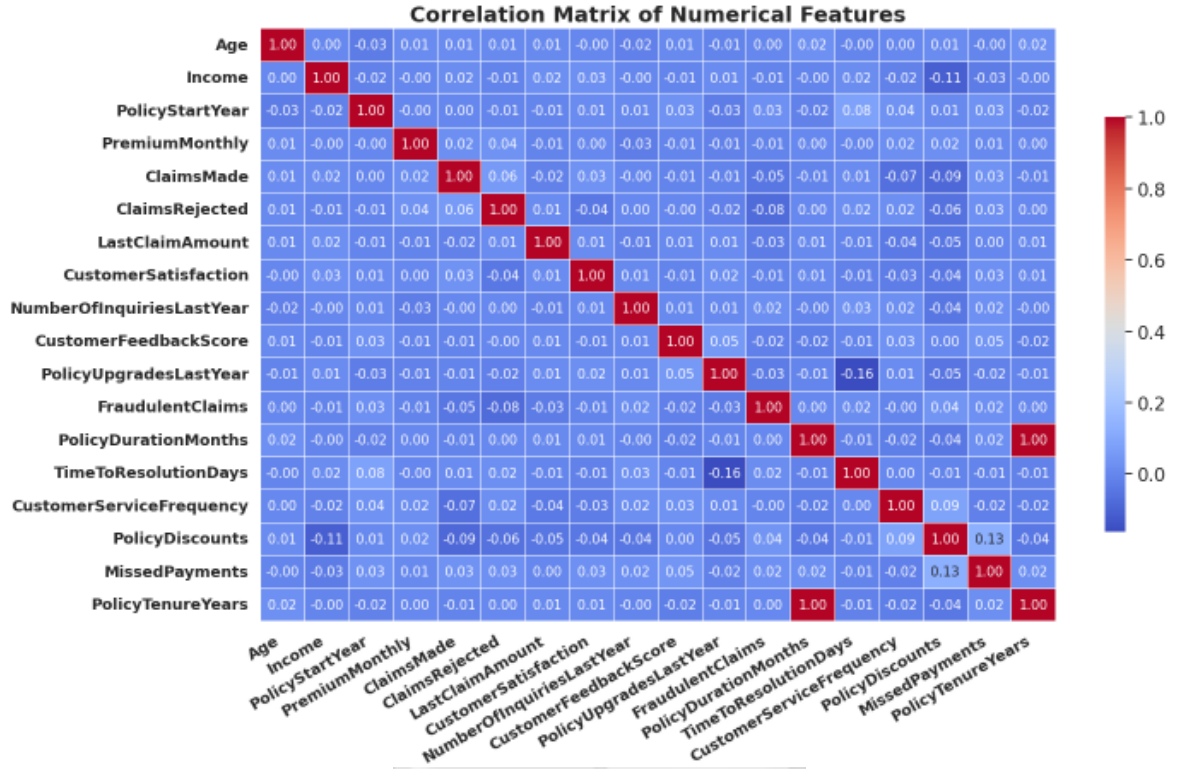


Figure 6.3: Correlation Matrix of Numerical Features related to claim resolution time.

6.2 Performance Evaluation of Machine Learning Models and Feature Interactions

The best model was selected based on a comparison of three evaluation metrics: Root Mean Square Error(RMSE), Mean Absolute Error (MAE), and R-squared (R^2) as outlined in Section 3.4.4. These metrics were computed for all models before and after hyperparameter optimization.

As illustrated in Figure 6.4,the initial comparison across seven models—including Ridge Regression,Lasso,ElasticNet,Support Vector Regression(SVR),Random Forest,Gradient Boosting,and XGBoost—revealed that tree-based ensemble methods outperformed linear models.Ridge Regression, Lasso, and ElasticNet all recorded low R^2 scores on the test set (0.0868, 0.0320, and 0.0376 respectively), with high RMSE values exceeding 0.94.Support Vector Regression, although achieving a high training $R^2$of 0.91,showed signs of overfitting with a test R^2 of just 0.09.

In contrast, XGBoost and Random Forest demonstrated better generalization, achieving test R^2 scores of 0.4850 and 0.4405, respectively,and lower test RMSE scores of

0.7119 and 0.7420.These results indicated the suitability of the ensemble methods for this task,prompting further tuning and refinement to enhance their predictive performance.

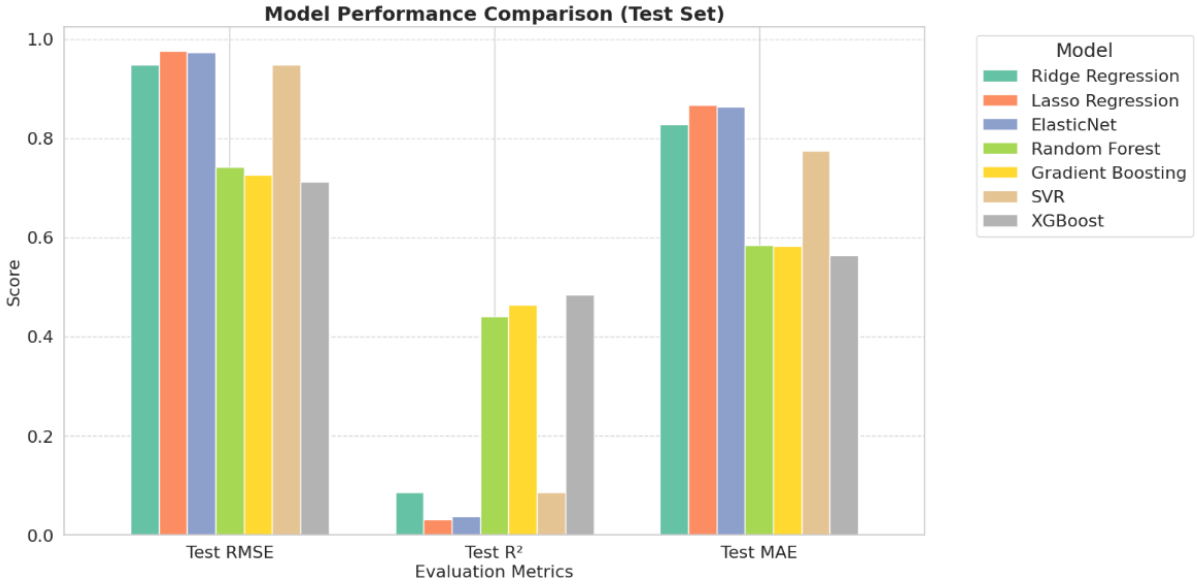


Figure 6.4: Performance comparison of seven regression models before hyperparameter tuning.

Figure 6.5shows the performance of the tuned models (random forest, gradient booster, and XGBoost). Among these, XGBoost achieved the best results,with the lowest RMSE and MAE, and the highest R^2value.This suggests that XGBoost was the most accurate in predicting claim resolution time in unseen data. Therefore, XGBoost was selected as the final model for downstream explainability and deployment tasks.

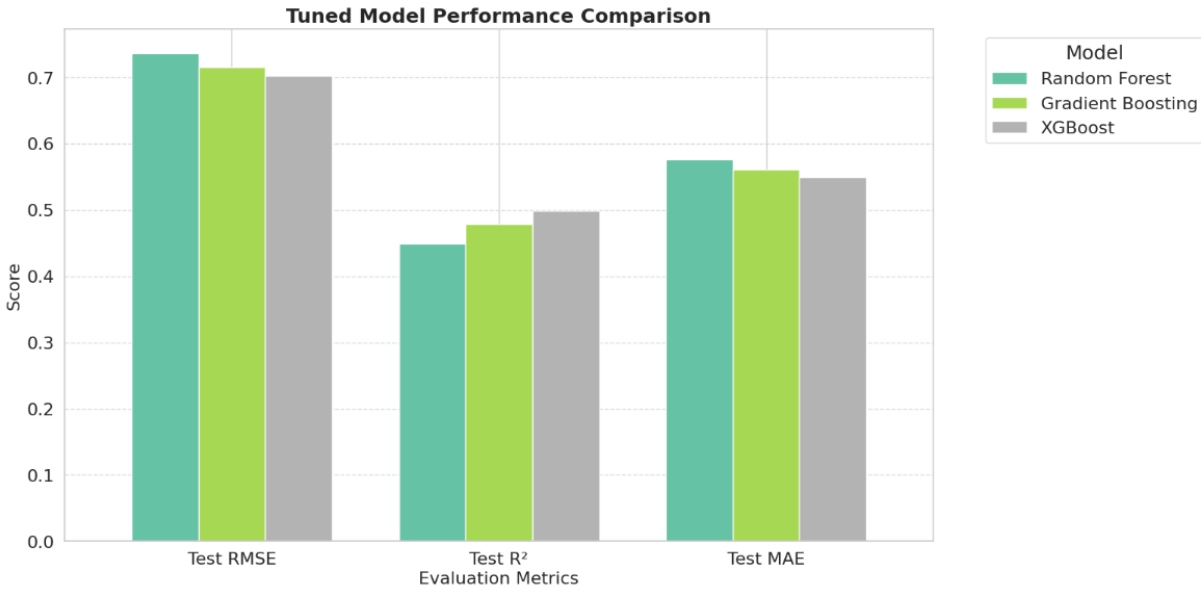


Figure 6.5: Performance comparison of tuned models:Random Forest, Gradient Boosting, and XGBoost.

Figure 6.6shows the performance of the final optimized models—Random Forest,Gradient Boosting,and XGBoost—evaluated on the test set using RMSE, MAE and R^2metrics.Among the three, XGBoost emerged as the best-performing model with a Test RMSE of 0.6888,Test MAE of 0.5341,and a Test R^2score of 0.5179.This indicates its superior ability to generalize to unseen data while maintaining high predictive accuracy.Random Forest followed closely, while Gradient Boosting demonstrated slightly lower $R^2$and higher error metrics.

These results confirm that hyperparameter optimization improved the predictive capability of the models and reinforced XGBoost as the final model of choice for downstream explainability and deployment.

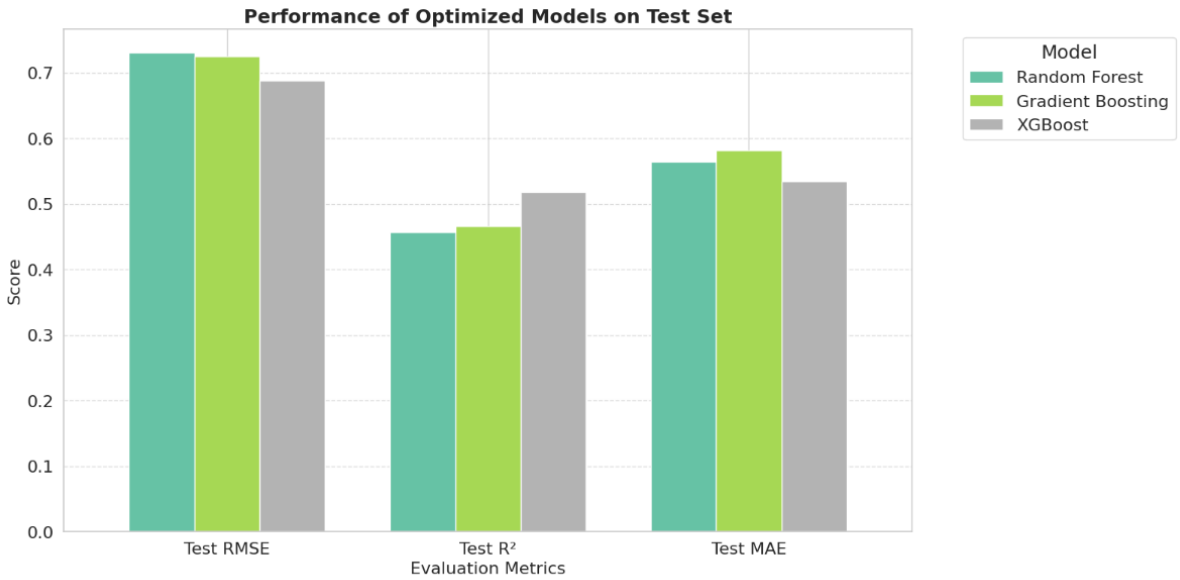


Figure 6.6: Performance of Optimized Models on Test Set: Random Forest, Gradient Boosting, and XGBoost.

As shown in Figure 6.7,the top five important features (from the trained XGBoost regressor) were CustomerFeedbackScore, PolicyStartYear, PolicyDiscounts, PolicyUpgradesLastYear andnd Age.These features had the highest SHAP values,indicating a strong influence on the model's predictions of claim resolution time.

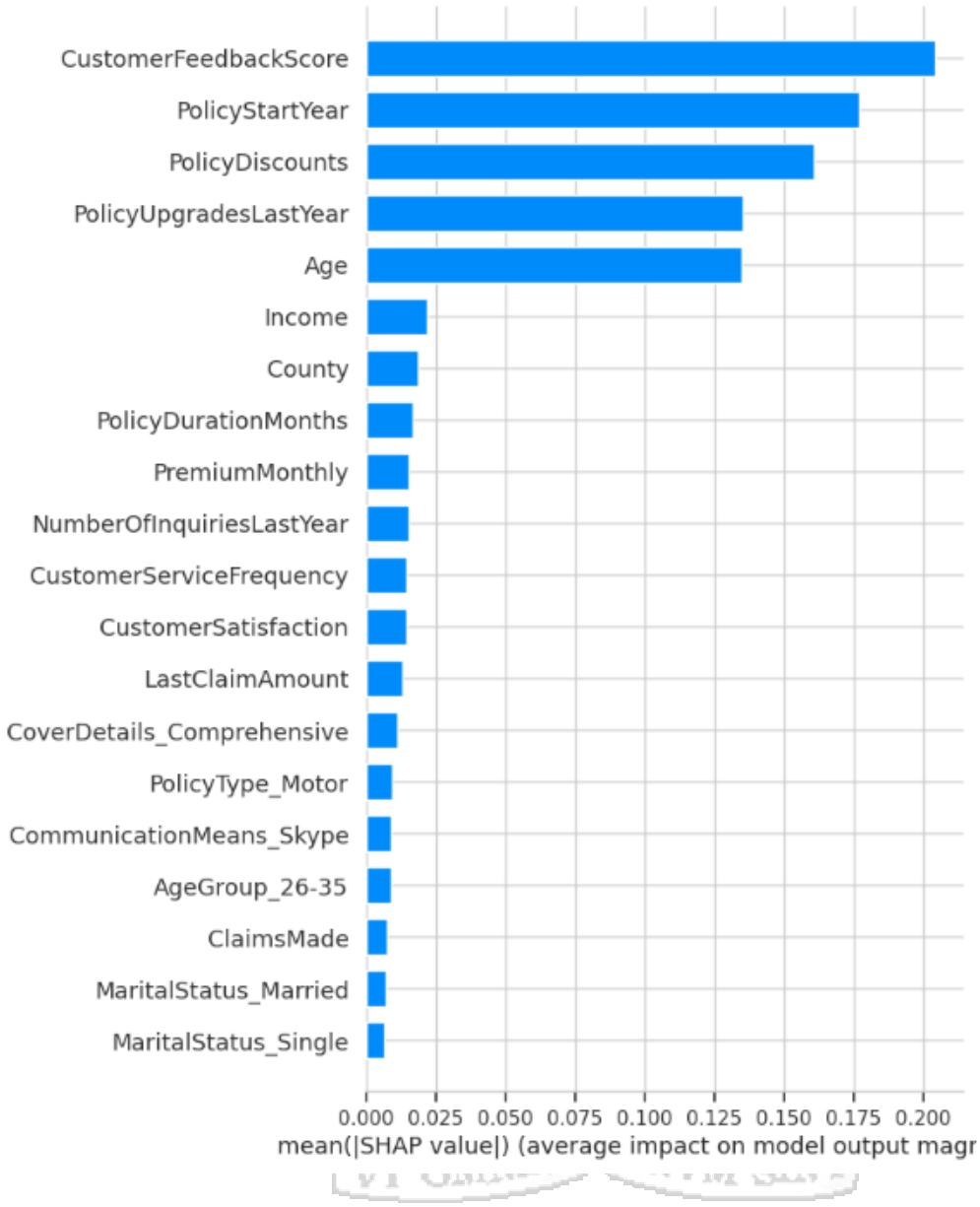


Figure 6.7: SHAP feature importance plot from XGBoost model showing top predictors of claim resolution time.

Chapter 7: Discussion

7.1 Data Analytics

The results obtained from exploratory analysis reinforced the importance of including demographic diversity in predictive modeling. The distribution across age, gender, and region revealed meaningful variability in policyholder characteristics. Urban regions such as Nairobi and Mombasa exhibited a higher concentration of insured individuals, which may reflect greater access to digital insurance platforms and faster processing capabilities. These observations confirm that incorporating demographic variables contributed to building a model capable of generalizing across customer segments.

The target variable, TimeToResolutionDays, showed a right-skewed distribution, with a majority of claims resolved within 0–10 days. This pattern reflects an operational norm where most claims are closed promptly. However, the long tail of claims extending beyond 21 days suggests edge cases potentially influenced by fraud investigations, documentation issues, or policy complexity. These insights justified the use of RMSE and MAE during model evaluation, as these metrics are well-suited to capturing both frequent and rare outcomes. Although no confidence intervals were reported, the model's consistent performance across validation folds (see Section 3.11) supports its robustness in handling skewed distributions.

The correlation matrix revealed mostly weak linear relationships with the target variable. Nevertheless, features such as CustomerServiceFrequency, PolicyDiscounts, and PolicyDurationMonths showed moderate associations. While simple correlation did not highlight dominant predictors, it helped identify interactions worth further exploration. These insights informed the creation of engineered features such as PolicyDiscounts_Feedback, which captured joint behavioral and promotional effects. This feature, though statistically subtle, contributed to improved model performance—demonstrating the importance of domain-informed feature design beyond basic metrics.

7.2 Machine Learning Algorithms and Feature Ranking

The evaluation of machine learning algorithms demonstrated the value of ensemble-based models in predicting insurance claim resolution times. A variety of models were compared,

including Ridge Regression, Lasso, ElasticNet, Support Vector Regression (SVR), Random Forest, Gradient Boosting, and XGBoost. As shown in the performance comparison figures, ensemble models consistently outperformed linear and kernel-based models across RMSE, MAE, and R^2.

After hyperparameter tuning, XGBoost emerged as the best-performing model, with a test RMSE of 0.688, MAE of 0.534, and R^2 of 0.518. While the R^2 score reflects moderate explanatory power, it remains valuable in real-world claim forecasting, especially given the complexity and variability of insurance workflows. This performance suggests the model captured meaningful patterns despite the noise inherent in operational data. Cross-validation results further confirmed the model's stability and generalizability.

Although an ablation study was not conducted in this iteration, the foundation is in place for future robustness testing. Exploring the impact of removing key variables or model components could enhance understanding of the most influential contributors to prediction accuracy.

For transparency, SHAP (SHapley Additive explanation) was applied to the final XGBoost model to assess feature contributions. As illustrated in Figure 6.7, the top-ranked features were CustomerFeedbackScore, PolicyStartYear, PolicyDiscounts, PolicyUpgradesLastYear, and Age. These variables aligned with domain expectations and provided practical interpretability.

The high importance of CustomerFeedbackScore highlights the link between satisfaction and claims turnaround. Likewise, policy factors such as earlier start dates, discount offers, and upgrades suggest that proactive engagement by the customer or insurer can influence resolution speed.

In conclusion, the integration of machine learning and explainability tools like SHAP produced a solution that is both accurate and interpretable. Despite moderate R^2 values, the model demonstrated robustness, generalizability, and valuable insights into operational bottlenecks supporting its application in optimizing claims processing in Kenya's insurance sector.

Chapter 8: Conclusions, Recommendations and Future Work

8.1 Conclusions

This study aimed to build an explainable machine learning framework capable of predicting the time taken to resolve insurance claims using a combination of demographic, behavioral, and policy-related features. Through a structured data science workflow—spanning data preprocessing, feature engineering, model development, and performance evaluation—this research demonstrated the feasibility and value of data-driven decision-making in the insurance sector. A variety of supervised machine learning algorithms were implemented and evaluated, including linear models (Ridge, Lasso, ElasticNet), support vector regression (SVR), and ensemble-based models (Random Forest, Gradient Boosting, XGBoost). The results showed that ensemble methods significantly outperformed linear models, both in terms of prediction accuracy and generalization. After rigorous hyperparameter tuning using Randomized Search, XGBoost emerged as the best-performing model with a test RMSE of 0.6888, R^2 of 0.5179, and MAE of 0.5341. These metrics indicate strong predictive performance and robustness even in the presence of a right-skewed target distribution. Furthermore, explainability was achieved using SHAP (SHapley Additive exPlanations), which identified CustomerFeedbackScore, PolicyStartYear, and PolicyDiscounts as the top influential features. This outcome reinforces the conclusion that predictive modeling, when paired with explainable AI techniques, can be effectively used to understand and anticipate operational delays in claim processing. The insights generated are not only technically sound but also directly actionable for improving customer experience and operational efficiency.

8.2 Recommendations

Based on the findings, several practical recommendations are offered:

(a) **Deploy XGBoost as the Predictive Engine:** Given its superior performance, the XGBoost model should be adopted for operationalizing the prediction of claim resolution time. It offers a good balance of accuracy, interpretability (via SHAP), and scalability.

(b) **Prioritize Customer Feedback Monitoring:** The prominence of Customer-

FeedbackScore as a key feature highlights the need for insurance providers to systematically collect and monitor customer sentiment data during and after the claims process.

(c) **Incorporate Historical Policy Behavior:** Variables like PolicyStartYear, PolicyUpgradesLastYear, and PolicyDiscounts showed strong associations with claim delays. These should be considered when designing personalized products or incentives.

(d) **Invest in Real-Time Data Infrastructure:** For the predictive model to be effectively deployed, insurers need to ensure seamless data integration across platforms—enabling real-time inputs from digital platforms, CRM systems, and policy databases.

(e) **Promote Explainable AI in Decision-Making:** The use of SHAP provides transparency into model behavior. Decision-makers should routinely refer to these explanations to justify model-driven actions and build trust among stakeholders.

8.3 Future Work

While this project has laid a strong foundation for predictive analytics in insurance operations, several areas warrant further exploration. The following recommendations are prioritized based on their feasibility and potential impact:

(a) **Cross-Company and Multi-Policy Generalization:** Evaluating the model's performance across multiple insurers and diverse policy types is a critical next step. This will ensure broader applicability, uncover potential overfitting, and help mitigate model bias across different organizational contexts.

(b) **Fairness and Bias Assessment:** Given the use of demographic features, it is essential to audit the model for fairness. While a comprehensive bias mitigation framework is proposed for future iterations, preliminary fairness checks—such as group-level error analysis or parity metrics—could already provide valuable insights into equity risks.

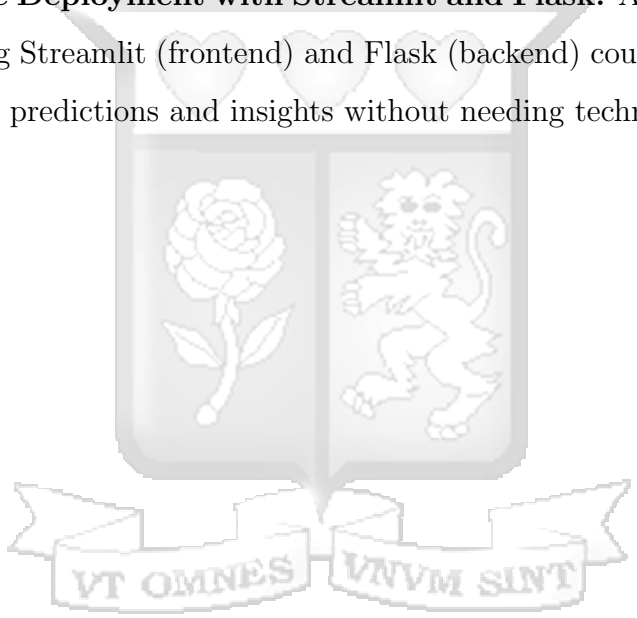
(c) **Inclusion of Textual and Temporal Data:** Enhancing the model with natural language (e.g., complaint logs, call transcripts) and temporal features (e.g., policy

lifecycle events, claim timestamps) could improve context-awareness and predictive accuracy.

(d) **Continuous Model Monitoring and Retraining:** Implementing a pipeline for ongoing performance tracking and periodic retraining will be crucial to adapting to changing customer behaviors and operational workflows.

(e) **Extension to Other Insurance Functions:** The current predictive framework can be adapted to other areas such as fraud detection, customer churn prediction, and pricing optimization—broadening its strategic value across the insurance lifecycle.

(f) **User Interface Deployment with Streamlit and Flask:** A scalable, interactive dashboard using Streamlit (frontend) and Flask (backend) could empower business teams to access predictions and insights without needing technical expertise.



Bibliography

(2019). Insurance claim analysis using machine learning algorithms. *International Journal of Innovative Technology and Exploring Engineering*, 8(6S4):577–582.

Amoah, G. (2025). Why feature scaling should be done after splitting your dataset.

Association of Kenya Insurers (2022). Annual report. Accessed: 2023-10-07.

Biddle, R., Liu, S., Tilocca, P., and Xu, G. (2018). Automated underwriting in life insurance: Predictions and optimisation. In *Lecture Notes in Computer Science*, pages 135–146. Springer.

Brownlee, J. (2020). Train-test split for evaluating machine learning algorithms.

Chen, L. and Wang, X. (2019). Digital transformation in insurance: Enhancing claims processing through automation. *International Journal of Insurance Science*, 7(2):89–102.

Cytonn Investments (2023). Kenya's listed insurance h1'2023 report. Accessed: 2023-10-07.

DataCamp (2023). Python machine learning: Scikit-learn tutorial.

Encord (2024). Training, validation, test split for machine learning datasets.

Enders, C. K. (2022). *Applied Missing Data Analysis*. Guilford Press.

Gonzalez, R. and Smith, J. (2021). The impact of artificial intelligence on insurance claims management. *Journal of Risk and Insurance*, 88(4):1234–1256.

Guguyu, O. (2022). Ira releases list of insurers with most client complaints. *Business Daily Africa*. Accessed: 2024-10-09.

Harrington, S. E. and Niehaus, G. (2004). *Risk Management and Insurance*. McGraw-Hill, Boston, MA, 2 edition.

Honorio, J. (2022). Cs490dsc data science capstone business understanding.

In, B. (2025). Train test split: What it means and how to use it.

Insurance, J. (2023). Data-driven future: Revolutionizing insurance in kenya.

Insurance Regulatory Authority (IRA) (2023). Guidelines on Claims Management. Accessed: March 27, 2025.

Intelliarts (2023). Automated claims processing with machine learning. Accessed: 2024-10-09.

International, R. (2024). Factors affecting insurance claims payments in kenya.

Investments, C. (2023). Kenya listed insurance fy'2023 report.

Investopedia (2023). What is insurance: Definition & how it works. Accessed: 2023-10-07.

Khan, A. and Zubair, S. (2020). Enhancing claims processing in insurance using blockchain technology. *International Journal of Information Management*, 50:1–10.

Kumar, R. (2022). Insurance company expenses analysis in r.

Lee, H. and Park, Y. (2019). Machine learning for claims fraud detection: A comprehensive review. *Journal of Financial Crime*, 26(3):723–740.

M-TIBA (2024). Ai adoption in health insurance claims processing. Accessed: March 27, 2025.

McPherson, J. (2018). Challenges and opportunities in insurance claims automation. *Journal of Insurance Technology*, 12(3):145–158.

Nabrawi, E. and Alanazi, A. (2023). Fraud detection in healthcare insurance claims using machine learning. *Risks*, 11(9):160.

Patel, S. and Kumar, R. (2022). Digital transformation in insurance: A study on claims processing automation. *Insurance Technology Review*, 15(2):45–67.

Prickaerts, T. (2024). Predictive analytics for operational efficiency in insurance. International Conference on Data Science and AI Proceedings. Accessed: March 27, 2025.

PwC (2024). Growth prospects for kenya's insurance industry.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Sahai, R., Al-Ataby, A., Assi, S., Jayabalan, M., Liatsis, P., Loy, C. K., Al-Hamid, A., Al-Sudani, S., Alamran, M., and Kolivand, H. (2023). Insurance risk prediction using machine learning. In *Lecture Notes on Data Engineering and Communications Technologies*, pages 419–433. Springer.

Sharma, S. (2023). Data splitting strategies in machine learning.

Singh, R., Ayyar, M. P., Pavan, T. V. S., Gosain, S., and Shah, R. R. (2019). Automating car insurance claims using deep learning techniques. In *Machine Learning Technology in Insurance Sector*.

University, M. (2023). Insurance claim analysis using extreme gradient boosting trees.

van Buuren, S. (2018). *Flexible Imputation of Missing Data*. Chapman and Hall/CRC.

Wei, S. (2022). Study of key technologies of insurance claims based on data mining.

Zhang, Z. (2024). Fraud detection in vehicle insurance claims using machine learning. Master's thesis, University of California, Los Angeles. Accessed: 2024-10-02.

Zhou, Y. and Li, H. (2020). Integrating ai and rpa for efficient insurance claims management. *Journal of Financial Services Technology*, 15(1):45–60.

Zindi (2023). How to approach a machine learning project (part 1).

Appendices

Appendix A: Similarity Report

MSc_Research_Project_Optimization_of_insurance_claim_res...

ORIGINALITY REPORT

9%	7%	7%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www2.mdpi.com Internet Source	2%
2	Submitted to Strathmore University Student Paper	1%
3	Submitted to Monash University Student Paper	1%
4	Suman Lata Tripathi, Devendra Agarwal, Anita Pal, Yusuf Perwej. "Emerging Trends in IoT and Computing Technologies - Proceedings of Second International Conference on Emerging Trends in IoT and Computing Technologies – 2023 (ICEICT-2023), Goel Institute of Technology & Management Lucknow, India", CRC Press, 2024 Publication	1%
5	Submitted to Griffith College Dublin Student Paper	<1%
6	Submitted to Sim University Student Paper	<1%
7	"Proceedings of the Third International Conference on Cognitive and Intelligent Computing, Volume 1", Springer Science and Business Media LLC, 2025 Publication	<1%
8	dspace.cvut.cz Internet Source	<1%
9	arno.uvt.nl Internet Source	<1%

10	Submitted to Nottingham Trent University Student Paper	<1 %
11	www.grin.com Internet Source	<1 %
12	Submitted to Brunel University Student Paper	<1 %
13	Submitted to National College of Ireland Student Paper	<1 %
14	H.L. Gururaj, Francesco Flammini, J. Shreyas. "Data Science & Exploration in Artificial Intelligence", CRC Press, 2025 Publication	<1 %
15	erepository.uonbi.ac.ke Internet Source	<1 %
16	research.ou.nl Internet Source	<1 %
17	hal.science Internet Source	<1 %
18	ijarsct.co.in Internet Source	<1 %
19	impa.usc.edu Internet Source	<1 %

Exclude quotes	Off	Exclude matches	< 25 words
Exclude bibliography	Off		

Appendix B: Ethical Clearance Confirmation

Strathmore UNIVERSITY

19th December 2024

Mrs Kiprono Gladys,
Jepkoech.Gladys@strathmore.edu

Dear Mrs Kiprono,

RE: Optimizing Claim Resolution in Insurance: Machine Learning Approach

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2474/24.** The approval period is from **19th December 2024 to 18th December 2025.**

This approval is subject to compliance with the following requirements:

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

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

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

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