PREDICTING LAPSE RATE IN LIFE INSURANCE USING MACHINE LEARNING ALGORITHMS: A CASE STUDY OF SANLAM NAMIBIA

By

Nelson Mitchell Kamati

222143851

Submitted in partial fulfilment of the requirements for the degree of

Master of Data Science

in the

Faculty of Computing and Informatics

Department of Informatics

At the

NAMIBIA UNIVERSITY OF SCIENCE AND TECHNOLOGY

Supervisor:

Dr Richard Maliwatu

META DATA

TITLE: PREDICTING LAPSE RATE IN LIFE INSURANCE USING MACHINE LEARNING ALGORITHMS: A CASE STUDY OF SANLAM NAMIBIA

STUDENT NAME: NELSON MITCHELL KAMATI

SUPERVISOR: DR RICHARD MALIWATU

DEPARTMENT: INFORMATICS

QUALIFICATION: MASTER OF DATA SCIENCE

SPECIALISATION: MACHINE LEARNING

KEYWORDS: SUPPORT VECTOR MACHINE, GRADIENT BOOST, RANDOM FOREST, LOGISTIC REGRESSION, LAPSE, MACHINE LEARNING, LIFE INSURANCE, LAPSE RISK

TYPE OF RESEARCH: EXPERIMENTAL RESEARCH

METHODOLOGY: QUANTITATIVE

STATUS: THESIS

SITE: NAMIBIA UNIVERSITY OF SCIENCE AND TECHNOLOGY

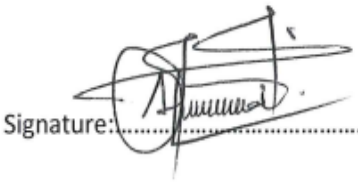
DECLARATION

I, **Nelson Mitchell Kamati,** hereby declare that the work included in the paper for the Master of Data Science, titled "Predicting Lapse Rate in Life Insurance Using Machine Learning Algorithms: A Case Study of Sanlam Namibia," represents my own original work, and I have fully acknowledged and referenced all sources used or quoted

I further confirm that I have not submitted this work, either in full or in part, at any university or educational institution for the purpose of earning a degree.

I also declare that, in accordance with the institution's guidelines, I will give full credit to all sources of information used for the research.

I confirm that the dissertation has undergone originality checking, and it satisfies the recognized criteria for originality.



Date: 07/10/2025

ACKNOWLEDGEMENTS

I would like to begin by thanking my supervisor, Dr Richard Maliwatu, for his continuous support, advice, and helpful feedback during this research. His expertise and assistance were crucial in completing this project.

Secondly, I would like to convey my heartfelt gratitude to my family for their ongoing support and encouragement throughout this journey. In addition, I am deeply grateful to the Lord for giving me the strength, tenacity, and knowledge to complete this research effectively.

Lastly, I want to sincerely thank Sanlam Namibia and the Namibia University of Science and Technology for their continuous support during this endeavour. Your constant support has been crucial to my research efforts.

ABSTRACT

The prediction of lapse rate in the life insurance industry holds significant importance for insurers aiming to maintain profitability and retain a strong customer base. Accurate lapse prediction will enable insurers to deploy personalised retention efforts, effective risk management, and financial planning depending on the model's results. The study explored the use of Machine Learning (ML) techniques to predict lapse rates in life insurance. The dissertation's main contribution was to empirically compare and benchmark 4 machine learning classifier models (Logistic Regression, Gradient Boost, Random Forest, Support Vector Machine). The study utilised secondary data that was obtained from Sanlam Namibia's database, including policyholder demographics, policy features, premium amounts, payment frequency, lapse status, and other relevant information that could influence the lapse rate. Different types of machine learning algorithms were used on the acquired data to assess their performance in forecasting lapse rate. The study's findings showed that machine learning methods can accurately predict the lapse rates in life insurance. The findings reveal the ensemble model's high predictive capacity. Gradient Boosting is the best overall classifier, with a 94% and F1-Scor of 94% respectively. Furthermore, some key policy and clients' characteristics were found to have substantial predictive potential for the lapse rate. However, the limitations of the study must be considered. Finally, the study recommends using ensemble models rather than single model classifiers because they are more effective at predicting life insurance lapses.

Keywords: Support Vector Machine, Gradient Boost, Random Forest, Logistic regression, lapse, machine learning, life insurance, lapse risk.

Table of Contents

List of figures

List of tables

ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
(XAI)	Integration with Explainable
CART	Classification And Regression Tree
EDA	Exploratory Data Analysis
FN	False Negative
FP	False Positive
GB	Gradient Boost
ILO	International Labour Organisation
LR	Logistic Regression
ML	Machine Learning
PAC	Prediction Accuracy
RBF	Radial Basis Function
RF	Random Forest
SVM	Support Vector Machine
TP	True Positive
TF	True False

CHAPTER ONE
INTRODUCTION

The prediction of lapse rates in the life insurance sector holds significant importance for insurers aiming to maintain profitability and retain a strong customer base. Accurately forecasting lapse rates, which refer to the percentage of policyholders discontinuing their coverage prematurely, is essential for effective risk management and financial planning (Smith, 2023; Johnson et al., 2019).

In the last few years, the application of machine learning algorithms has shown promise in improving the accuracy of such predictions, offering potential advancements in the field (Brown & Green, 2020; Lee, 2021). Both insurers and policyholders depend on accurate insurance lapse forecasting. Ensuring financial stability for insurers, it acts as a proactive risk management method (Loisel et al., 2021). By providing individualised solutions, this strategy enables the retention of clients, thereby increasing client satisfaction (Hsieh & Chen, 2025). Accurate lapse estimates also make it easier to comply with regulations and preserve the insurer's standing in the market (Loisel et al., 2021). Due to accurate underwriting procedures, balanced risk portfolios and improved pricing strategies are produced (Johnson et al., 2019).

Insurance companies can stay competitive in a changing market by adjusting their product lineups and marketing tactics in response to lapse trends (Bauer et al., 2023). By refocusing efforts on retaining loyal consumers or obtaining new ones, resources can be allocated more effectively (Ferrario et al., 2023). The ability to adjust to market changes and consumer behaviour is facilitated by data-driven decision-making in lapse prediction (Frees et al., 2022). In the end, this strategy increases trust between insurers and policyholders and adds to long-term profitability (Hernandez & Ma, 2019).

This study investigated the usefulness of Machine Learning (ML) classification models in forecasting policy lapse risk in life insurance. The investigation trained and tested statistical models using policyholder data, including demographics and coverage details. The cross-validation and bootstrapping used to verify results are reliable. The study analysed the historical backdrop of policy lapse risk in the sector, as well as how policyholder factors like age and occupation affect lapse rate. This analysis was limited to a single dataset and did not

consider external factors that may impact policy lapse risk. However, the study's findings have potential benefits for life insurance companies.

1.1 Motivation of the study

The challenges posed by insurance lapses in the insurance sector are economically significant since a lapse has a detrimental influence on the financial well-being of the role players. This research can benefit industry participants such as financial intermediaries, insurance underwriters, and administrators who operate in dynamic, unpredictable, and complex environments. It is widely recognised that acquiring new clients is significantly more costly than retaining an old client base (Poufinas & Michaelide, 2018). The study provides valuable insights into the factors affecting lapse rates, thereby helping stakeholders to develop effective strategies for managing high lapse rates and related risks. Lowering lapse rates helps stakeholders retain customers and sustain their current client base, thus decreasing the expenses related to acquiring new clients to replace those lost to lapses.

1.2 Problem statement

An increase in lapse rates directly affects various aspects of insurance operations, including the company's financial stability, pricing strategies, and competitiveness. High lapse rates can lead to adverse consequences such as the need for premium increases, potential financial instability, damage to the company's reputation, and even insolvency (Huang & Sun, 2017; Turner & Jenkins, 2018). To address these challenges and implement personalised retention strategies, insurers must accurately predict lapse risks and understand cancellation behaviour. The use of accurate predictive models is crucial for decision-makers in the insurance sector to stay well-informed about future trends and make informed decisions. For Sanlam Namibia, this is important for it to achieve its goals and ensure profitability and customer retention. A better understanding and accurate prediction of policy lapse risks are essential for improving and effectively managing lapse risk, as it has a significant impact on the financial solvency of life insurance companies. This research employed ML techniques to predict the lapse rate using historical data.

1.3 Aim of the study

The study aimed to enhance the accuracy of predicting life insurance lapses through a comprehensive evaluation of machine learning classification models and feature selection techniques.

1.4 Research objectives and questions

1.4.1 Research objective

The main objective of this study was to assess the applicability and effectiveness of ML classification models in accurately predicting policy lapse rates using historical Sanlam data.

The specific sub-objectives of the research were as follows:

1. Evaluate the effectiveness of different commonly used ML classification models for forecasting life insurance lapses.
2. Identify and assess the most influential features in predicting life insurance lapses through in-depth data analysis.
3. Examine the impact of specific policyholder characteristics, such as age and occupation, on lapse likelihood.

1.4.2 Research questions

The main question guiding this research is: How can machine learning classification models and feature selection techniques predict life insurance lapses?

The specific sub-questions addressed in this research were as follows:

1. Among the four commonly used machine learning classification models, namely logistic regression, support vector machine, random forest and gradient boost, which model demonstrates the highest effectiveness in predicting life insurance rates?
2. Which key features in the dataset significantly influence the prediction of life insurance lapses?
3. How do policyholder characteristics such as age and occupation affect the likelihood of a life insurance policy lapse?

1.5 Significance of study

The study's findings can assist Sanlam Namibia in making effective retention decisions based on the models' results, hence reducing the risks associated with client loss. It can assist

Sanlam in efficiently planning finances, reducing prediction uncertainty, and providing early warnings of cancellations. Clients can benefit from lower premium rates and other benefits associated with retention initiatives if model forecasts are right. The findings also add to the existing academic literature.

1.6 Study scope and delineation

The study only focused on Sanlam Namibia; other insurance companies in Namibia were not part of it. The study focused on developing and executing ML algorithms to predict life insurance policy lapse rates within the specific context of the Sanlam Namibia environment. The research obtained relevant data such as policy specifics, payment histories, and demographic data from Sanlam Namibia's databases. After processing, the data were used to create machine learning models. A predictive model was constructed by the study using a variety of methods, and it was then trained and validated using various data subsets. To evaluate the model's performance, evaluation measures were applied. The objective was to provide Sanlam Namibia with informed recommendations on potential lapse mitigation measures by identifying and analysing the key factors influencing lapse rates. The study emphasised ethical considerations, admitted any shortcomings, and recommended directions for future research.

1.7 Limitations of the study

The study focused on developing and executing machine learning algorithms to predict life insurance policy lapse rates within the specific context of the Sanlam Namibia environment. The population was limited to policyholders within Sanlam’s policyholders. This study did not include policyholders from other insurance companies in Namibia. Instead, it relied on relevant secondary data, including policy details, payment history, and demographic information extracted from Sanlam Namibia’s database.

This study focused on predicting lapse rates in the life insurance industry by employing machine learning algorithms as a predictive tool. The study was conducted within the context of Sanlam Namibia, a prominent insurance company, aiming to enhance its understanding and forecasting of lapse rates.

Chapter one outlines the significance of accurate lapse rate predictions and the use of ML in this context. The problem statement, questions, and objectives of the research are introduced, and the limitations of the study are outlined.

Chapter two provides an in-depth review of the concepts of lapse rates in life insurance, the factors influencing them, and the broader landscape of machine learning applications in the insurance sector. Additionally, it surveys previous studies and case studies relevant to lapse rate prediction.

Chapter three outlines the steps taken to gather and preprocess data from Sanlam Namibia. It details the selection of variables, feature engineering techniques, and the choice of machine learning algorithms for predictive modelling. Various evaluation metrics are described for assessing model performance. Explanatory variables present an overview of numerous insights and assumptions about lapse rates that have been evaluated in the literature. The chapter presents a table listing all variables that have demonstrated or are expected to experience a significant relationship with lapse rate.

Chapter four presents descriptive statistics and the outcomes of model training and validation. These results are interpreted, thereby providing valuable insights into the effectiveness of the chosen machine learning algorithms in predicting lapse rates.

Chapter five presents a comparison of the findings with existing literature and offers implications for Sanlam Namibia and the broader insurance industry. Finally, the study concludes by summarising key findings, acknowledging any limitations, and suggesting directions for future research in this area.

1.8 Chapter summary

The chapter underscored the importance of accurately predicting life insurance policy lapses for Sanlam Namibia and the broader industry. High lapse rates negatively impact profitability, solvency, and acquisition costs, while precise predictions aid in reserving retention strategies and regulatory compliance. The chapter highlighted recent literature, thereby suggesting that machine learning (ML) classifiers such as logistic regression, support vector machines, random forests, and gradient boosting can outperform traditional actuarial methods in predicting lapses. The study aimed to evaluate the effectiveness of these ML models using Sanlam's historical data, guided by three sub-objectives: comparing ML models, identifying key predictive features, and assessing the influence of policyholder attributes.

Research questions were formulated to address these objectives, emphasising the practical significance for Sanlam and its customers. The scope was limited to Sanlam Namibia's portfolio, acknowledging dataset constraints and excluding external macroeconomic factors. Key limitations and ethical considerations were outlined to set realistic expectations for the findings' generalisability. The chapter also previewed the structure of the dissertation, including the literature review, methodology, results, discussion, and conclusion chapters. With the research problem, objectives, and questions clearly defined, the groundwork is laid for a thorough literature review in Chapter Two.

CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction

The insurance sector plays a crucial role in the financial services industry, as it impacts economic growth, promotes efficient resource allocation, reduces transaction costs, creates liquidity, simplifies economies of scale in investment, and spreads financial losses. Insurance is a contract between a policyholder and an insurer, in which the policyholder transfers the risk of financial loss to the insurer, who undertakes the associated uncertainty (Haiss & Sümegi, 2020).

The prediction of lapse rates in life insurance is a critical concern for insurance companies, as it directly impacts their financial stability, pricing strategies, and overall competitiveness. The increasing use of ML algorithms offers a promising approach to enhancing the accuracy of these predictions. This literature review explores various machine learning techniques, such as support vector machine, logistic regression, gradient boosting, and Random Forest, used in predicting lapse rates in life insurance, specifically focusing on a case study of Sanlam Namibia, as well as the difficulties of implementing machine learning in the life insurance sector.

The life insurance sector is essential to risk management, financial security for people and families, and financial planning. To remain profitable and retain a strong customer base, insurers must accurately predict lapse rates, which refer to the percentage of policyholders who discontinue their coverage early. In the insurance industry, applying machine learning algorithms has recently shown promise in terms of improving predictive accuracy. This literature review primarily focuses on lapse risk in life insurance, commonly referred to as churn, and how machine learning can be used to forecast lapse rate in life insurance.

In conducting this review, a comprehensive search was performed using academic articles, abstracts, and resources from Google Scholar and the Namibia University of Science and Technology (NUST) library. The main keywords guiding this review include Random Forest, Logistic Regression, Lapse, Machine Learning, Linear Model-Logistic Regression-Binomial Classification, Support Vector Machine-Kernel Function, and gradient boosting-Boosting Ensemble. These keywords were chosen to ensure a broad and detailed examination of the relevant methodologies and their applications in lapse rate prediction.

This chapter also lays the groundwork for the research given in the remaining sections of the thesis, including the models examined and the difficulties in applying machine learning to the life insurance sector. This literature review aims to explore the existing body of research concerning the utilisation of machine learning techniques for predicting lapse rates in life insurance.

2.2 Sanlam Namibia - Insurance company

Sanlam Namibia is a significant financial services provider. It is an important player in Namibia's life insurance business, providing a variety of policies designed to protect consumers' financial well-being. As part of the larger Sanlam Group, the Namibian branch has substantial issues in controlling and forecasting the lapse rate of life insurance contracts, which can have a considerable impact on profitability and client retention tactics. Under these circumstances, using machine learning algorithms to anticipate lapse rates helps to uncover the potential for improving Sanlam Namibia's ability to identify high-risk policies and implement target interventions. Sanlam can better understand the causes of policy lapses and optimise its risk management processes by using predictive analytics, resulting in more sustainable business operations. The study sought to investigate the application of machine learning approaches in predicting lapse rates, using Sanlam Namibia as a case study, to improve the accuracy of risk assessment and strengthen customer retention efforts.

2.3 Life insurance

A policy agreement between an insurer and a policyholder that offers protection for individuals, groups, and pension plans is known as life insurance. A life insurance contract is a formal arrangement that promises to make payments if one or more people pass away. The insurer, the insured (whose lifetime determines the payment of benefits), the policyholder (who begins the contract and pays the premium), and the beneficiary (who receives the benefits) are the parties involved in the contract (Anifalaje, 2021). Typically, a life insurance policy's policyholder pays premiums during their lifetime in exchange for the insurer's promise to pay a certain amount to designated beneficiaries after the policyholder passes away (Fontinelle, 2021).

2.4 Life insurance products

Life insurance differs from other types of insurance by offering long-term benefits. These products may offer benefits such as survival, death, or both. Common life insurance products include pure endowment, life annuities, term, entire, and endowment insurance. A life insurance contract has monetary components such as premiums, benefits, and expenses (Anifalaje, 2021). Term life insurance, often known as temporary life insurance, provides coverage for a specific number of years. This type of insurance provides a payment to specified beneficiaries equal to the policy's face amount if the policyholder dies within its term. If the policyholder does not die during the policy period, no payments will be made.

The policyholder is obligated to pay the insurance company regularly for the duration of the policy or until death (Mariyam, 2025). Insurance companies have attempted to increase the efficiency of product sales over time by using underwriting processes (Towiwat & Swierczek, 2024).

2.5 Insurance lapse and lapse rate

A lapse in the insurance sector is the termination of policy coverage because of unpaid premiums (Baoilan et al., 2025). Every time a premium is late, a policy does not always automatically lapse. Before the payment expires, the policyholder has a grace period to make it. Should a claim be submitted within the grace period, the insurer is responsible for paying the customer the benefit. To remain profitable and retain a strong customer base, insurers must accurately predict lapse rates, which refer to the percentage of policyholders who discontinue their coverage early (Baoilan et al., 2025).

In a broader sense, lapse covers partial and early terminations of contracts, as well as changes to the frequency and size of premium payments. Due to the severity of this risk, lapse hereafter refers to the complete and early termination of the contract initiated by the policyholder. Given that the premium calculations are based on the contract continuing (and paying premium) until the end of the specified term, it has a significant impact on an insurer's asset liability management. If the contract is terminated early, then high acquisition costs may remain unpaid. An insurance firm faces a serious risk because of higher-than-expected lapse rates.

The possibility of a lapse is believed to account for up to 50% of a contract's fair value and is one of the largest components of regulatory capital (Biagini et al., 2021). Therefore,

accurately predicting lapse rates is significant. The likelihood of a lapse can impact contract pricing, the insurer's required liquidity, and the regulatory capital that must be maintained (Biagini et al., 2021).

Policyholders have the right to cancel a contract, a process known as a lapse. A key issue with early lapses is that insufficient premium payments may have been made to cover the costs associated with the insurance (Biagini et al., 2021). Losses from expired contracts are incorporated into the calculation of new insurance prices to mitigate the negative effects of lapses. Consequently, future rates will be adjusted to reflect the risk, which will be shared by both current and future policyholders. According to Biagini et al. (2021), the option to lapse may account for up to 50% of the contract's fair value in certain cases.

Policy-specific factors like entrance age, contract duration, gender, or sum insured can be used to analyse a particular lapse behaviour in a certain contract as well as an individual. However, since policyholder-specific data is frequently treated confidentially, there is a limited number of studies that utilise such data. Eling and Kochanski (2022) provide a comprehensive summary of the research on life insurance lapse, identifying 12 empirical publications and 44 works on theoretical lapse models. Among the 12 empirical studies, seven focused on macroeconomic variables, while four examined product and policyholder characteristics.

2.6 Importance of predicting insurance lapse

Policy lapses introduce uncertainty and risk for insurers. This risk, known as lapse risk, stems from the unpredictability and potential magnitude of policy terminations. Regulatory bodies mandate insurers to maintain specified buffers of capital, termed regulatory capital, to guarantee their ongoing operation and stability (Deng, 2022).

Both insurers and policyholders depend on accurate insurance lapse forecasting. Ensuring financial stability for insurers acts as a proactive risk management method (Braun et al., 2019). By providing individualised solutions, this strategy enables the retention of clients, thereby increasing client satisfaction (Wade et al., 2016). Accurate lapse estimates also make it easier to comply with regulations and preserve the insurer's standing in the market (Eling & Kochanski, 2022). Owing to accurate underwriting procedures, balanced risk portfolios and improved pricing strategies are produced (Deng, 2022).

Insurance companies can stay competitive in a changing market by adjusting their product lineups and marketing tactics in response to lapse trends (West et al., 2020). By refocusing efforts on retaining loyal consumers or obtaining new ones, resources can be allocated more effectively (Wuthrich & Buser, 2019). The ability to adjust to market changes and consumer behaviour is facilitated by data-driven decision-making in lapse prediction (Bruun et al., 2024). In the end, this strategy increases trust between insurers and policyholders and adds to long-term profitability (Eling & Kochanski, 2022).

2.7 An analysis of the factors influencing life insurance

This section explores the different factors that have an influence on the pricing and underwriting of life insurance products. Insurance employs explanatory factors such as age, gender, occupation, level of education, product type, payment method, income level, and premium amount to assess the risk of insuring an individual and to reveal the likelihood of their demise. The impact that these variables may have on the price of life insurance plans and the processes used to calculate premiums is, therefore, examined. Understanding these explanatory factors is crucial for both individuals purchasing a policy and insurance companies' pricing and underwriting them.

2.7.1 Demographics

Age, gender, education level, occupation, physical health, etc, are demographic variables that affect the risk profile of potential life insurance and pension policyholders. Based on these qualities, insurance companies set risk-reflective premiums using various premium tables (Braun et al., 2019).

Using risk-rating parameters, life insurers group policyholders into risk categories, a process known as risk pooling (Masiuk, 2025). To avoid exorbitant costs, the rating variables must be impartial, verifiable, and affordable to calculate premiums (Masiuk, 2025). Both individuals buying a policy and insurance companies pricing and underwriting them need to be aware of these explanatory elements.

Risk-based pricing and risk sharing are the two guiding ideas for the supply of private insurance based on demographics (Braun et al., 2019). As a result of higher underwriting expenses, the risk-based pricing technique uses medical and specialised exams and questions to calculate the cost for each policyholder. Risk categorisation can help to resolve this problem (Xin & Huang, 2024). The second premise, known as risk sharing, entails paying

premiums to share risk among people participating in risk pools, thereby enabling insurers to diversify their risk (Tarr et al., 2023; Xin & Huang, 2024).

2.7.2 Age

Age is often considered a key risk factor in life insurance and pension pricing, as it is a quantitative variable that is simple to assess and inexpensive to acquire (Lim et al., 2023). Younger people are thought to have fewer health conditions that could cause mortality because life expectancy declines with age. As a result, insurance companies charge older applicants for insurance policies greater premiums and may not give insurance policies past a certain age (Xin & Huang, 2024).

2.7.3 Gender

In the pricing structure, gender is also employed as a rating element, mostly for pension products, and combined with more thorough medical underwriting (Fong, 2015). As a result of research demonstrating that men have a higher mortality rate than women, insurers utilise this data to grade insurance policies that cover risks that differ for men and women (Xin & Huang, 2024). However, it is crucial to remember that the European Court of Justice issued an EU Gender Directive in 2004 that called for equal treatment of men and women in the supply and access to goods and services (Böök et al., 2025).

2.7.4 Socio-economic factors

Socio-economic factors, including income level, education level, culture, health, way of life, and occupational risks, have an impact on insurance costs. Insurance companies provide new solutions to address these changing risks (Braun et al., 2019). For instance, accidents are a common risk in many occupations. Based on estimates from the International Labour Organization (ILO), accidents account for 30% of all work-related deaths worldwide, with diseases accounting for the remaining 70% (Pega et al., 2021). Therefore, even if medical and other criteria are positive, people in high-risk occupations may still have to pay a higher premium (Xin & Huang, 2024). Overall, identifying the risk profile of policyholders and establishing adequate risk-reflective rates depend greatly on demographic considerations.

2.8 Machine learning for lapse prediction

In recent years, there has been an increased interest in investigating the use of machine learning (ML) approaches for lapse prediction in life insurance. ML algorithms provide various advantages over traditional statistical methods, including the capacity to handle complex

nonlinear relationships and the inclusion of several data formats, and potentially enhance generalisability to previously unexplored data (Kim & Kim, 2020).

Several studies have proved the efficacy of machine learning for lapse prediction. Kim and Kim (2020) found that Gradient Boosting models outperformed logistic regression in predicting lapse behaviour. Their findings demonstrated the capacity of ML models to capture complicated interactions between policyholder characteristics, product features, and previous payment history. Random Forests were proven to effectively identify high-risk policyholders based on demographic, behavioural, and economic characteristics (Rana & Gupta, 2020). Regression models are commonly used to forecast lapse rates based on policy and customer data, including age, gender, insurance type, and premium payment history (Andaria et al., 2025; Ferrario et al., 2023).

These studies highlight the potential of ML for improving lapse prediction accuracy compared to traditional methods. However, some gaps remain in the current literature, which necessitated further exploration of advanced ML techniques in the present study.

2.9 Machine learning approaches

More uses of artificial intelligence (AI) and machine learning (ML) are likely to become commonplace. According to the theory behind machine learning (ML) (Yang, 2023), robots should pick up new skills and get better with practice. According to Burri Rama Devi (2019), learning is generally divided into three types: supervised, unsupervised, and reinforcement learning. An ML process, called supervised learning, employs a training dataset to instruct the mapping from input features to target labels (labelled answers) by studying a variety of input-output examples (Burri Rama Devi, 2019). In unsupervised learning, machine learning models analyse input data to uncover patterns without the guidance of labelled targets (Burri Rama Devi, 2019). The most popular machine learning (ML) technique in the insurance sector is supervised learning (Burri Rama Devi, 2019).

Supervised learning is often divided into two types: classification and regression (Burri Rama Devi, 2019). The response variable, whether numeric or categorical, distinguishes between the two. Classification predicts categorical responses based on input-output examples, whereas regression predicts continuous numeric responses (Ayodele, 2021). In supervised learning, unbalanced data can occur when one class has fewer occurrences than the others.

Researchers have focused on this scenario because of its prevalence in real-world applications (Omar, 2022). The present study aimed to optimise binary classification issues.

Unsupervised algorithms, unlike supervised learning, analyse input data without the use of predefined labels (Burri Rama Devi, 2019). Without supervision, it seeks out patterns and structure in the data on its own. Unsupervised learning aims to find regularities, irregularities, links, similarities, and linkages in input data by examining unlabelled data (Burri Rama Devi, 2019). The algorithm investigates the overall correlation between data clusters (Burri Rama Devi, 2019; Ayodele, 2021). The most well-known challenges in unsupervised learning are clustering and association rules (Ayodele, 2021).

There are further related difficulties, such as outlier and anomaly detection. This technique identifies data instances that deviate significantly from predicted behaviour and patterns, attempting to uncover exceptional occurrences that stand out from the norm (Omar, 2022).

The insurance business has benefited considerably from big data technologies, which enable more efficient data collecting, processing, analysis, and management (Blier-Wong & Marceau, 2021). ML is a fast-developing topic with enormous potential for further application in actuarial science (Blier-Wong & Marceau, 2021; Burri Rama Devi, 2019; Maier et al., 2019). The present study employed supervised learning as a machine learning (ML) technique.

2.10 Machine Learning Models (MLM)

There have been extensive studies on predicting lapse rates in life insurance (Shamsuddin et al., 2022). The current research used machine learning to analyse policyholder data for patterns and trends that may indicate policy lapses.

Regression models are commonly used to forecast lapse rates based on customer and policy variables, including age, gender, insurance type, and premium payment history (Andaria et al., 2025). Previous researchers used decision trees to categorise policyholders as "likely to lapse" or "not likely to lapse" (Malali, 2023; Wang et al., 2019). Insurance professionals typically use Generalized Linear Models (GLMs), namely Logistic Regression (Xong & Kang, 2019; Loisel et al., 2021; Maynard et al., 2019; Wang, 2021a).

Other researchers suggest that ensemble models, which incorporate predictions from many models, can enhance prediction accuracy (Grize et al., 2020; Loisel et al., 2021). Boosting and bagging strategies can be applied to train multiple models on various data subsets and

combine their predictions (Diana et al., 2019; Grize et al., 2020; Andaria et al., 2025; Loisel et al., 2021).

Some studies have applied advanced ML methods such as support vector machines to estimate lapse rates in life insurance (Groll et al., 2022; Xong & Kang, 2019). While promising, these strategies may require additional data and computing resources to apply (Bolancé et al., 2016; Diana et al., 2019; Grize et al., 2020; Maynard et al., 2019).

Numerous studies have employed various algorithmic techniques, including Regression and Classification trees (Grize et al., 2020; Groll et al., 2022; Loisel et al., 2021), Elastic Net regularisation (Groll et al., 2022; Loisel et al., 2021), Naive Bayes (Scriney et al., 2020), and Random Forests (Bärtl & Krummaker, 2020; Diana et al., 2019; Groll et al., 2022).

Several academic studies have assessed the performance of various models, including Bärtl and Krummaker (2020), Diana et al. (2019), Grize et al. (2020), Groll et al. (2022), Maynard et al. (2019), Scriney et al. (2020), Yung Wang (2021b), Wuthrich and Buser (2019), Xong and Kang (2019), and Groll et al. (2022). These have compared logistic regression to CART, neural networks, Elastic-net, extreme gradient boosting, Support Vector Machine, and random forest models for predicting churn lapses.

2.10.1 Logistic Regression (LR)

Logistic Regression is a modelling approach that utilises a link function to forecast the probability of a binary outcome using one or more independent variables (Kgare, 2019). This implies two possible results, such as 0 or 1; in the current study, these outcomes are lapse or non-lapse (stay active) (Kgare, 2019). When working with a binary dependent variable (0 or 1), logistic regression becomes a crucial analytical method. In logistic regression, independent variables (x) are linearly combined using weights or coefficient values to predict the dependent variable (y). According to Kgare (2019), the logistic regression equation is expressed as shown in Equation 2.1

$$\text{logit}(y) = \ln\left(\frac{P}{1-P}\right) = \beta_0+\beta_1\chi_1 +.....+\beta_k\chi_k \,, \qquad (2.1)$$

Where,

- y, represents the dichotomous outcome.
- χ_1, … …, χ_k, are the predictor variables such as education, premium rate,
- entry age, occupation, gender, and product type.

- β_0, β_1, … …, β_k, are the regression coefficients for the model, β_0 is the intercept.
- p is the proportion of the data with lapse outcome, $1 - p$ is the probability of no lapse (active).

2.10.2 Support Vector Machines (SVM)

Boser et al. (1992) introduced the support vector machine (SVM) at the fifth annual workshop of the Association for Computing Machinery. SVM research has grown in popularity within the field of machine learning. It is well-known for its high data classification abilities (Ndungi, 2023).

SVMs address both linear and nonlinear problems using kernel functions that transform data into separable forms (Ndungi, 2023). Common kernels include linear, RBF, polynomial, and sigmoid. SVMs require careful parameter tuning, including the regularisation parameter (C) and kernel-specific parameters like gamma, with grid search being a common method for optimisation abilities (Statinfer, 2019).

SVMs offer several advantages, such as the ability to work in high-dimensional spaces, low risk of overfitting, effectiveness when class separation is clear, and the capacity to handle complex problems with suitable kernel functions. However, they also have disadvantages, including the challenge of selecting an appropriate kernel function and the tendency to underperform when the number of attributes exceeds the number of data samples (Statinfer, 2019).

2.10.3 Random Forest

It is an ensemble learning model introduced by Breiman (2001). It combines multiple decision trees (DTs) using bootstrap sampling, where each tree is built by randomly selecting features and training on different data subsets. The trees vote on the most popular class, and the overall accuracy improves through the aggregation of individual trees' predictions (Groll et al., 2022).

Advantages of RF include its ability to handle large, high-dimensional datasets, robustness to overfitting, insensitivity to outliers, and capacity to work with both classification and regression problems (Groll et al., 2022). RF does not require feature scaling, can manage missing data, and is easy to set up. However, it can be slow to run and may show bias towards categorical values (Groll et al., 2022).

2.10.4 Gradient Boosting

It is a classification model that combines decision trees (DTs) sequentially, with each tree correcting the errors of the previous one, unlike Random Forest (RF), where trees are built independently (Bentéjac et al., 2019). This process reduces the ensemble's overall loss function through gradient descent. However, GB's iterative nature requires careful tuning of regularisation parameters to prevent overfitting (Bentéjac et al., 2019).

The methods frequently outperform traditional machine learning models. Moreover, there is no need to normalise or standardise features. However, these methods can be time-consuming to execute, are enhanced sequentially rather than concurrently, and are vulnerable to outliers (Bentéjac et al., 2019).

2.11 Machine learning challenges

This section explores common machine learning issues and the approaches used to address them. Specifically, it examines overfitting, class imbalance, and missing data. Methods for handling missing data in machine learning models are discussed, along with the limitations of text analysis and strategies for its effective application within machine learning contexts.

2.11.1 Overfitting

Overfitting refers to a model's ability to perform a certain task and generate accurate predictions on unknown data (Guido, 2016). A model's ability to learn and perform well on training data can have an impact on its ability to predict fresh data (Guido, 2016). This is observed when a model excels on training data but underperforms on testing data. Inconsistencies in the dataset, limited training sets, and classifier complexity can induce overfitting, while underfitting happens when the model is too simple and has low accuracy (Journal of Physics & Ying, 2019; Low et al., 2025). Cross-validation can help prevent overfitting (Low et al., 2025). This method generates multiple smaller subsets from the training set and leverages them to improve the model (Low et al., 2025).

2.11.2 Class imbalance

Unbalanced data is a common challenge in machine learning classification problems, where observations are not evenly distributed among classes (Arnold et al., 2025). Imbalance can negatively affect the performance of machine learning algorithms, which need a balanced representation of classes to produce correct predictions. Real-world datasets frequently have

imbalanced class distributions, with one or more classes being under- or over-represented (Tsai et al., 2022).

Imbalanced data refers to a majority class with more samples than the minority class (Tsai et al., 2022). This can result in a biased prediction model that favours the majority class over the smaller class. Certain strategies have existed to address these matters, including using data-sampling methods to alter the training data (Kim & Hwang, 2022). These approaches can restore balance by randomly under-sampling the majority or over-sampling the minority (Kim & Hwang, 2022; Tsai et al., 2022).

SMOTE (Synthetic Minority Over-sampling Technique) is a prominent approach for over-sampling (Xie et al., 2019). SMOTE creates synthetic minority samples by interpolating existing samples (Xie et al., 2019). Shumaly et al. (2020) discovered that under-sampling outperformed random sampling for predicting churn on unbalanced datasets.

2.11.3 Missing data

Missing data poses a challenge for machine learning, potentially reducing model accuracy (Roy, 2019). It can be addressed through deletion or imputation. Deletion methods, such as listwise deletion, are suitable for data missing completely at random (MCAR) in large datasets but risk losing valuable information in smaller datasets (Norazian, 2021; Roy, 2019). MCAR data can be imputed using mean or mode values (Gelman, 2010). For data missing at random (MAR), where missingness relates to observed variables, multiple imputation is effective, thereby generating and combining multiple predictions (Norazian, 2021). For missing not at random (MNAR) data, where missingness depends on unobserved values, specialised handling is required (Norazian, 2021).

2.12 Chapter summary

Numerous studies (including those Bärtl & Krummaker, 2020; Dana et al., 2019; Grize et al., 2020; Loisel et al., 2021; Maynard et al., 2019; Scriney et al., 2020; Wang et al., 2019; Wuthrich & Buser, 2019; Xong & Kang, 2019) have compared the performance of various models for lapse prediction in insurance. For instance, Groll et al. (2022) evaluated logistic regression, Classification and Regression Tree (CART), neural networks, Elastic-net, extreme gradient boosting, Support Vector Machine, and random forest models for churn management-related lapse prediction.

Accurate lapse rate prediction is critical in insurance operations, as high lapse rates can lead to increased premiums, financial instability, reputational harm, and even insolvency (Johnson et al., 2019). To mitigate these risks, insurers are adopting machine learning techniques to develop personalised retention strategies (Johnson et al., 2019). Machine learning has demonstrated its effectiveness across diverse fields, making it a promising tool for predicting lapse rates (Rana & Gupta, 2020). These models utilise historical data, such as demographics and purchasing behaviours, to enhance predictive accuracy Li & Chen, 2021; Wang et al., 2019). Techniques like decision trees, random forests, and neural networks have shown significant applicability in this domain (Hernandez & Ma, 2019; Kim & Kim, 2020).

Despite their promise, implementing machine learning in life insurance presents challenges, including issues with data quality, model interpretability, and the need for continuous updates to align with changing market conditions (Peddamukkula, 2024). A study by Shamsuddin et al. (2022) found a growing trend in research on life insurance lapse risk, with 84.4% of 178 analysed papers discussing this topic. Popular approaches include machine learning, pricing strategy, agent behaviour analysis, stochastic interest rate modelling, and logistic regression.

Research has addressed questions such as the determinants of lapse rates and the potential of machine learning classification models for prediction. However, gaps remain, particularly in the development of advanced predictive models. Despite these advances, gaps remain in interpretability, feature engineering, and integration with explainable AI (XAI) techniques (Peddamukkula, 2024). Addressing these gaps can enhance model transparency, compliance, and actionable insights, paving the way for more effective and responsible applications of ML in the insurance industry. Future studies could focus on integrating cutting-edge machine learning algorithms to enhance accuracy and adaptability. Current research remains in its early stages, necessitating further work to improve model performance and understanding of the factors influencing lapse rates. This research aimed to advance lapse rate prediction through sophisticated machine learning algorithms using a case study on Sanlam Namibia. This study sought to benefit the company while contributing to industry-wide improvements in managing lapse rates, thereby offering a more precise and timely approach to addressing this critical issue.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Introduction

This chapter outlines the approach and techniques used to predict life insurance policy lapse rates at Sanlam Namibia. The study focused on forecasting whether a policyholder's insurance policy will lapse or remain active, which is vital for managing risks and optimising profitability. The section outlines the research design, data collection, pre-processing, feature selection, model development, and evaluation methods used in the study.

The methodology for addressing this problem involved several key steps. A comprehensive literature review (Chapter Two) was conducted to identify important explanatory variables and challenges associated with lapse risk. Following the literature review, a dataset comprising policyholder information, including demographics, financial details, and policy variables, was collected and pre-processed. This dataset was then utilised to train and evaluate machine learning algorithms for predicting lapse risk.

In essence, this chapter provides a detailed overview of the methodology used to address the binary classification problem of lapse risk in the life insurance industry. It covers identifying relevant explanatory variables, preprocessing large datasets, evaluating and optimising machine learning models, and interpreting results.

3.2 Research paradigm

This study adopted a positivist research paradigm, which is rooted in the belief that reality is objective and can be measured through empirical observation and statistical analysis. The positivist paradigm is particularly suitable for studies that seek to explain phenomena using quantifiable data and objective measurements (Dudovskiy, 2021). Given that the goal of this research was to predict lapse rates in life insurance policies using machine learning algorithms based on historical policyholder data, the positivist approach was appropriate. It allowed for a structured and systematic investigation of the relationships between input variables and policy lapse behaviour (Dudovskiy, 2021).

3.3 Research approach

A quantitative research approach was employed to investigate the problem. Quantitative methods are typically used when the aim is to measure the extent or frequency of certain phenomena and test hypotheses using statistical tools (Dudovskiy, 2021). In this study, historical data from Sanlam Namibia were analysed using machine learning algorithms to develop predictive models. The quantitative approach facilitated the use of mathematical computations, accuracy assessments, and performance comparisons across various models, allowing for an evidence-based determination of the best-performing algorithm for predicting policy lapses.

3.4 Research design

The study utilised an experimental and predictive modelling design. This design involved the collection and transformation of historical data, the application of various machine learning techniques, and the systematic evaluation of the models. Predictive modelling design is appropriate for studies that aim to forecast future events based on historical patterns (Oscar, 2021). In this case, the model was trained on labelled data to identify patterns associated with policy lapses. The design allowed for rigorous testing and comparison of machine learning algorithms such as logistic regression, support vector machine, random forest, and gradient boosting, using predefined performance metrics.

3.5 Data collection

Data were collected from the internal database of Sanlam Namibia, a leading life insurance provider. The dataset consisted of records from January 2022 to August 2023 and included information on 16,975 unique policyholders classified as major members. Variables captured in the dataset included policyholder demographics (such as age and gender), policy details (such as policy start and end dates), premium amounts, payment history, lapse status, and other relevant features. The data were obtained with the necessary permission and in adherence to ethical standards, thereby ensuring the use of reliable and authentic sources. The collected dataset served as the foundation for training, validating, and testing the predictive models.

3.6 Data preparation

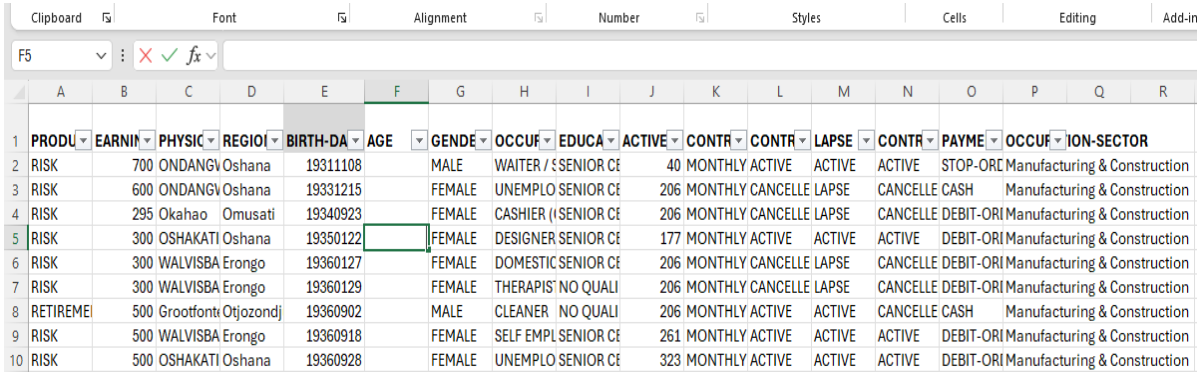
Data preparation is an important procedure in ML that converts raw data into usable input. The structure and scale of input data greatly impact the effectiveness of machine learning models and approaches. Inconsistencies in real-world datasets can negatively impact model performance. Pre-processing improves the quality and usability of data for machine learning analysis (García et al., 2015). The researcher used descriptive statistics to better comprehend the data structure. To prepare for modelling, a data audit was conducted to remove inconsistencies such as outliers, mistakes, erroneous date formats, and duplication.

3.6.1 Data source

Historical lapse data from the Sanlam Namibia database were used as model inputs. The dataset comprised policyholder demographics, policy details, premium amounts, payment history, lapse status, and other variables potentially influencing lapse rates. Initially, the dataset included 10 categorical and 3 numerical variables, representing a total of 16,975 unique policyholders who were classified as major members. To enhance the predictive power and resilience of the models, additional feature engineering was performed. This involved creating new variables and transforming existing ones. For instance, region and occupation sector variables were derived from town names and job titles, respectively. The dataset covered policy records from January 2022 to August 2023.

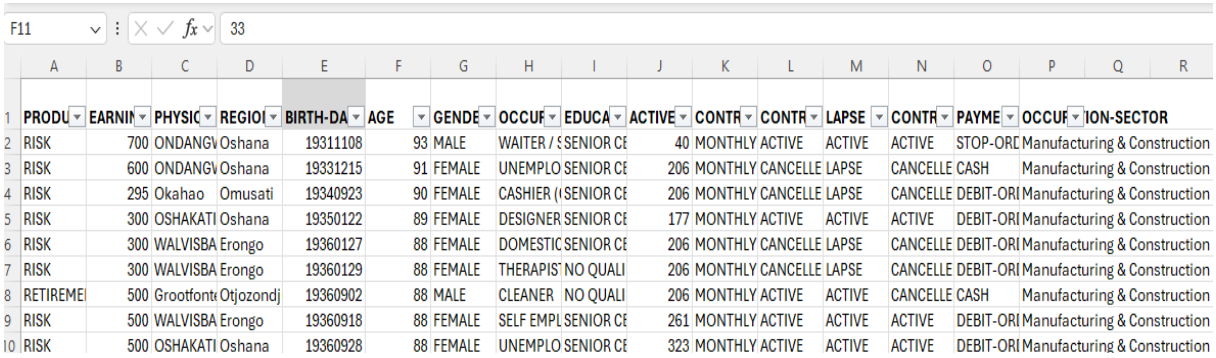
3.6.2 Imputation of missing data

In the dataset, there were missing values for key attributes such as age. To address this issue, the researcher employed an imputation strategy using the available information. Specifically, for instances where the age attribute was missing, the date of birth (DOB) variable was utilised to calculate the age as indicated in Figures 2 and 3. This was done by subtracting the year of birth from the current year or the date of data collection to derive an accurate estimate of age. This approach allowed for the effective completion of missing data fields, thereby enhancing the dataset's completeness and reliability. By leveraging the contextual clues provided within the existing data, the researcher was able to minimise the impact of missing values, ensuring a more robust data analysis and valid research outcomes.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	PRODU	EARNI	PHYSIC	REGIO	BIRTH-DA	AGE	GENDE	OCCUF	EDUCA	ACTIVE	CONTR	CONTR	LAPSE	CONTR	PAYME	OCCUF	ION-SECTOR	
2	RISK	700	ONDANGV	Oshana	19311108		MALE	WAITER / S	SENIOR CE	40	MONTHLY	ACTIVE	ACTIVE	ACTIVE	STOP-ORI	Manufacturing & Construction		
3	RISK	600	ONDANGV	Oshana	19331215		FEMALE	UNEMPLO	SENIOR CE	206	MONTHLY	CANCELLE	LAPSE	CANCELLE	CASH	Manufacturing & Construction		
4	RISK	295	Okahao	Omusati	19340923		FEMALE	CASHIER (I	SENIOR CE	206	MONTHLY	CANCELLE	LAPSE	CANCELLE	DEBIT-ORI	Manufacturing & Construction		
5	RISK	300	OSHAKATI	Oshana	19350122		FEMALE	DESIGNER	SENIOR CE	177	MONTHLY	ACTIVE	ACTIVE	ACTIVE	DEBIT-ORI	Manufacturing & Construction		
6	RISK	300	WALVISBA	Erongo	19360127		FEMALE	DOMESTIC	SENIOR CE	206	MONTHLY	CANCELLE	LAPSE	CANCELLE	DEBIT-ORI	Manufacturing & Construction		
7	RISK	300	WALVISBA	Erongo	19360129		FEMALE	THERAPIS	NO QUALI	206	MONTHLY	CANCELLE	LAPSE	CANCELLE	DEBIT-ORI	Manufacturing & Construction		
8	RETIREME	500	Grootfont	Otjozondj	19360902		MALE	CLEANER	NO QUALI	206	MONTHLY	ACTIVE	ACTIVE	CANCELLE	CASH	Manufacturing & Construction		
9	RISK	500	WALVISBA	Erongo	19360918		FEMALE	SELF EMPL	SENIOR CE	261	MONTHLY	ACTIVE	ACTIVE	ACTIVE	DEBIT-ORI	Manufacturing & Construction		
10	RISK	500	OSHAKATI	Oshana	19360928		FEMALE	UNEMPLO	SENIOR CE	323	MONTHLY	ACTIVE	ACTIVE	ACTIVE	DEBIT-ORI	Manufacturing & Construction		

Figure 1: Dataset with missing values before imputation



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	PRODU	EARNI	PHYSIC	REGIO	BIRTH-DA	AGE	GENDE	OCCUF	EDUCA	ACTIVE	CONTR	CONTR	LAPSE	CONTR	PAYME	OCCUF	ION-SECTOR	
2	RISK	700	ONDANGV	Oshana	19311108	93	MALE	WAITER / S	SENIOR CE	40	MONTHLY	ACTIVE	ACTIVE	ACTIVE	STOP-ORI	Manufacturing & Construction		
3	RISK	600	ONDANGV	Oshana	19331215	91	FEMALE	UNEMPLO	SENIOR CE	206	MONTHLY	CANCELLE	LAPSE	CANCELLE	CASH	Manufacturing & Construction		
4	RISK	295	Okahao	Omusati	19340923	90	FEMALE	CASHIER (I	SENIOR CE	206	MONTHLY	CANCELLE	LAPSE	CANCELLE	DEBIT-ORI	Manufacturing & Construction		
5	RISK	300	OSHAKATI	Oshana	19350122	89	FEMALE	DESIGNER	SENIOR CE	177	MONTHLY	ACTIVE	ACTIVE	ACTIVE	DEBIT-ORI	Manufacturing & Construction		
6	RISK	300	WALVISBA	Erongo	19360127	88	FEMALE	DOMESTIC	SENIOR CE	206	MONTHLY	CANCELLE	LAPSE	CANCELLE	DEBIT-ORI	Manufacturing & Construction		
7	RISK	300	WALVISBA	Erongo	19360129	88	FEMALE	THERAPIS	NO QUALI	206	MONTHLY	CANCELLE	LAPSE	CANCELLE	DEBIT-ORI	Manufacturing & Construction		
8	RETIREME	500	Grootfont	Otjozondj	19360902	88	MALE	CLEANER	NO QUALI	206	MONTHLY	ACTIVE	ACTIVE	CANCELLE	CASH	Manufacturing & Construction		
9	RISK	500	WALVISBA	Erongo	19360918	88	FEMALE	SELF EMPL	SENIOR CE	261	MONTHLY	ACTIVE	ACTIVE	ACTIVE	DEBIT-ORI	Manufacturing & Construction		
10	RISK	500	OSHAKATI	Oshana	19360928	88	FEMALE	UNEMPLO	SENIOR CE	323	MONTHLY	ACTIVE	ACTIVE	ACTIVE	DEBIT-ORI	Manufacturing & Construction		

Figure 2: Dataset with missing values after imputation

3.6.3 Replacing outliers

The dataset had an outlier with a basic salary value of -11470.42. To address the outlier in the Earning Basic Salary attribute, specifically the single negative value of -11470, the researcher removed it from the dataset. Negative salaries are not realistic and likely represent a data entry error. Since it could not be corrected due to a lack of reliable information, removing the outlier ensured that it would not distort the statistical analysis. This change improved the dataset's correctness and integrity, allowing for more accurate research conclusions.

3.6.4 Feature engineering

In the data transformation phase of the pre-processing for the analysis, the researcher employed feature engineering techniques to enhance the dataset's usability for machine learning. The researcher derived two new variables by grouping towns based on their respective regions and categorising occupations according to their occupational sectors. This was done through existing attributes in the dataset, such as the names of towns and the specific job titles of individuals. This transformation not only enhanced the dataset's structure and interpretability but also expanded the information accessible for future modelling,

allowing for more in-depth assessments of regional and occupational dynamics within the study.

3.6.5 Encoding of categorical variables

Categorical variables were converted to numerical variables using one-hot encoding, which is ideal for nominal categories. This method creates dummy variables (0, 1) based on the categories of a feature. First, categorical variables are converted into integers, typically based on their alphabetical order, with integer values ranging from 0 to n-1, where n is the number of unique classes.

For example, as per Figure 4 below, gender has two distinct characteristics: male and female. If a code of 0 means "not female," then a value of 1 means "female". According to Kuhn and Johnson (2013), a code of 1 indicates "male," while a code of 0 indicates "not male."

One-hot encoding is an effective approach for transforming categorical data into numerical values for machine learning models. However, this technique can result in strongly linked characteristics called collinearity. This happens when the same information is represented several times using the new binary variables. Collinearity might complicate interpreting the individual effects of predictor factors and impact parameter estimates. To reduce collinearity, it may be required to delete one-hot encoded features with a strong correlation. For example, if the male column is 0, recording a 1 for male already indicates whether the person is female. Double representation can cause instability in modelling (Kuhn & Johnson, 2013).

Gender
Male
Female
Male
Male
Female
Male

→

Male	Female
1	0
0	1
1	0
1	0
0	1
1	0

Figure 3: One-Hot encoding example. Source: Hutcheson (2011)

3.7 Feature Selection

Reducing dimensionality in machine learning can significantly improve algorithm performance. Feature selection is a frequent pre-processing way to do this. This method

identifies key dependencies and correlations between input and target features to reduce redundant features, reduce learning time, improve classification accuracy, and prevent overfitting (Haar et al., 2019).

In supervised learning, identifying the most essential characteristics for predicting the target feature is crucial, as each feature is assigned a class label. Feature selection involves selecting a subset of original features based on specific evaluation criteria. Feature selection offers benefits such as improved data quality, reduced computing time, enhanced prediction performance, and efficient data collection.

3.7.1 Chi-Square

The Chi-Square filter approach is commonly used for feature selection in machine learning. The Chi-Square statistical test assesses dependence between two categorical variables. This test is used in feature selection to evaluate the correlation between input features and target variables in classification problems. The test uses the Chi-Square statistic to determine which traits are most useful and relevant for classification. This strategy is frequently used in conjunction with other feature selection techniques to produce the best results (Singhal & Rana, 2015).

Despite having only 12 variables after dummy construction, feature engineering methods were applied to select relevant features for predicting the target variable. The Chi-Square test was used to compare models by evaluating the relationship between factors and the target, ranking the variables based on their importance, and assessing their dependence on the target.

3.8 Procedure of ML application

Machine learning relies heavily on using various predictive algorithms. Predictive modelling uses mathematical approaches to analyse a dataset with a target variable and predictors (Cheek et al., 2023). The approach aims to use statistical tools to select the optimal model for predicting performance. A model's performance is assessed by its ability to accurately forecast unknown data (Breiman, 2001; Kuhn & Johnson, 2013). According to Kuhn and Johnson (2013) and Cheek et al. (2023), the process of using machine learning to predict lapse rates in life insurance involves the following important steps:

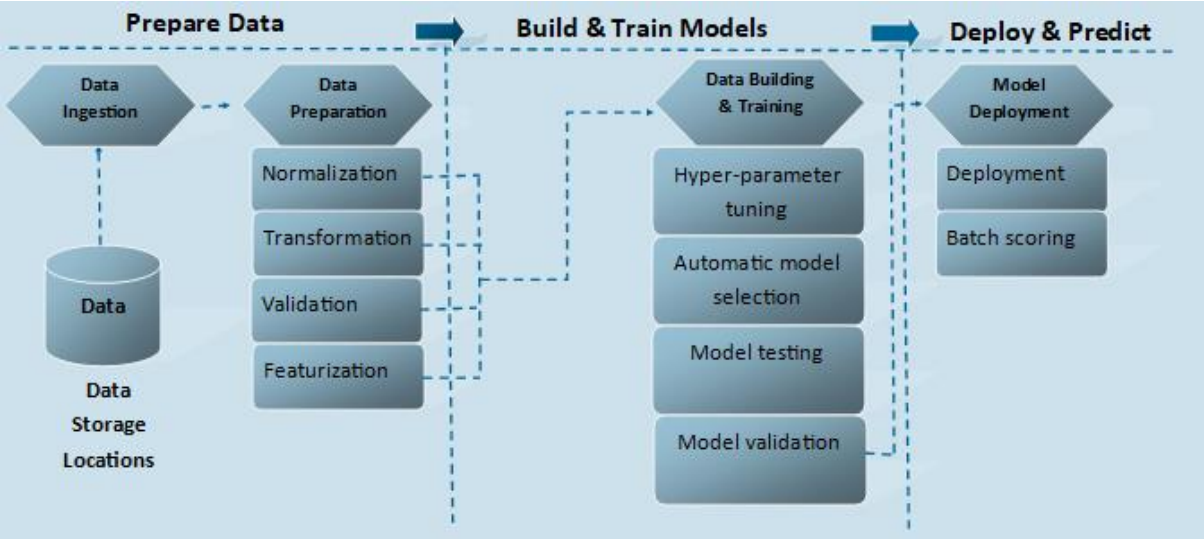


Figure 4: Machine learning steps: Source: Author

3.8.1 Data collection

The initial stage of the study involved the collection and organisation of relevant data, which included information on policyholders, policy details, and historical lapse records. This data was obtained directly from Sanlam Namibia, serving as the primary and trusted source to ensure the quality and reliability of the prediction model. The data collection process played a critical role in the development of an effective machine learning model, as high-quality input data is essential for generating accurate and meaningful predictions (Hastie et al., 2009). As emphasised in the literature, gathering data from reliable and authoritative sources enhances the validity and robustness of predictive modelling efforts (Smith, 2023; Johnson et al., 2019).

3.8.2 Data pre-processing

This involves handling missing values, outliers, cleaning and structuring the data, as well as introducing additional variables or features relevant to the prediction task. This step collectively forms a crucial foundation for effective data analysis and modelling.

3.8.3 Feature selection

This pertains to the use of feature selection approaches to determine the most significant aspects for lapse rate prediction. This can be accomplished by techniques such as correlation analysis, recursive feature elimination, or other feature selection algorithms (Armas et al., 2022).

3.8.4 Splitting the dataset

The dataset was divided into two sets: training and testing. The training set was used to train the machine learning model, while the test set was used to assess its performance (Lawson & Smith, 2023).

3.8.5 Model selection and training

Considering the features of the dataset and the unique needs of the prediction task, quite a few machine learning algorithms were evaluated, including Linear Regression, Random Forests, and Gradient Boosting. The models were trained on a portion of the dataset. At this phase, the selected machine learning model was trained using the training dataset. The model would learn the patterns and relationships between the input features (past data) and the target variable.

3.8.6 Model evaluation

To evaluate the trained models, metrics such as accuracy, precision, recall, and F1 score were used. This evaluation was performed on a separate test subset to ensure an unbiased performance assessment on new data.

3.8.7 Model improvement

Iterating and refining the model is an important aspect of designing an accurate and reliable prediction machine learning model. If the model's performance is not adequate, it may be essential to return to the previous steps and iterate on them to enhance the model. This could include gathering more data using an alternative algorithm or fine-tuning the model's parameters. It is vital to note that the algorithm's performance is dependent on the quality and representativeness of the dataset, as well as the machine learning model's proper selection and tuning. Regularly monitoring and upgrading the algorithm with new data will assist in ensuring its usefulness in predicting the lapse rate in the future.

3.8.8 Model deployment and monitoring

The greatest model was put into operation to forecast lapse rates in life insurance policies at Sanlam Namibia. This could include incorporating the model into an insurance company's risk management processes or developing a user-friendly interface for policyholders to access predictions. Continuous monitoring of the model's performance can be implemented, and any necessary adjustments or refinements will be made to ensure its accuracy and

effectiveness over time. This iterative process of monitoring and adaptation ensures that the model remains reliable and aligned with the evolving dynamics of the insurance landscape.

3.9 Model selection and training

In this study, an 80/20 data partitioning strategy was used, where 80% of the data was allocated to train the models and 20% for validation. The training set allowed the models to learn patterns and relationships within the data, while the validation set was used to assess model performance and prevent overfitting. This approach also facilitated hyperparameter tuning to improve the models' predictive accuracy and generalisation to unseen data, thereby ensuring robust and reliable outcomes.

3.9.1 Performance evaluation criteria

All models were evaluated based on their accuracy and statistical power. The models were critically evaluated using a confusion matrix, also known as a classification matrix. A confusion matrix evaluates classifier models and offers insights into their predictions. This study thoroughly examined numerous critical metrics for evaluating predictive performance. It looked specifically at accuracy, sensitivity, specificity, positive predictive values, and negative predictive values. This extensive investigation offers an in-depth comprehension of the prediction capabilities of the model.

The accuracy rate, which indicates the proportion of correct predictions out of all cases. As stated by Rung et al. (2020), was calculated using the classification matrix and the formula below.

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

Where:

TP = True Positive

TN = True Negative

FP =False Positive

FN = False Negative

Precision or positive predicted value represents the ratio of true positives to the total number of positive predictions. It indicates how many of the predicted positives were correct.

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

Sensitivity or recall is defined as the ratio of true positives to the total of actual positives. A low recall signifies a high rate of false negatives.

$$Recall = \frac{TP}{TP+FN\ n},$$

Specificity is determined by dividing the number of true negatives by the total number of negatives. A higher specificity suggests that the model performs well in detecting true negatives.

$$Specificity = \frac{TN}{TN+FP},$$

Misclassification error (ME) represents the fraction of misclassified predictions in relation to the total number of observations.

$$ME = \frac{TN+FN}{TN+TP+FN+FP},$$

The F1-score, also known as the F-measure, quantifies the balance between precision and recall, and is given by the formula below.

$$F1_{score} = 2 * \left(\frac{Precison*Recall}{Precision\ +Recall}\right),$$

3.10 Data analysis process

The data analysis process involved multiple stages. First, data pre-processing was conducted to handle anomalies such as missing values and outliers. For instance, a negative basic salary value (−11,470.42) was identified and removed as it was unrealistic and could not be corrected. Missing age values were imputed by calculating the age from the date of birth, thereby improving data completeness and consistency. Feature engineering techniques were then applied to create new variables, such as region and occupation sector, derived from

existing attributes like town names and job titles. These transformations enhanced the predictive capacity of the models.

Following data preparation, the dataset was split into training and testing sets using an 80:20 split ratio, and cross-validation was used to ensure robust model performance. Machine learning algorithms, including logistic regression, support vector machine, random forest, and gradient boosting, were trained and tested on the dataset. Models were evaluated using performance metrics such as accuracy, precision, recall, and F1-score. These metrics provided insights into each model's ability to correctly classify policyholders as likely to lapse or not, supporting objective comparisons.

To ensure the validity of the findings, real-world data from a credible source (Sanlam Namibia) were used, and model evaluations followed industry-standard practices. Reliability was maintained through consistent application of preprocessing methods and algorithm parameters, as well as repetition of analyses to check for consistency in outcomes.

3.11 Software

In this study, Power BI and Python were used as tools for Exploratory Data Analysis (EDA), thereby providing useful insights into the information through visualisations and summaries, and assisting in the identification of patterns, trends, and irregularities. Julia programming language was used as a tool to build and evaluate predictive models by exploiting its comprehensive libraries and computational efficiency. These tools enabled the development of accurate and scalable models, which provided deeper insights into the study concerns.

3.12 Ethical considerations

Ethical considerations were strictly observed throughout the research process. The data used in this study were obtained with formal permission from Sanlam Namibia, and all datasets were anonymised to protect the privacy of individual policyholders. No identifiable personal information was included in the analysis. Ethical clearance was obtained from the university's research ethics committee prior to commencing the study. The principles of confidentiality, informed consent, and non-maleficence were upheld, and the data were stored securely with limited access to ensure confidentiality. Since the study did not involve direct interaction with human subjects, the risk to participants was minimal, and ethical integrity was preserved throughout.

3.13 Chapter summary

This chapter outlined the methodological framework adopted to predict lapse rates in life insurance policies using machine learning, with a specific focus on Sanlam Namibia. Guided by a positivist paradigm and a quantitative approach, the study employed an experimental and predictive modelling design. The methodology was structured into multiple phases, including data collection, preprocessing, feature selection, model training, evaluation, and deployment. Data obtained from Sanlam Namibia served as the foundation for model development, which was conducted using robust data analysis techniques and supported by ethical research practices. This structured approach ensured that the research was scientifically sound, ethically responsible, and aligned with best practices in predictive analytics. Furthermore, the methodology was designed to meet the research objectives by enhancing prediction accuracy and supporting risk management strategies within the insurance sector. The table below summarises the relationship between the study components, tools used, and the anticipated outcomes.

Table 1: Summarized methodology

Research Questions	**Tools/Tasks/Activities**	**Outcome/Criteria**
What is the performance of the four commonly used ML models, namely logistic regression, support vector machine, random forest and gradient boost in predicting life insurance lapses?	Data preprocessing using Julia Model training and evaluation using algorithms (e.g., Logistic Regression, Random Forest, Gradient Boost) Model performance analysis using metrics like accuracy, precision, recall, F1-score	Identification of the best-performing ML model based on evaluation metrics
Which key features in the dataset significantly influence the prediction of life insurance lapses?	Feature importance analysis using chi-square or decision	List of ranked key features and analysis of their

	tree	contribution to predictive performance improvements
How do policyholder characteristics such as age and occupation affect the likelihood of a life insurance policy lapse?	Exploratory Data Analysis (EDA) using Python. Statistical analysis (e.g., correlation and crosstabulation) on policyholder data Predictive modelling using relevant features	Insights into the role of policyholder characteristics and their predictive power in assessing lapse risk

CHAPTER FOUR

RESULTS ANALYSIS

4.1 Introduction

This chapter discusses the process of developing predictive models. The chapter begins with an Exploratory Data Analysis (EDA) to identify relevant patterns and relationships in the dataset. The EDA serves as the foundation for comprehending the data and guiding the model development process. Following the EDA, the chapter evaluates the predictive models developed. The findings section contains a detailed analysis and evaluation of the models' performance during the training and validation stages. The discussion focuses on the relevance of these findings, which provide insight into the models' effectiveness and limitations. Comparisons of training and validation results are also shown, highlighting the models' capacity to generalize to new data and their potential practical applications.

4.2 Impact of policyholder characteristics on lapse rate

4.2.1 Target variable

The target variable was constructed as a binary variable, indicating whether an insurance policy status had lapsed (coded as 0) or not (coded as 1) at the date of analysis. Figure 5 below presents the distribution of active contract policies against lapsed ones. The comparison between active and lapsed policies revealed a significant retention issue, with 10,547 lapsed policies compared to 38% (6,428) active ones, indicating that over 62% of policies have lapsed. This imbalance highlights the urgent need for initiatives to reduce the high lapse rate and increase policyholder retention.

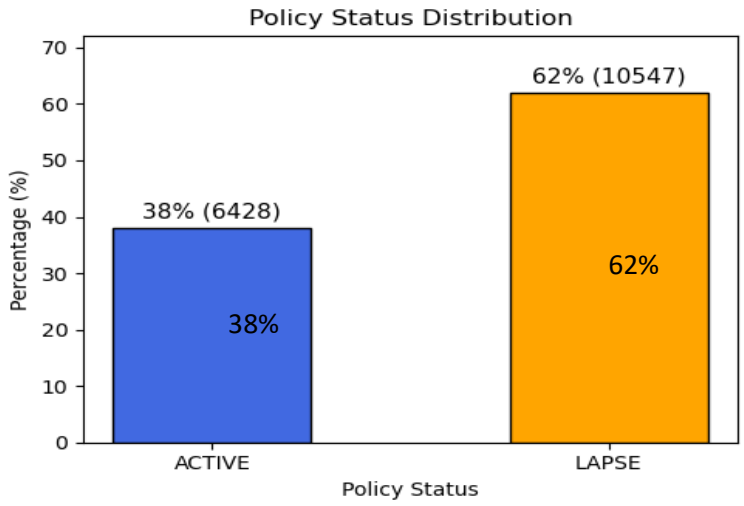


Figure 5: Target variable. Source: Author

4.2.2 Demographics distribution

Figure 6 below shows the gender distribution of lapsed life policies, with 53% held by males and 47% by females. This indicates that male policyholders are slightly more likely to lapse their policies than females, suggesting possible gender-related trends in policy management or financial behaviour, though the difference is not substantial.

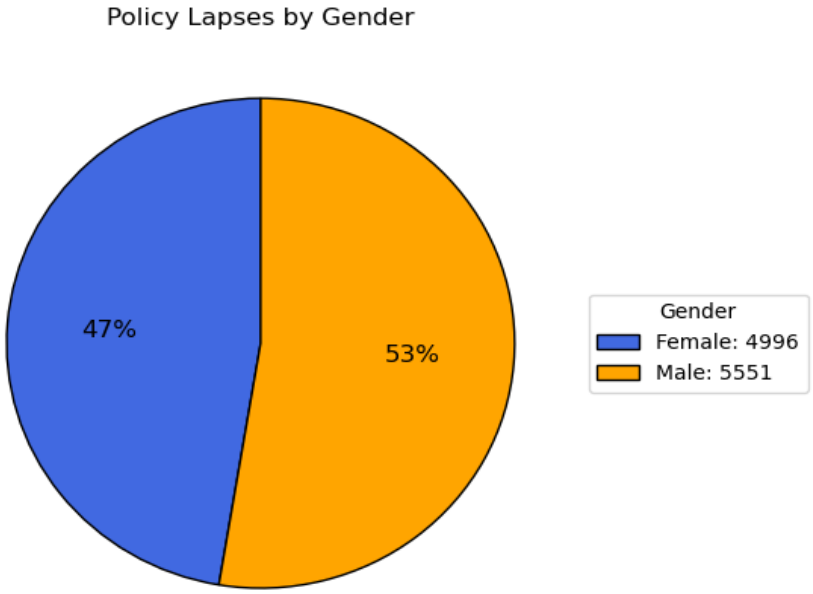


Figure 6: Lapses and Gender. Source: Author

Figure 7 below illustrates the distribution of life insurance policy lapses across various regions in Namibia, with Khomas having the highest number of lapses (31%), likely due to its large population or economic factors. Erongo and Oshana also show relatively high lapse rates, with 15% and 14% lapses, respectively. In contrast, regions like Zambezi, Kavango West, and the international category have significantly lower lapse percentages, reflecting regional differences in policy management or economic conditions. Additionally, 21% of lapses fall under the "No Address" category, possibly due to incomplete or outdated client information.

Figure 8 below shows a map of Namibia with its respective regions, thus providing a clear visual representation of the country's geographical layout. This map highlights the spatial distribution of various regions across the country. These visualisations collectively emphasise the geographical variation in policy lapses, which is crucial for understanding regional trends and formulating strategies to address lapses in life insurance policies.

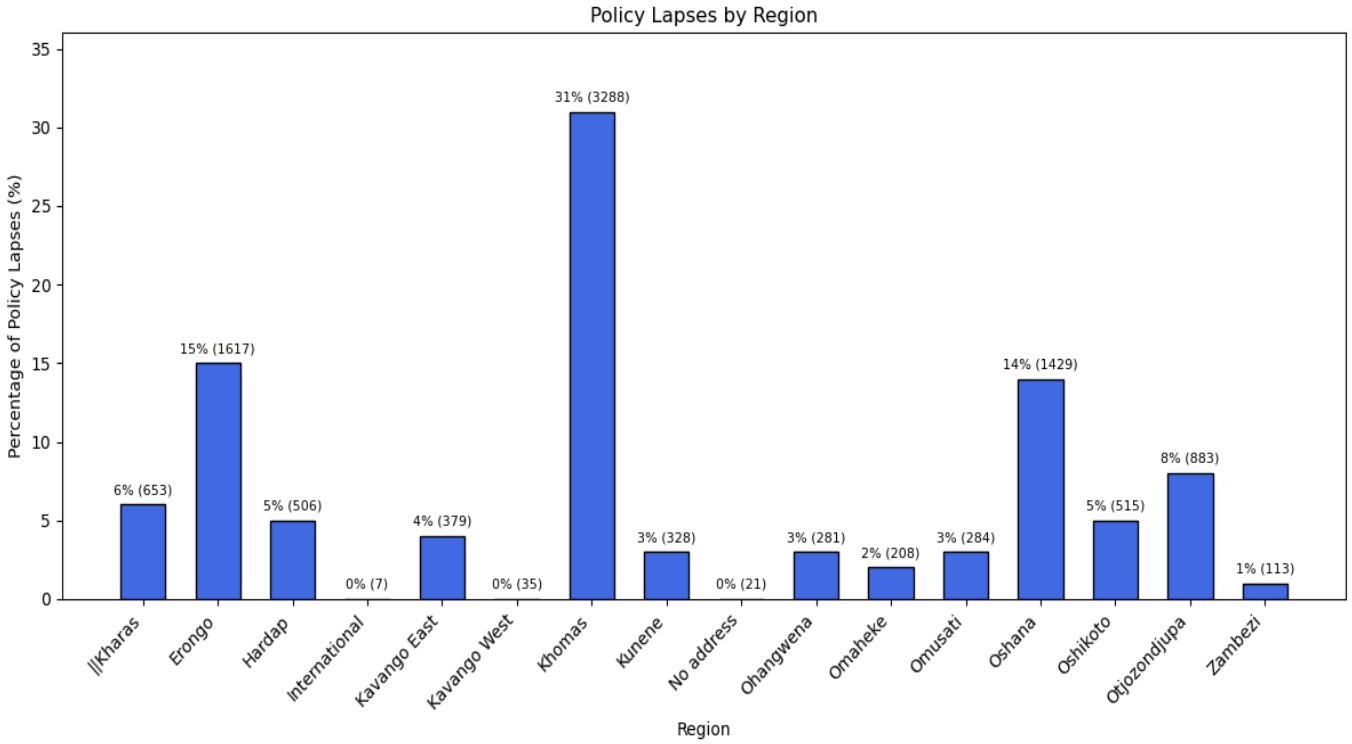


Figure 7: Lapses by Region. Source: Author

Figure 8: Namibia Map with regions: Source: (www.editablemaps.com)

Figure 9 below depicts the distribution of life insurance policy lapses across various age groups, revealing that the 50+ age group experiences the highest lapse rate, with 35% lapses. This trend may be attributed to life stage transitions such as retirement or reduced income, which can affect the ability to maintain premium payments. The 40-44 and 35-39 age groups also demonstrate elevated lapse counts, with 17% and 16% lapses, respectively, indicating potential financial pressures during midlife. In contrast, the youngest age groups (0-19 years) show negligible lapses, which is likely due to lower policy ownership among this demographic. This distribution suggests that age plays a critical role in policy lapse patterns, possibly influenced by economic and life stage considerations.

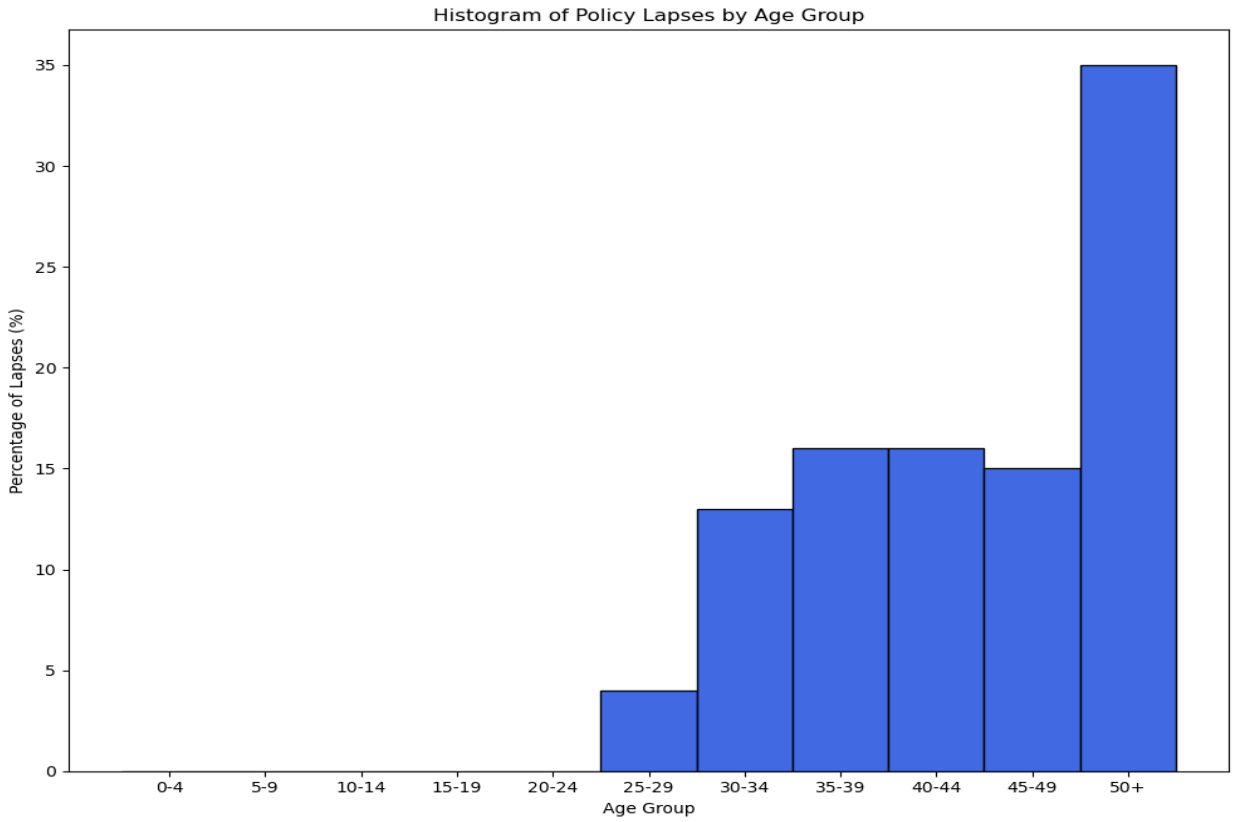


Figure 9: Lapses by Age group. Source: Author

4.2.3 Socio-economic factors distribution

Figure 10 below presents the distribution of policy lapses based on the education level of policyholders, revealing that individuals with a Senior Certificate (Grade 12) account for the highest number of lapses, totalling 56%. Those with a 3-year Diploma also represent a significant portion, with 26% lapses. Policyholders without any formal qualification show 9% lapses, suggesting a correlation between lower education levels and higher lapse rates. Individuals with Doctorate and Associate qualifications, on the other hand, have the lowest

lapse rates, showing that higher educational achievement may correlate with better policy retention.

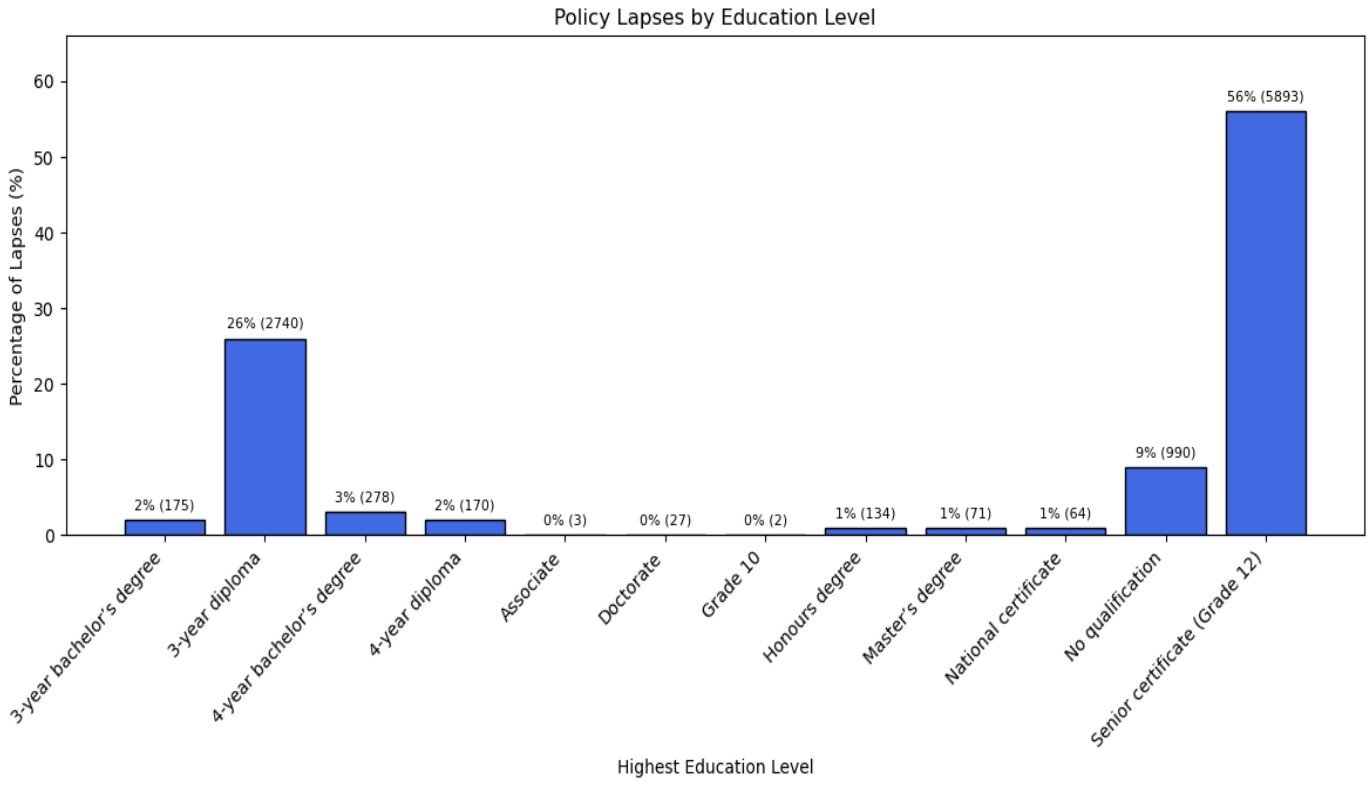


Figure 10: Lapses by Education Level. Source: Author

The bar graph in Figure 11 below presents the distribution of insurance policy lapses by various income brackets. The highest percentage of lapses occurs in the 10,000–19,999 income bracket, with 55% lapses, followed by the less than 10,000 bracket, which has 31% lapses. Policyholders in higher income brackets, particularly those with basic salary earnings between 20,000 and 29,999 and above, exhibit significantly fewer lapses, with lapses recorded in the income groups, respectively. The lowest lapse percentage is seen among individuals earning above 50,000. This distribution suggests a potential inverse relationship between income level and policy lapse rates, where lower-income earners are more likely to lapse their life insurance policies, potentially due to financial strain or affordability issues. Higher-income earners demonstrate better policy retention, reflecting greater financial stability.

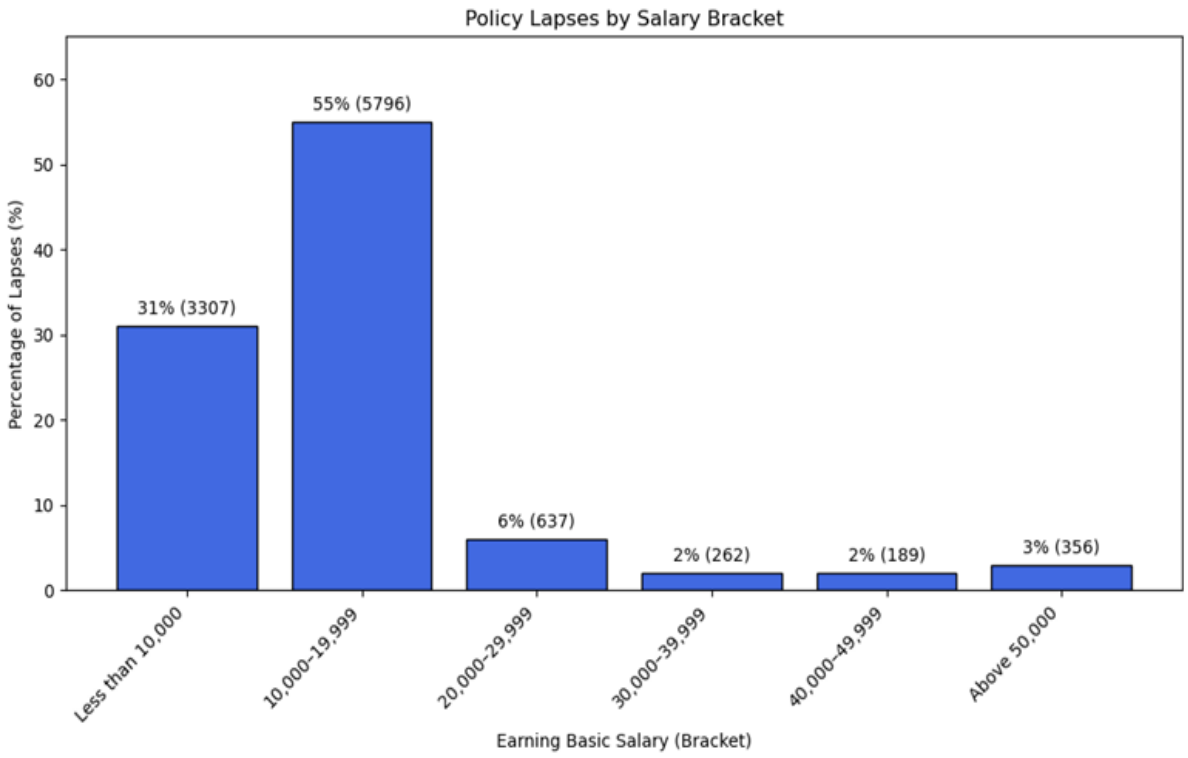


Figure 11: Lapses by Earning Basic Salary. Source: Author

Table 11 provides a breakdown of policy lapses among various occupational sectors. Occupations were grouped according to their sectors to form occupational sectors. The table revealed critical insights into how employment type may influence policy retention. The Hospitality & Tourism sector recorded the highest number of lapses, with 32% lapses, and this might be potentially due to the sector's volatility and seasonality, which can lead to income instability. This is followed by the Finance & Business sector, with 20% lapses, indicating that even within more stable sectors, policy lapses occur at a higher rate.

Sectors such as General Work and Government & Public Sector had high lapse numbers of 7% and 4%, respectively, indicating differing levels of financial stability within these industries. Importantly, 5% lapses are associated with unemployment, demonstrating a strong association between financial insecurity and policy lapses. The data indicate that occupational stability and income security play a key role in life insurance policy retention, with individuals in less stable sectors or who are unemployed being more likely to lapse their policies.

Table 2: Lapses by occupation sectors

Occupation Sector	Number of Lapses	Proportion of Lapses
Agriculture & Environmental	144	1%
Automotive Repair & Maintenance	18	0%
Beauty and Grooming	32	0%
Creative Arts & Media	36	0%
Education	1116	11%
Engineering & Technical	304	3%
Finance & Business	2050	20%
Food and Beverage	7	0%
General work	695	7%
Government & Public	441	4%
Healthcare	350	3%
Hospitality & Tourism	3354	32%
Information Technology	52	1%
Legal	65	1%
Manufacturing & Construction	430	4%
Public Safety & Security	193	2%
Sales & Marketing	438	4%
Transport & Logistics	273	3%
unemployed	549	5%
Total	**10547**	**100%**

4.2.4 Policy characteristics distribution

Figure 12 describes the distribution of policy lapses by several payment methods, highlighting differences in lapse rates. The debit-order payment method had the most lapses at 56%, implying that the automated nature of this payment method may lead to a lack of active engagement in policy management as policyholders get disengaged over time. Stop order payments and cash have smaller lapse numbers (23% and 21%, respectively). These methods, which require more conscious and deliberate action by policyholders, may encourage more thoughtful policy administration, thus potentially lowering the probability of lapses when compared to debit orders. However, the lapse rates for these manual payment methods remain high, implying that the payment method alone may not entirely account for the policyholder's chance of lapse.

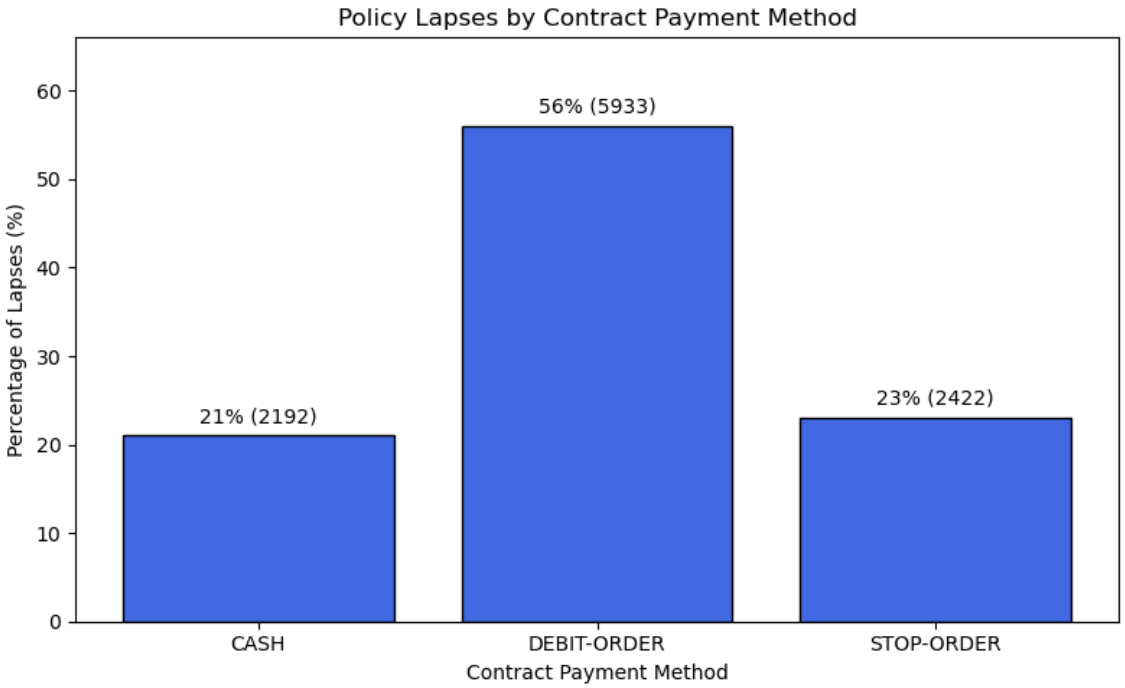


Figure 12: Lapses by Payment methods. Source: Author

The analysis of policy lapse data by product type in Table 3 revealed that most lapses occurred under the Risk product, which accounted for 6,362 policies, representing 60% of all lapses. This was followed by the Savings product, which contributed 1,395 lapses (13%), and Risk-and-savings-whole life, with 1,020 lapses (10%).

Products such as Retirement and Risk-and-savings-term contributed 321 (3%) and 395 (4%) lapses, respectively. Other product types like Capricorn-risk, Life-plan, Sansave, and Welwitschia recorded relatively low proportions of lapses, ranging between 1% and 2% each.

In contrast, several products experienced minimal lapse activity, each constituting 0% of total lapses, such as Income-protector (6 lapses), Risk-provider-term (1 lapse), Sanchild (16 lapses), Sanra (27 lapses), and Whole life (15 lapses).

These findings suggest that traditional risk-based products were more susceptible to lapses, possibly due to affordability, policyholder risk profiles, or product design. In contrast, products with low lapse rates may have exhibited stronger retention characteristics or targeted more stable client segments. The trends show that risk-based and savings-oriented products have higher lapse risks, which could be attributed to economic or financial restraints, although certain long-term plans retain more customers.

Table 3: Lapses by product type

Product Type		Number of Lapses	Proportion of Lapses
Capricorn-risk		125	1%
Children		184	2%
Fundamental needs (risk)		231	2%
Income-protector		6	0%
Life-plan		74	1%
Retirement		321	3%
Risk		6362	60%
Risk-and-savings-term		395	4%
Risk-and-savings-whole life		1020	10%
Risk-provider-term		1	0%
Risk-provider-whole life		66	1%
Sanchild		16	0%
Sanra		27	0%
Sansave		25	1%
Savings		1395	13%
Term		24	0%
Welwitschia		260	2%
Whole life		15	0%
Total		**10547**	**100%**

4.3 Key feature impacting lapse prediction

4.3.1 Features importance

The feature importance Figure 13 shows that payment-frequency, contract-payment-status-reason, and contract-status-reason are the most influential features, showing that they have a large impact on the model's predictive performance. Moderate contributors include educational level, gender, occupation, and occupation sector, whereas physical-town, product-type, region, age, and active-premiums-rate have little influence. This implies that the model's forecasts are primarily reliant on payment behaviour and contractual specifics, with demographic and regional factors contributing less. Low-importance features should be considered for exclusion during optimisation; however, this should be evaluated to avoid affecting the model's accuracy.

This depicts the feature variables and their impact on the insurance policy lapse rate. Therefore, it is crucial to analyse the data and how they affect the desired predicted outcome.

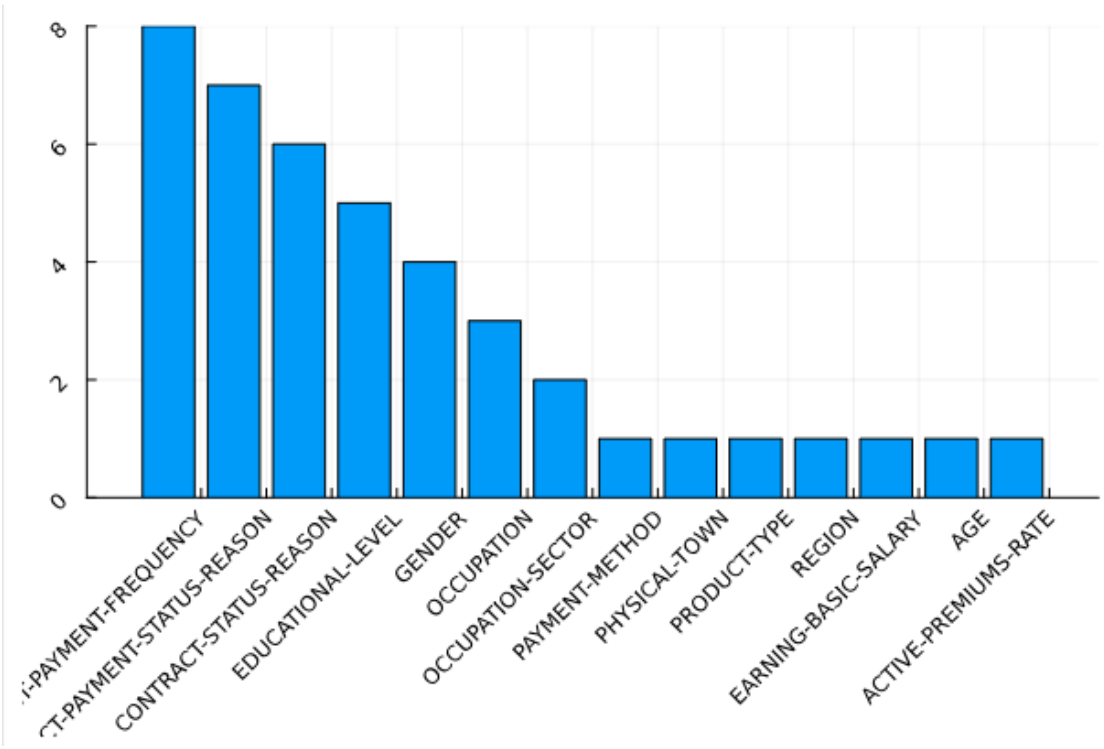


Figure 13: Feature Importance Selection Summary. Source: Author

4.4 Effectiveness of machine learning classification models

4.4.1 Logistic regression

After the data cleaning and pre-processing procedure was accomplished, the processed data were then split into the Training dataset and the Test dataset using the 'train_test_split' function from the MLUtils library in Julia, as depicted in Figure 14.

```
begin
Xtrain=X[:,1:Int(round((datasetSize*0.80)))]
Xtest=X[:,Int(round((datasetSize*0.20))):end]
end
```

```
2×3396 OneHotMatrix(::Vector{UInt32}) with eltype Bool:
 1  1  1  1  1  1  1  1  1  1  1  ⋅  1  1  …  1  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  1  1
 ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  1  ⋅  ⋅     ⋅  1  1  1  1  1  1  1  1  1  1  ⋅  ⋅
```

```
begin
Ytrain = Y[:,1:Int(round((datasetSize*0.80)))]
Ytest= Y[:,Int(round((datasetSize*0.20))):end]
end
```

32.0 ms

Figure 14: Splitting the processed data. Source: Author

The training set was used to train the Logistic Regression Model and learn the patterns within the data. Thereafter, the model was fed the Test data which serves as unseen or unbiased

data that helps us to gauge the performance of the newly trained model. The Logistic Regression model was trained and evaluated using the Test set data, and it recorded a prediction accuracy score of 48%. The Logistic Regression model's performance is very low, indicating that it performs poorly, as depicted in Figure 15, which shows that the model is not reliable and cannot be deployed for practical usage.

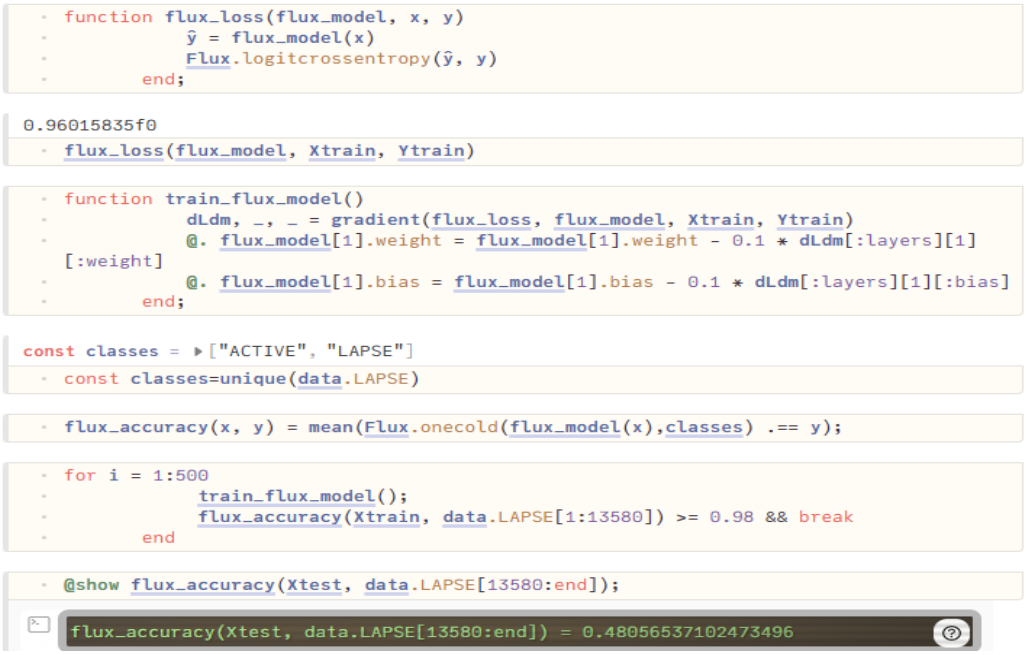


Figure 15: Logistic Regression Model training and testing. Source: Author

4.4.2 Support vector machine

The Support Vector Machine (SVM) model was evaluated using processed data as indicated in Figure 16, which provided an accuracy score of 87%, suggesting that the model successfully captures key decision boundaries. However, while the model demonstrates satisfactory performance on the training dataset, its ability to predict unseen data with the same level of accuracy remains moderate. This suggests that despite its strong performance in controlled settings, the SVM model may require further optimisation and validation to ensure its reliability and effectiveness in practical applications.

```
model =
  SVM(SVC, RadialBasis::KERNEL = 2, nothing, 1528, 13580, 2, [String7("LAPSE"), String7("AC
```

```
1 model= svmtrain(Xtrain,Ytrain)
```

```
1 ŷ, decision_values = svmpredict(model, Xtest);
```

```
1 # # Compute accuracy
2 @info "Accuracy: %.2f%%\n" mean(ŷ .== Ytest) * 100
```

```
Accuracy: %.2f%%
mean(ŷ .== Ytest) * 100: 86.77856301531213
```

Figure 16: Support Vector Machine training and testing. Source: Author

4.4.3 Random Forest

The processed data was also fed to the Random Forest model, and the performance of the model was evaluated. The performance of the Random Forest is very high, with an accuracy score of 93% in predicting lapse rates in life insurance policies. This indicates that the model correctly predicted whether a policy would lapse (or not) for 93% of the cases in the test dataset, as shown in Figure 17. The trained model's performance showed a high predictive accuracy on unseen data, thus proving its reliability in practical settings. Such a high accuracy suggests that the model is effective in identifying the patterns that influence the likelihood of policy lapses, which is crucial for assessing risks and improving retention strategies.

```
accuracy = nfoldCV_forest(Ytrain, Xtrain', n_folds, n_subfeatures)
```

```
Fold 1
Classes:  InlineStrings.String7["ACTIVE", "LAPSE"]
Matrix:   2×2 Matrix{Int64}:
 1322   249
   75  2880

Accuracy: 0.9284136102518781
Kappa:    0.8378530072666406

Fold 2
Classes:  InlineStrings.String7["ACTIVE", "LAPSE"]
Matrix:   2×2 Matrix{Int64}:
 1298   277
   60  2891

Accuracy: 0.9255413168360583
Kappa:    0.8304710273069263
```

Figure 17: Random Forest training and testing. Source: Author

4.4.4 Gradient Boosting

The Gradient Boosting model achieved an accuracy of 94% in predicting lapse rates in life insurance policies, as indicated in Figure 18. This suggests that the model correctly predicted whether a policy would lapse (or not) in 94% of the cases in the test dataset. This extremely high accuracy indicates that the Gradient Boosting model is performing well, capturing the underlying patterns and relationships that influence policy lapses, which is valuable for improving risk management and retention efforts within the insurance industry.

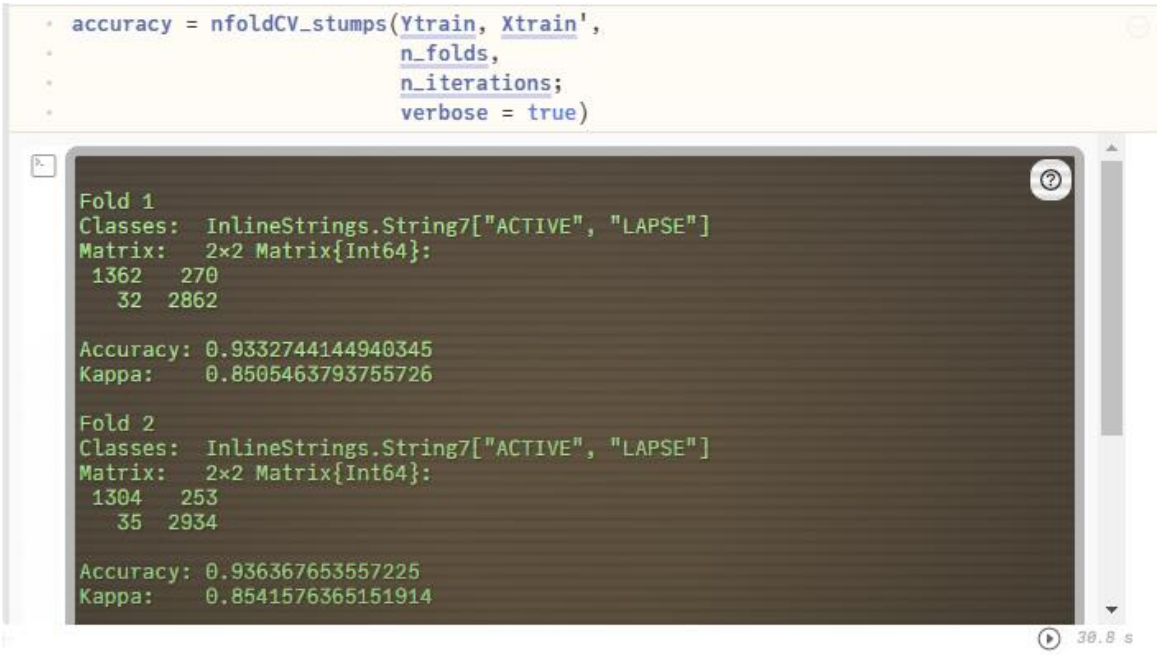


Figure 18: Gradient Boosting training and testing. Source: Author

4.4.5 Model comparisons

The trained models: Logistic regression model, Support Vector machine models, Random Forest, and Gradient boosting, respectively, were tested and evaluated so the best performing model can be selected. Figure 19 below indicates that Gradient Boosting and Random Forest are the most effective models, with accuracies of 94% and 93%, respectively. These models excel due to their ability to handle complex relationships and non-linear patterns in the data, which are often present in lapse prediction scenarios. The support Vector Machine performs well with an 87% accuracy, suggesting that it captures some key decision boundaries. However, Logistic Regression, with an accuracy of 48%, struggles to model the complexities of lapse rates, likely due to its linear nature. Ensemble methods like Random Forest and Gradient Boosting are recommended for robust predictions in this case. On this

basis, the Gradient Boosting was selected to perform the predictions for the insurance lapse rate.

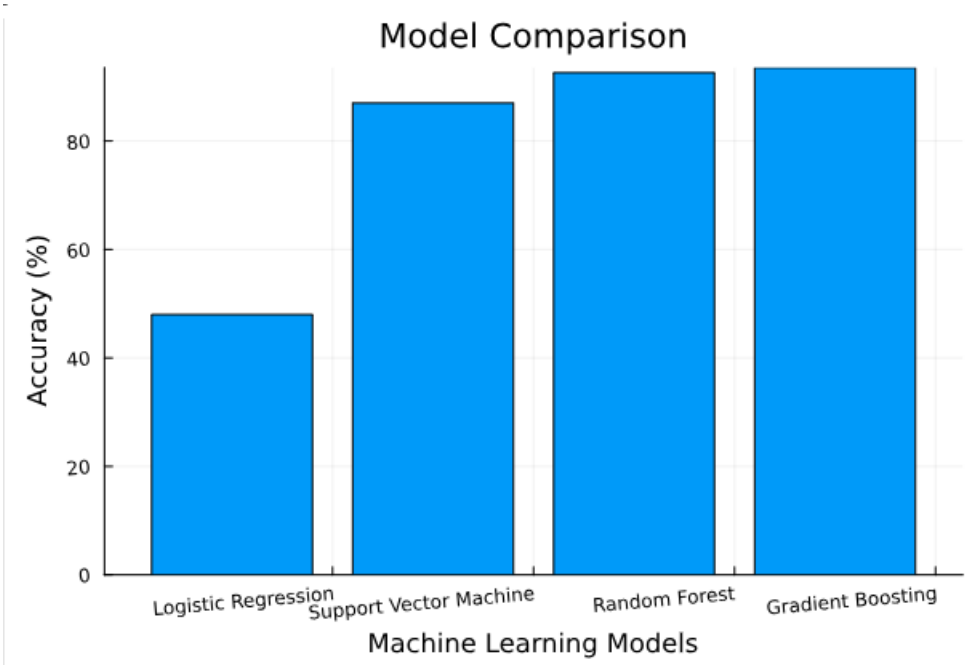


Figure 19: Model evaluation and selection. Source: Author

The efficacy of models for predicting policy lapses in Table 4 below varied significantly. Logistic Regression performed badly, with an accuracy of 48%, a precision of 47%, a recall of 50%, and an F1 score of 48%, indicating that it has difficulty distinguishing between active and inactive policies. The Support Vector Machine (SVM) fared significantly better, with an accuracy of 87%, a precision of 86%, a recall of 90%, and an F1 score of 89%, indicating strong dependability and sensitivity. Random Forest performed even better, with an accuracy of 93%, a precision of 93%, a recall of 94%, and an F1 score of 93%, indicating strong performance in anticipating lapses. Gradient Boost was the top model, achieving the highest accuracy (94%), precision (93%), recall (95%), and F1 score (94%), demonstrating remarkable balance and predictive capacity. Overall, Gradient Boost is the best model for predicting policy lapses, with SVM and Random Forest providing strong options, while Logistic Regression is unsuitable for the task.

Table 4: Model's performance on policy lapses prediction

Models		Logistic Regression (LR)		Support Vector Machine (SVM)		Random Forest		Gradient Boost	
Confusion Matrix		**Lapses**	**Active**	**Lapses**	**Active**	**Lapses**	**Active**	**Lapses**	**Active**
	Lapses	1172	2300	1254	224	1406	244	1354	254
	Active	53	1000	87	2774	73	2803	43	2875
Misclafication		52%		13%		7.4%		7.5%	
Accuracy		48%		87%		92.6%		93.5%	
Precision		47%		86%		93%		93%	
Recall		50%		90%		94%		95%	
F1 score		48%		89%		93%		94%	

4.5 Chapter summary

This chapter was centred on the analysis of the results from the study on predicting lapse rates in life insurance using machine learning algorithms, thereby offering a thorough overview of the findings obtained from the study. The primary goal of this research was to design and develop an accurate and reliable model for predicting lapse rates in life insurance using machine learning algorithms to effectively manage and retain active life insurance policies.

The beginning of the chapter presented an analysis of information about the dataset from which the feature variables were derived and considered within the research. This was trailed by a comparative analysis of the four models created: Logistic Regression model, Support Vector Machine, Random Forest, and Gradient Boosting model used Life Insurance lapse rate, highlighting the high performance of the two models. The results indicated that the Gradient Boosting model outperforms the rest of the models by a higher accuracy score of 94%. To conclude, the chapter offered valuable insight, emphasizing the effectiveness, high accuracy, and reliability of machine learning models' utilisation for predicting insurance lapse rate.

Furthermore, the analysis indicated that more than 60% of policies had lapsed, with considerable variances seen across demographic, economic, and product variables. Age, income level, education, occupation, and payment methods all have a substantial impact on lapse rates, but some product categories, particularly risk-based and savings-oriented plans, have a larger lapse risk. These findings emphasised the importance of focused tactics for addressing budgetary restrictions, increasing engagement in policy management, and designing products that better meet the requirements and conditions of policyholders. Addressing these difficulties is critical for increasing policyholder retention and decreasing lapse rates.

CHAPTER FIVE
RESEARCH CONCLUSION

5.1 Introduction

This chapter provides a comprehensive summary of the study, emphasising the key findings and conclusions derived from the research on predicting lapse rates in life insurance using machine learning algorithms. The chapter revisits the primary objectives of the study and highlights how they were achieved through the development and evaluation of predictive models.

Furthermore, this chapter discusses the significant contributions of the study to both academic literature and practical applications in the insurance industry, particularly in enhancing policyholder retention strategies. The chapter also outlines the limitations encountered during the research, which may have influenced the scope and generalisability of the results.

In addition, recommendations for future research and practical implementation are presented, offering insights into how insurance companies can leverage advanced analytics to better manage lapse rates. The chapter concludes with a summary that encapsulates the overall outcomes and implications of the study, reaffirming the importance of machine learning techniques in addressing policy lapse challenges.

5.2 Findings

Sub-research question one: *What is the performance of the four commonly used machine learning models, namely Logistic Regression, Support Vector Machine, Random Forest, and Gradient Boost, in predicting life insurance lapses?*

The outcome of the study showed that the Gradient Boost and Random Forest models outperform the Logistic Regression and Support Vector Machine models machine learning algorithm in terms of accuracy and computational efficiency. The two models can successfully predict the insurance lapse rate with minimal errors, thereby empowering stakeholders to make informed decisions pertaining to insurance policies.

The features were analysed and evaluated based on the impact they had on predicting the lapse rate of insurance policies. The features that impacted the lapse rate of insurance policies and were used to train the selected model from the data set were the gender of policy holders, age, geographical location, socio-economic factors (education level, financial

earnings, and occupation), and policy characteristics (Payment frequency, contract payment status, product type, premium rate, and contract status reasons).

Sub-research question two: *What are the key features in the dataset that significantly influence the prediction of life insurance lapses?*

Payment frequency, contract payment status, and contract status-reasons were identified as the most impactful features influencing model performance. These variables directly reflected the policyholder's engagement with premium payments and provided critical insights into lapse behaviour. Irregular payment frequencies and contract statuses indicating payment delinquency or inactivity strongly correlated with higher lapse rates, making them essential predictors for accurate modelling. The predictive strength of these features stems from their ability to capture real-time behavioural signals that precede policy termination, making them essential for early lapse detection and intervention strategies.

Education level, gender, and occupation were found to moderately influence lapse prediction. These demographic and socio-economic variables did not have a direct transactional link with the insurance product but affected policyholder behaviour through indirect pathways. For instance, individuals with higher education levels may possess greater financial literacy, enabling better understanding and appreciation of long-term insurance benefits, thus reducing the likelihood of lapses. Similarly, gender and occupation may influence income stability, financial planning behaviour, and attitudes towards risk, which in turn affect the consistency of premium payments. While these features were not the top contributors, their inclusion improved the model's robustness by capturing underlying behavioural patterns.

In contrast, features such as physical town, region, and product type had limited predictive power in the model. Although these variables added geographic or categorical detail, they lacked strong correlations with lapse outcomes. Their minimal variation across policyholders or weaker associations with financial behaviour reduced their overall contribution to the model's accuracy. Similarly, while product type offered some differentiation, particularly for risk-only versus savings-linked products, it did not consistently predict lapses with the same reliability as payment behaviour indicators.

In summary, payment behaviour-related features, including payment frequency, contract payment status, and contract status reasons stood out as the most impactful drivers of policy lapse predictions. Demographic variables contributed moderately, while geographic and

product-specific factors held lesser importance. This hierarchy of feature importance highlighted the need for insurers to monitor and intervene based on dynamic behavioural indicators to effectively mitigate lapse risks.

Sub-research question three: *How do policyholder characteristics such as age and occupation affect the likelihood of life insurance policy lapses?*

The analysis revealed that age had been a significant determinant in predicting the likelihood of life insurance policy lapses. The 50+ age group exhibited the highest lapse rate, accounting for 35% of all recorded lapses. This trend was likely due to life stage transitions such as retirement, reduced disposable income, or changing financial priorities, which may have influenced the ability or willingness of policyholders to continue paying premiums. The 40–44 and 35–39 age groups followed closely, with 17% and 16% of lapses, respectively. These age groups appeared to experience considerable financial pressures during midlife, including supporting dependents, repaying home loans, and managing healthcare expenses. Such obligations may have compromised their capacity to maintain long-term financial commitments like life insurance policies.

In contrast, the younger age groups, particularly those aged 0–19 years, recorded negligible lapse rates. This was likely because insurance policies for children were typically owned and managed by parents or guardians, who were responsible for maintaining the payments. As a result, these policies were less vulnerable to lapses driven by the insured individual's financial behaviour. These findings suggested that policyholder age and associated life stage factors had played a critical role in influencing policy lapse behaviour.

Occupational stability significantly influences life insurance policy retention. The Hospitality and Tourism sector recorded the highest lapse rate at 32%, likely due to the industry's income volatility and seasonal employment patterns. This instability can make it challenging for individuals to consistently meet premium obligations. Similarly, the Finance and Business sector accounted for 20% of policy lapses, demonstrating that even among more stable professions, financial overextension or shifting priorities may lead to discontinuation. Additional sectors such as General Work (7%) and the Government and Public Sector (4%), showed notable lapse rates, indicating that financial obligations or inadequate financial planning may affect even those with perceived job security.

Moreover, unemployment was associated with 5% of total lapses, underscoring the critical role of financial security in maintaining life insurance coverage. Individuals without stable income streams are more vulnerable to policy lapses due to affordability constraints. These findings highlight the need for insurance providers to consider policyholder occupation and employment status when assessing lapse risk, and to design interventions such as flexible payment structures or targeted financial literacy programmes for those in unstable or lower-income sectors to enhance policy retention.

The analysis of policy lapses by education level revealed a strong relationship between educational attainment and the likelihood of discontinuing life insurance policies. Policyholders with a Senior Certificate (Grade 12) had the highest lapse rate at 56%, followed by those with a 3-year Diploma at 26%, suggesting that individuals with mid-level education may face financial or informational barriers that hinder long-term policy retention. Additionally, 9% of lapses were recorded among those with no formal qualification, indicating that lower education levels may be associated with financial vulnerability and limited awareness of insurance benefits.

In contrast, individuals with higher qualifications, such as a Doctorate and associate degrees, showed the lowest lapse rates, pointing to a possible correlation between advanced education, financial literacy, and consistent premium payment behaviour. These findings emphasise that education plays a significant role in influencing insurance policy continuity and highlight the potential value of financial literacy programmes aimed at supporting less-educated policyholders.

Income level further demonstrated a strong inverse relationship with lapse rates. Most lapses occurred among policyholders earning between N$10,000 and 19,999 (55%), followed by those earning less than N$10,000 (31%). This indicates that individuals in lower income brackets are more likely to lapse on their policies due to affordability issues and financial constraints. Conversely, higher-income groups, particularly those earning above N$20,000 exhibited significantly lower lapse rates, with the lowest found in the N$50,000+ category. These findings underscore the importance of income stability in sustaining insurance coverage and suggest the need for flexible payment structures to support lower-income policyholders.

The analysis of payment methods revealed significant insights into how these factors influenced the likelihood of life insurance policy lapses. The findings showed that policyholders who used the debit-order payment method accounted for the highest proportion of lapses, at 56%. This method, being automated, required minimal ongoing effort from policyholders after initial setup. Although convenient, the automation appeared to contribute to a lack of active engagement in policy management. Over time, policyholders may have become disengaged, leading to missed payments or a general neglect of policy responsibilities. This suggested that the ease and passive nature of debit orders may have contributed to higher lapse rates, as policyholders were less likely to monitor and manage their policies actively.

In contrast, policyholders who paid via stop orders and cash demonstrated lower lapse rates, at 23% and 21% respectively. These manual payment methods required more deliberate and conscious action from the policyholders, which may have encouraged a higher level of attentiveness and responsibility in policy management. This more active involvement was likely associated with greater commitment to maintaining the policy, thereby reducing the risk of lapses. However, despite the relatively lower lapse rates for stop-order and cash payments, the findings indicated that lapse rates remained high across all payment methods. This implied that the payment method alone did not fully explain lapse behaviour. Other factors, such as demographic characteristics, financial literacy, income stability, and broader economic conditions, likely played a significant role. Therefore, while the payment method was an important factor, a comprehensive understanding of policy lapse required considering multiple interacting characteristics.

The analysis of policy lapses by product type showed that certain life insurance products had been more prone to lapses than others, indicating a strong relationship between product type and lapse risk. The Risk product recorded the highest number of lapses, accounting for 60% of all lapse cases, followed by the Savings product with 13%, and Risk-and-savings-whole life with 10%. These products, often associated with higher premiums or shorter-term financial goals, appeared more vulnerable to lapses, possibly due to financial strain or reduced long-term engagement from policyholders. In contrast, products such as Retirement and Risk-and-savings-term contributed smaller proportions of lapses (3% and 4%, respectively), while several niche products such as Income-protector, Sanchild, Sanra, and Whole life exhibited minimal lapse activity, each accounting for 0% of total lapses.

These findings suggested that traditional risk-based and savings-oriented products had higher lapse risks, potentially due to affordability issues, less stable client segments, or product features that did not promote long-term commitment. On the other hand, products with low lapse rates may have been better structured to retain customers or may have targeted more financially stable policyholders. Overall, the results indicated that policyholder characteristics as reflected in product selection had a significant impact on lapse behaviour, and that thoughtful product design and alignment with client needs were crucial for improving retention rates.

5.3 Contribution of the study

The study advances data science by demonstrating the superiority of ML algorithms, particularly Gradient Boost and Random Forest, in predicting lapse rates in life insurance. It highlights the importance of feature selection and shows how these models outperform traditional methods, offering insights into the practical application of ML in the insurance industry.

Furthermore, the study contributes significantly to the field of life insurance analytics by demonstrating the superiority of Gradient Boost and Random Forest models over traditional algorithms such as Logistic Regression and Support Vector Machines in predicting lapse rates with greater accuracy and computational efficiency. It identified key predictive features, including payment frequency, contract payment status, and contract status reasons, while highlighting the moderate influence of demographic and socio-economic factors such as education, income, and occupation. Additionally, the research revealed that risk-based and savings-oriented insurance plans had been more susceptible to lapses, providing actionable insights for targeted interventions. Focusing on the Namibian context, specifically Sanlam Namibia, the study offered localised knowledge on the drivers of policy lapses, enabling stakeholders to make data-driven decisions regarding policy design and risk management. The research also highlighted the practical application of machine learning (ML) in real-world settings, demonstrating how theoretical data science methodologies had been effectively applied to address industry-specific challenges in the life insurance sector.

5.4 Limitations

It is imperative that the limitations of the study are acknowledged and understood, as they are essential in the implementation of the designed ML model. The availability of quality data impacts the accuracy and dependability of the model's prediction significantly. Inaccurate, missing, or incomplete data entries could have influenced the model's predictive powers.

It is also vital to recognise the limitations concerning data availability, model selection, and other factors that can influence the model's effectiveness.

During the study, several data-related limitations were encountered and successfully addressed to ensure the integrity and reliability of the analysis. Firstly, the dataset contained an outlier in the Earning Basic Salary attribute, with a negative value of -11,470.42. Since negative salaries were unrealistic and likely indicated a data entry error, and because no reliable information was available to correct it, the outlier was removed. This decision prevented distortion in the statistical analysis and improved the overall quality of the dataset. Secondly, missing values were present in key attributes such as age. To resolve this, an imputation strategy was applied: the missing ages were calculated using the date of birth (DOB) by subtracting the year of birth from the year of data collection. This approach leveraged available contextual information to ensure completeness and accuracy of the dataset. Lastly, during the data transformation phase, feature engineering techniques were employed to improve the dataset's structure and usability for machine learning. Two new variables were derived by grouping towns into their respective regions and categorizing job titles into broader occupational sectors. These transformations enhanced the dataset's interpretability and provided additional insights for modelling, particularly in examining regional and occupational influences on lapse behaviour. Through these interventions, the researcher ensured a high-quality dataset that supported robust and valid analytical outcomes.

5.5 Recommendations

From the study findings, it is recommended to further explore and refine the model by discovering additional feature variables that can enhance the accuracy of the prediction model.

Moreover, implementing an iterative model validation process will be vital for ensuring reliability. Furthermore, regularly testing and updating the predictive models with real-world data will help identify potential biases or shortcomings in the algorithms used.

Lastly, collaborating with industry professionals could provide valuable insights into practical applications of these predictions in risk management strategies within insurance companies. Such collaboration ensures that academic research remains aligned with real-world needs and challenges.

5.6 Conclusion

In conclusion, applying machine learning (ML) methods to predict insurance policy lapse rates represents a significant advancement in the ML industry. ML models can leverage vast datasets to identify patterns and factors influencing policyholder behaviour. This predictive capability is particularly crucial in today's dynamic insurance landscape, where understanding customer retention is essential for maintaining profitability and sustainability.

The findings from these investigations underscore the importance of integrating technological innovations into traditional actuarial practices. By utilising ML algorithms, insurers can enhance their ability to forecast lapse rates, thereby enabling more informed decision-making regarding product offerings and customer engagement strategies.

Although challenges remain such as data quality issues and model interpretability, the potential benefits of employing ML in predicting lapse rates are substantial. The integration of these advanced methodologies not only promises improved operational efficiencies but also fosters greater resilience within the insurance sector in changing economic conditions. Future research should focus on refining these models further and exploring their applicability across diverse markets to fully realise their potential.

References

Andaria, R., Sumarti, N., & Puspita, D. (2025). Observation of calculated lapse rate for Sharia Life Insurance Data: A case study. *Jambura Journal of Mathematics, 7*(1), 79-83. *Jambura Journal of Mathematics, 7(1), 79-83.*, *7*(1), 79–83.

Anifalaje, K. (2021). Changing legal perspectives of the requirement of insurable interest in insurance contracts. *Commonwealth Law Bulletin*, *47*(2), 363–393.

Armas, E., Arancibia, H., & Neira, S. (2022). *Identification and forecast of the insurance lapses rate*. https://www.bing.com/ck/a?!&&p=5b202e5a4b1a11fcae646b996ce2e2751f8e9df4a82b55857052bce3ed72ea07JmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly93d3cubWRwaS5jb20vMjQxMC0zODg4LzcvNC8yMDQ

Arnold, O., Aghware, F. O., Okpor, M. D., Akazue, M. I., Malasowe, B. O., Aghaunor, T. C., & Onyemenem, S. I. (2025). Effects of data balancing in diabetes mellitus detection: A comparative XGBoost and random forest learning approach. *NIPES-Journal of Science and Technology Research*, *7*(1), 1–11.

Ayodele, T. O. (2021). New advances in machine learning. *New Advances in Machine Learning*, 19–20. https://www.bing.com/ck/a?!&&p=983d892e1e81f0b8efbddc439c3ae5b86486da9a1ea41f1f82aa35cfff4864dcJmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly93d3cuZGl2YS1wb3J0YWwub3JnL3NtYXNoL2dldC9kaXZhMjoxNzk2MzI3L0NPVkVSMDEucGRm

Baoilan, A. S., Barret, D. Z. D., Bohol, L. J. N., Bulala, F. M. C., Calderon, G. M. D., Conoman, C. N. P., & Flores, A. J. (2025). Evaluating Policy Lapse Behavior on Life Insurance Policyholders in Digos City. *International Journal of Research and Innovation in Social Science*, *9*(4), 4092–4111.

Bärtl, M., & Krummaker, S. (2020). Prediction of claims in export credit finance: A comparison of four machine learning techniques. *Risks*, *8*(1), 22. https://doi.org/10.3390/risks8010022

Bauer, D., Gao, J., Moenig, T., Ulm, E. R., & Zhu, N. (2017). Policyholder exercise behavior in life insurance: The state of affairs. *North American Actuarial Journal, 21(4)*, 485–501.

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2019). *A comparative analysis of XGBoost.* Universidad Autónoma de Madrid.

Biagini, F., Huber, T., Jaspersen, J., & Mazzon, A. (2021). Estimating extreme cancellation rates in life insurance. *Geneva Risk and Insurance Review*. https://www.bing.com/ck/a?!&&p=88bceeeb71a2c1c644c1a694bdf450c05d86f7edd5ac5a795b451277d6fe4079JmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly9vbmxpbmVsaWJyYXJ5LndpbGV5LmNvbS9kb2kvZnVsbC8xMC4xMTExL2pvcmkuMTIzMzY_bXNvY2tpZD0zMzA0Yjg0NTJjN2Q2OGU1MjgwMmFiZmYyZGJiNjkwNQ

Blier-Wong C., C.-H. L. L., & E. Marceau. (2021). *Machine learning in PC insurance: A review for pricing and reserving.* https://www.bing.com/ck/a?!&&p=4970dba7e5e7e99e7040cd1e518de691a685518d4db8a76ce37408ad3e830c74JmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly93d3cucmVzZWFyY2hnYXRlLm5ldC9wdWJsaWNhdGlvbi8zNDc5Nzc2NDVfTWFjaGluZV9MZWFybmluZ19pbl9QQ19JbnN1cmFuY2VfQV9SZXZpZXdfZm9yX1ByaWNpbmdfYW5kX1Jlc2VydmluZw

Bolancé, C., Guillen, M., & & Padilla-Barreto, A. E. (2016). Predicting the probability of customer churn in insurance. *In Modeling and Simulation in Engineering, Economics and Management: International Conference, MS 2016, Teruel, Spain,* 82–91.

Böök, B., Timmer, A., & de la Corte-Rodríguez, M. (2025). *A comparative analysis of gender equality law in Europe 2024: The 27 EU Member States, Iceland, Liechtenstein, Norway, and the United Kingdom compared.* https://www.bing.com/ck/a?!&&p=6b55a9a032987a3c319eb717758014a52ad5274c0cc8b21e9450335e39a56281JmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly9yZXNlYXJjaC1wb3J0YWwudXUubmwvZW4vcHVibGljYXRpb25zL2EtY29tcGFyYXRpdmUtYW5hbHlzaXMtb2YtZ2VuZGVyLWVxdWFsaXR5LWxhdy1pbi1ldXJvcGUtMjAyNC10aGUtLTI

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *In COLT '92: Proceedings of the Fifth Annual 28 Workshop on Computational 80 Learning Theory*. https://www.bing.com/ck/a?!&&p=4f15e8cf779ea4a9f0e17479d95ec6238c84844f7d20f0abffa063790c0bc922JmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly9kbC5hY20ub3JnL2RvaS8xMC4xMTQ1LzEzMDM4NS4xMzA0MDE

Braun, A., Cohen, L. H., Malloy, C. J., & Malloy, C. J. (2019). *Saving face: A solution to the hidden crisis for life insurance policyholders*. https://www.bing.com/ck/a?!&&p=be60ff37a31c165ad12f159d1f796406a1950d84dae767e420de48e70ae28586JmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly93d3cudndybS5ydy5mYXUuZGUvZmlsZXMvMjAxOC8wOC9FMV8xLUJyYXVuLUNvaGVuLU1hbGxveS1YdS5wZGY

Breiman, L. (2001). Random forests. *Machine Learning*, 1–33. https://www.bing.com/ck/a?!&&p=6fa5cd07baad0f566547de7918e66287cc5b23b261a749222ae270ccbe010f49JmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly93d3cuc2NpcnAub3JnL3JlZmVyZW5jZS9yZWZlcmVuY2VzcGFwZXJzP3JlZmVyZW5jZWlkPTM5MzAwMzQ

Brown, A., & Green, P. (2020). Machine learning in insurance. *The British Actuarial Journal*, 285–339.

Bruun, S. B., Lioma, C., & Maistro, M. (2024). Recommending target actions outside sessions in the data-poor insurance domain. ACM transactions on recommender systems. *ACM Transactions on Recommender Systems*, *3*(1), 1–24.

Burri Rama Devi, B.-R. B. R. R. and S. R. B. (2019). Insurance claim analysis using Machine Learning Algorithms. *International Journal of Innovative Technology and Exploring Engineering*, *8*(6S4), 577–582. https://doi.org/10.35940/ijitee.F1118.0486S419

Deng, Y. (2022). *Longevity risk modelling, securities pricing, and other related issues.* https://www.bing.com/ck/a?!&&p=3fcfc52b98f31ee8eea21ed74889dceab20e1ae8919ea080a46bf0c4d728532eJmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly93d3cuc2VtYW50aWNzY2hvbGFyLm9yZy9wYXBlci9Mb25nZXZpdHktcmlzay1tb2RlbGluZyUyQy1zZWN1cml0aWVzLXByaWNpbmctYW5kLURlbmcvN2JiZGJkOGViMjM5ZTAxMWJlYjYxNzhkZDJjZjU5OTdiMWQ0OTI5Mg

Diana, A., J. E., Griffin, J., J. Oberoi, & J. Yao. (2019). *Machine-learning methods for insurance applications: A survey.* Society of Actuaries.

Dudovskiy, J. (2021). Implications of individual resistance to change. *Research methodology*. https://www.bing.com/ck/a?!&&p=a3e45dc004ce91caa9a0cb5df2bfdf206fd08a4314fcce58f93b610a0d013389JmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly93d3cucmVzZWFyY2hnYXRlLm5ldC9wdWJsaWNhdGlvbi8zMzkxOTAzMzZfUmVzaXN0YW5jZV90b19DaGFuZ2VfQ2F1c2VzX2FuZF9TdHJhdGVnaWVzX2FzX2FuX09yZ2FuaXphdGlvbmFsX0NoYWxsZW5nZQ

Eling, M., & Kochanski, M. (2013). Research on lapse in life insurance: What has been done and what needs to be done? *The Journal of Risk Finance*, *14*(4), 392–413. https://doi.org/10.1108/JRF-12-2012-0088

Ferrario, A., Noll, A., & Wuthrich, M. V. (2023). Insights from inside neural networks. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3226852

Fong, J. H. (2015). Beyond age and sex: Enhancing annuity pricing. *The Geneva Risk and Insurance Review, 40(2)*, 133–170.

Fontinelle, A. (2021). Life insurance guide to policies and companies. *Life Insurance Guide to Policies and Companies.* https://www.bing.com/ck/a?!&&p=5f0410aaf5a9f4db653d63b01038444675903e751b093bb277492f24ad5bec09JmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly93d3cuc3R1ZG9jdS5jb20vZW4tdXMvZG9jdW1lbnQvdW5pdmVyc2l0eS1vZi1zb3V0aC1kYWtvdGEvaW5zdXJhbmNlL2NvbXByZWhlbnNpdmUtZ3VpZGUtdG8tbGlmZS1pbnN1cmFuY2UtcG9saWNpZXMtYW5kLXByYWN0aWNlcy8xMjUzMjExODI

Frees, E. W., Derrig, R. A., & Meyers, G. (2022). *Predictive modelling in actuarial science. Predictive modelling applications in actuarial science*. https://www.bing.com/ck/a?!&&p=73e416a4a6a0f28051f3c5d4e52166b55f79b9d2b55448db5f60d1de54c583c7JmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly93d3cuY2FtYnJpZGdlLm9yZy9jb3JlL2Jvb2tzL3ByZWRpY3RpdmUtbW9kZWxpbmctYXBwbGljYXRpb25zLWluLWFjdHVhcmlhbC1zY2llbmNlLzgyQTdFNUI5NTQ1QjFCRUREN0Y2QTc0OTEyNjUxMTNB

García, S., J., Luengo, & F. Herrera. (2015). Data preprocessing in data mining. *American Actuarial Journal*, 34–49.

Gelman, A. (2010). Missing-data imputation*. Data Analysis Using Regression and Multilevel/Hierarchical Models*, 529–544.

https://www.bing.com/ck/a?!&&p=73e416a4a6a0f28051f3c5d4e52166b55f79b9d2b55448db5f60d1de54c583c7JmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly93d3cuY2FtYnJpZGdlLm9yZy9jb3JlL2Jvb2tzL3ByZWRpY3RpdmUtbW9kZWxpbmctYXBwbGljYXRpb25zLWluLWFjdHVhcmlhbC1zY2llbmNlLzgyQTdFNUI5NTQ1QjFCRUREN0Y2QTc0OTEyNjUxMTNB

Groll, A., Wasserfuhr, C., & Zeldin, L. (2022). *Churn modelling of life insurance policies via statistical and machine learning methods–Analysis of important features.* https://www.bing.com/ck/a?!&&p=21d58050ec0ab4f3c85dbd5d150e282e081154b9de634ad88e86b94ca9e5bcbaJmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly9hcnhpdi5vcmcvYWJzLzIyMDIuMDkxODI

Guido, A. (2016). *Introduction to machine learning with python.* https://www.bing.com/ck/a?!&&p=a8ab348edcc3e6675bf947b850be35e4b98e4b1036d58350eb1824701b6546aaJmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-

2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly9maWxlcy5hZGRpY3Rib29rcy5jb20vd3AtY29udGVudC91cGxvYWRzLzIwMjQvMDcvSW50cm9kdWN0aW9uLXRvLU1hY2hpbmUtTGVhcm5pbmctd2l0aC1QeXRob24ucGRm

Haar, L., K., Anding, K., Trambitckii, & Notni, G. (2019). *Comparison between supervised and unsupervised feature selection methods*. https://www.bing.com/ck/a?!&&p=25daa73bac27e33ea3a35803a5bb3da9c32feaf4fc3f500e85646fb69205dc9cJmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly9wZGZzLnNlbWFudGljc2Nob2xhci5vcmcvOWFhOS8wMmEyNWMyZWM3NzUxOWUwMWY5ZGQyMDA3Njg3Zjg1YTNhYzYucGRm

Haiss, P., & Sümegi, K. (2008). The relationship between insurance and economic growth in Europe: a theoretical and empirical analysis. *Empirica*, *35*(4), 405–431. https://doi.org/10.1007/s10663-008-9075-2

Hernandez, J. (2019). A data-driven method for individual life insurance policy lapse rate prediction. *Journal of Risk and Insurance, 86*(4), 967-992.

Hsieh, A.-T., & Chen, W.-H. (2025). Examining the impact of service quality and hedonic value on customer satisfaction and retention in online auction platforms. *Journal of Retailing and Consumer Services, 72,* 103377. https://Doi.Org/10.1016/j.Jretconser.2024.103377.

Huang, H., & Sun, J. (2017). The dynamics of life insurance policy lapse behavior. *North American Actuarial Journal*, 461–478.

Johnson, M. E., Luo, Y., Shi, B., & Bai, S. (2019). Policyholder surrender behavior analysis using machine learning: A case study of the Chinese insurance market. *Journal of Risk and Insurance, 86*(1), 133-164.

Journal of Physics, & Ying, X. (2019). *An overview of overfitting and its solutions*. 1–7. https://www.bing.com/ck/a?!&&p=414072d141049c01e93a57b9538833fbba60322e530ec3a6f8786bdbb8deab66JmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly9pb3BzY2llbmNlLmlvcC5vcmcvYXJ0aWNsZS8xMC4xMDg4LzE3NDItNjU5Ni8xMTY4LzIvMDIyMDIy

Kgare, M. (2019). *Predicting lapse rate in life insurance using machine learning algorithms*. https://www.bing.com/ck/a?!&&p=5f34a8a503d16e8740b29fce70e8cbb27004218ecee688e1900fc5d58a522bc9JmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly9pci51bmlzYS5hYy56YS9oYW5kbGUvMTA1MDAvMjg5OTY

Kim, J., & Kim, S. (2020). A machine learning approach for identifying voluntary lapses in long-term care insurance policies. *North American Actuarial Journal, 24*(3), 381-397.

Kim, M., & Hwang, K. B. (2022). An empirical evaluation of sampling methods for the classification of imbalanced data. *PLoS One*, *17*(7).

Kuhn, M., & K. Johnson. (2013). Applied predictive modeling. *Springer.* 3–8.

Lee, J. M. (2021). *Machine learning applications in actuarial science.* https://www.bing.com/ck/a?!&&p=0a22851f0620003930cf6f3bed5dd477695e63f99584b24f1190ef6d0ecbfb4eJmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly93d3cucmVzZWFyY2hnYXRlLm5ldC9wdWJsaWNhdGlvbi8zODkxMzIwNjRfTWFjaGluZV9MZWFybmluZ19pbl9BY3R1YXJpYWxfU2NpZW5jZV9FbmhhbmNpbmdfUHJlZGljdGl2ZV9Nb2RlbHNfZm9yX0luc3VyYW5jZV9SaXNrX01hbmFnZW1lbnQ

Li, H., & Chen, W. (2021). A machine learning approach to modeling insurance policyholder surrender behavior. Insurance. *Mathematics and Economics*. https://www.bing.com/ck/a?!&&p=8237456f4889368572b24d7bb243328638997b9b8688d2cae3380466bd3a068bJmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly9kbC5hY20ub3JnL2RvaS8xMC4xMDE2L2ouZXN3YS4yMDIxLjExNjI2MQ

Lim, S., Oh, T., & Ngayo, G. (2023). Analyzing factors affecting risk aversion: Case of life insurance data in Korea. *Heliyon*, *9*(10).

Loisel, S., Piette, P., & Tsai, C.-H. J. (2021). Applying economic measures to lapse risk management with machine learning approaches. *ASTIN Bulletin*, *51*(3), 839–871. https://doi.org/10.1017/asb.2021.10

Low, Z. W., Rana, M. E., & Hameed, V. A. (2025). Cloud-based deep learning: Technologies, challenges and impact on modern data science. *International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)*, 1–9.

Maier, M., Carlotto, H., Sanchez, F., Balogun, S. & Merritt, S. (2019). Transforming Underwriting in the life insurance industry. *Proceedings of the AAAI Conference on Artificial Intelligence*, 9373–9380.

Malali, N. (2023). Predictive AI for identifying lapse risk in life insurance policies. *Using Machine Learning to Foresee and Mitigate Policyholder Attrition.* https://www.bing.com/ck/a?!&&p=ce7f445a9812a224bf9f30bf0fbcd422ffb4d8122a94d47a35499aafc61a7fb5JmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly93d3cucmVzZWFyY2hnYXRlLm5ldC9wcm9maWxlL05paGFyLU1hbGFsaS9wdWJsaWNhdGlvbi8zOTA4NzQ5MjdfUHJlZGljdGl2ZV9BSV9mb3JfSWRlbnRpZnlpbmdfTGFwc2VfUmlza19pbl9MaWZlX0luc3VyYW5jZV9Qb2xpY2llc19Vc2luZ19NYWNoaW5lX0xlYXJuaW5nX3RvX0ZvcmVzZWVfYW5kX01pdGlnYXRlX1BvbGljeWhvbGRlcl9BdHRyaXRpb24vbGlua3MvNjgwMTFiZDZiZDNmMTkzMGRkNWZhZjE1L1ByZWRpY3RpdmUtQUktZm9yLUlkZW50aWZ5aW5nLUxhcHNlLVJpc2staW4tTGlmZS1JbnN1cmFuY2UtUG9saWNpZXMtVXNpbmctTWFjaGluZS1MZWFybmluZy10by1Gb3Jlc2VlLWFuZC1NaXRpZ2F0ZS1Qb2xpY3lob2xkZXItQXR0cml0aW9uLnBkZg

Mariyam, S. (2025). The urgency of the principle of interest in insurance agreements. *PAMALI: Pattimura Magister Law Review*, *5*(2), 230-243.

Masiuk, I. (2025). *The role of insurance in investment protection and developing financial markets.* https://www.bing.com/ck/a?!&&p=2c948b72b18cbea9bc771464dccd32b69212cb3c50fbd199ce4fc501a88b36c0JmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly9kc3BhY2UuZHNhdS5kcC51YS9oYW5kbGUvMTIzNDU2Nzg5LzEyMjMy

Maynard, T., Bordon, A., Berry, J. B., Baxter, D. B., Skertic, W., Gotch, B. T., & Patel, P. A. (2019). *What Role for AI in insurance pricing*. https://www.bing.com/ck/a?!&&p=51d00e78b07a9fc53413a529576aa864259bed158a2dd573e2a9ba555ce8a3d0JmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly93d3cucmVzZWFyY2hnYXRlLm5ldC9wcm9maWxlL1RyZXZvci1NYXluYXJkLTMvcHVibGljYXRpb24vMzM3MTEwODkyX1dIQVRfUk9MRV9GT1JfQUlfU5fSU5TVVJBTkNFX1BSSUNJTkdfQV9QUkVQUklOVC9saW5rcy82MzcxZjYzNjQzMWIxZjUzMDA5OTY3NGQvV0hBVC1ST0xFLUZPUi1BSS1JTi1JTlNVUkFOQ0UtUFJJQ0lORy1BLVBSRVBSSU5ULnBkZg

Ndungi, R. (2023). *Machine learning algorithm using classification and regression trees, linear gaussian naive bayes, support vector machines, cart and K-nearest neighbors to aid cancer research firms*

in data grouping and better analysis. https://www.bing.com/ck/a?!&&p=5d6e7a492c3ada6019a73ab11cf53e0a1a661f97f357cd7597fb4df375749233JmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly93d3cucmVzZWFyY2hnYXRlLm5ldC9wdWJsaWNhdGlvbi8zNjcyNzY3NTdfTWFjaGluZV9MZWFybmluZ19BbGdvcml0aG1fVXNpbmdfQ2xhc3NpZmljYXRpb25fYW5kX1JlZ3Jlc3Npb25fVHJlZXNfTGluZWFyX0dhdXNzaWFuX05haXZlX0JheWVzX1N1cHBvcnRfVmVjdG9yX01hY2hpbmVzX2NhcnRfYW5kX3stbmVhcmNlbnRyb2lkX2NsYXNzaWZpZXI5SZXNlYXJjaF9Qb3J0dyb3VwaW5nX2FuZF9EYXRhX0dyb3VwaW5nX2FuZF9wbl9EYXRhX0dyb3VwaW5nX2FuZF9wbl9SZXNlYXJjaF8Qb3J0dyb3VwaW5nX2FuZF9wbl9EYXRhX0dyb3VwaW5nX2FuZF9wbl9SZXNlYXJjaF8wZGY

Norazian, M. (2013). Roles of imputation methods for filling the missing values: A review. *Advances in Environmental Biology* (pp. 3861–3869). https://www.bing.com/ck/a?!&&p=06f1f8e3099286a231f114e46499afa780f401a0a42ff13edcf9329742cfbe0bJmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly93d3cucmVzZWFyY2hnYXRlLm5ldC9wcm9maWxlL05vcmF6aWFuLU1vaGFtZWQtTm9vci9wdWJsaWNhdGlvbi8yNTk3NzI5NzVfUm9sZXNfb2ZfSW1wdXRhdGlvbl9NZXRob2RzX2Zvcl9GaWxsaW5nX3RoZV9NaXNzaW5nX1ZhbHVlc19BX1Jldmlldy5sZXMtb2YtSW1wdXRhdGlvbi1NZXRob2RzLWZvci1GaWxsaW5nLXRoZS1NaXNzaW5nLVZhbHVlcy1BLVJldmlldy5wZGY

Omar, M. (2022). Malware anomaly detection using local outlier factor technique. *Innovative Deep Learning Solutions.* Springer International Publishing.

Peddamukkula, P. K. (2024). The impact of AI-driven automated underwriting on the life insurance industry. *International Journal of Computer Technology and Electronics Communication*, *7*(5), 9437–9446.

Pega, F., Náfrádi, B., Momen, N. C., Ujita, Y., Streicher, K. N., Prüss-Üstün, A. M., Descatha, A., Driscoll, T., Fischer, F. M., Godderis, L., Kiiver, H. M., Li, J., Magnusson Hanson, L. L., Rugulies, R., Sørensen, K., & Woodruff, T. J. (2021). Global, regional, and national burdens of ischemic heart disease and stroke attributable to exposure to long working hours for 194 countries, 2000–2016: A systematic analysis from the WHO/ilo joint estimates of the work-related burden of disease and injury. *Environment International*, *154*, 106595. https://doi.org/10.1016/j.envint.2021.106595

Poufinas, T., & Michaelide, G. (2018). Determinants of life insurance policy surrenders. *Modern Economy, 9(8)*, 1400–1422.

Rana, R., & Gupta, S. (202 C.E.). Machine learning algorithms for life insurance surrender prediction: A comparison. *International Journal of Computational Intelligence Systems,* 870–881.

Scriney, M., Nie, D., & Roantree, M. (2020). *Predicting customer churn for insurance data.* https://doi.org/10.1007/978-3-030-59065-9_21

Shamsuddin, S. N., Ismail, N., & Roslan, N. F. (2022). *What we know about research on life insurance lapse: A bibliometric analysis. Risks, 10(5), 97.*

Shumaly, S., Neysaryan, P., & Guo, Y. (2020). Handling class imbalance in customer churn prediction in the telecom sector using sampling techniques, bagging and boosting trees. *International Conference on Computer and Knowledge Engineering (ICCKE)*, 082–087.

Singhal, R., & R. Rana. (2015). Chi-square test and its application in hypothesis testing. *Journal of the Practice of Cardiovascular Sciences*. https://www.bing.com/ck/a?!&&p=485c9323782c45aaf10820961d6bcd0962afb6e4bd751b704dc0dae563773932JmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly93d3cucmVzZWFyY2hnYXRlLm5ldC9wdWJsaWNhdGlvbi8yNz

c5MzU5MDBfQ2hpLXNxdWFyZV90ZXN0X2FuZF9pdHNfYXBwbGljYXRpb25faW5faHlwb3RoZXNpc190ZXN0aW5n

Smith, A. (2023). The role of data and machine learning in insurance. *British Actuarial Journal*. https://www.bing.com/ck/a?!&&p=7220120d1fbefec78ee1fec3edbe404efaadc09b376afa010b4c2a34d674c0d2JmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly93d3cuc2NpZW5jZWRpcmVjdC5jb20vc2NpZW5jZS9hcnRpY2xlL3BpaS9TMjQwNTkxODgyMzAwMDE4MQ

Statinfer. (2019). *SVM: Advantages, disadvantages and applications*. https://www.bing.com/ck/a?!&&p=c7477a8f2aa64da35ace311e43237676f750e8dbcf9649e0f5ca6f69924386c8JmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly9yb2JvdGljc2Jpei5jb20vcHJvcy1hbmQtY29ucy1vZi1zdXBwb3J0LXZlY3Rvci1tYWNoaW5lLXN2bS8

Towiwat, M. S., & Swierczek, F. W. (2024). *Assessing potential AI solutions in the traditional insurance brokerage firm* (Doctoral dissertation, Thammasat University). https://www.bing.com/ck/a?!&&p=d573258f46f39acff740fd6bd7444e56ee0e17c81e1c3645431d84ab07ab530fJmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cDovL2V0aGVzaXNhcmNoaXZlLmxpYnJhcnkudHUuYWMudGgvdGhlc2lzLzIwMjQvVFVfMjAyNF82NjAyMDQzMjIzXzE5Nzk2XzI5NzkyLnBkZg

Tsai, H.-H., T.-W. Yang, W.-M. Wong, & C.-F. Chou. (2022). *A hybrid approach for binary classification of imbalanced data.* https://www.bing.com/ck/a?!&&p=5efaaa2f552c70787b3f9832938125445ed0b69508aee22d6f89c08cf12b057cJmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly9wdWJzLmFpcC5vcmcvYWlwL2FwbC9hcnRpY2xlLzEwNi8yMS8yMTM1MDUvMjc0ODgvQ29tcGxlbWVudGFyeS1yZXNpc3RpdmUtc3dpdGNoaW5nLWJlaGF2aW9yLWluZHVjZWQ

Turner, J. A., & Jenkins, L. (2018). Assessing interest rate risk in the life insurance sector. *Journal of Banking & Finance*, 409–426.

Wade, C., Liebenberg, A., & Blau, B. M. (2016). Information and insurer financial strength ratings: do short sellers anticipate ratings changes? *Journal of Risk and Insurance, 83*(2), 475–500.

Wang, W., Gu, B., Zhang, Z., & Wang, X. (2019). *Research on the prediction of policy lapse rate based on the random forest model.* https://www.bing.com/ck/a?!&&p=2feaa9afff47a51d1a2c0013b15af344398eaf997b58867fe8c011f65c47e2d5JmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly90cmFuc2xhdGlvbmFsLW1lZGljaW5lLmJpb21lZGNlbnRyYWwuY29tL2FydGljbGVzLzEwLjExODYvczEyOTY3LTAyNS0wNzA0NS02

Wang, Y. (2021a). Predictive machine learning for underwriting life and health insurance. *Actuarial Society of South Africa.* https://www.bing.com/ck/a?!&&p=b42751a992ee30e6523faab0e9d89cc8d419863e27ffc755fa09c64fc8bc5630JmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly9hY3R1YXJpYWxzb2NpZXR5Lm9yZy56YS9jb252ZW50aW9uL3dwLWNvbnRlbnQvdXBsb2Fkcy8yMDIxLzEwLzIwMjEtQVNTQS1XYW5nLUZJTi1yZWR1Y2VkLnBkZg

Wang, Y. (2021b). Predictive machine learning for underwriting life and health insurance. *Actuarial Society of South Africa*. https://www.bing.com/ck/a?!&&p=b42751a992ee30e6523faab0e9d89cc8d419863e27ffc755fa09c64fc8bc5630JmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly9hY3R1YXJpYWxzb2NpZXR5Lm9yZy56YS9jb252ZW50aW9uL3dwLWNvbnRlbnQvdXBsb2Fkcy8yMDIxLzEwLzIwMjEtQVNTQS1XYW5nLUZJTi1yZWR1Y2VkLnBkZg

West, D. C., Ford, J. B., & Ibrahim, E. (2020). *Strategic marketing: creating competitive advantage*.

Wuthrich, M. V., & Buser, C. (2019). *Data analytics for non-life insurance pricing.*

Wuthrich, M. V, & C. Buser. (2019). *Data analytics for non-life insurance pricing.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2870308.

Xie, G., Liang., Dong, Z., Tan. B, & B. Zang. (2019). *An improved oversampling algorithm based on the samples' selection strategy for classifying imbalanced data.* https://www.bing.com/ck/a?!&&p=f2fcde5c2cfdf8b4c642929f8ab443dad205b3bde0e109c330f9a6b4486e4375JmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly9vbmxpbmVsaWJyYXJ5LndpbGV5LmNvbS9kb2kvZnVsbC8xMC4xMTU1LzIwMTkvMzUyNjUzOT9tc29ja2lkPTMzMDRiODQ1MmM3ZDY4ZTUyODAyYWJmZjJkYmI2OTA1

Xin, X., & Huang, F. (2024). Antidiscrimination insurance pricing: Regulations, fairness criteria, and models. *North American Actuarial Journal*, *28*(2), 285–319.

Xong, L. J., & Kang, H. M. (2019). A comparison of classification models for life insurance lapse risk. *International Journal of Recent Technology and Engineering, 7(5)*, 245–250.

Yang, Q. (2023). *Robot skill acquisition through prior-conditioned reinforcement learning* (Doctoral dissertation, Örebro University). https://www.bing.com/ck/a?!&&p=983d892e1e81f0b8efbddc439c3ae5b86486da9a1ea41f1f82aa35cfff4864dcJmltdHM9MTc1OTcwODgwMA&ptn=3&ver=2&hsh=4&fclid=3304b845-2c7d-68e5-2802-abff2dbb6905&u=a1aHR0cHM6Ly93d3cuZGl2YS1wb3J0YWwub3JnL3NtYXNoL2dldC9kaXZhMjoxNzk2MzI3L0NPVkVSMDEucGRm

Appendix A: Ethics Clearance Certificate.

NAMIBIA UNIVERSITY OF SCIENCE AND TECHNOLOGY

FACULTY RESEARCH ETHICS COMMITTEE (F-REC)

DECISION/FEEDBACK ON THE RESEARCH PROPOSAL

Dear **Nelson M. Kamati (222143851)**

RESEARCH TOPIC: PREDICTING LAPSE RATE IN LIFE INSURANCE USING MACHINE LEARNING ALGORITHMS (CASE STUDY OF SANALAM NAMIBIA).

Supervisor (if applicable): **Dr Richard Maliwatu**

Qualification registered for (if applicable): Master of Data Science

(Reference number of applications: **FACULTY RESEARCH ETHICS COMMITTEE REGISTRATION NUMBER: FREC - 11/24)**

Re: Ethical screening application No: **FREC - 11/24**

The Faculty of **Computing and Informatics** Ethics Screening Committee of the Namibia University of Science and Technology reviewed your application for the above-mentioned research. The research as set out in the application has been:

Approved	X

(Indicate with an X, and N/A if not applicable and proceed)

We would like to point out that you, as a researcher, are obliged to maintain the ethical integrity of your research, adhere to the ethical guidelines of NUST, and remain within the scope of your research proposal and supporting evidence as submitted to the F-REC. Should any aspect of your research change from the information as presented to the F-REC, which could affect the possibility of harm to any research subject, you are under the obligation to report it immediately to your supervisor or F-REC as applicable in writing. Should there be any uncertainty in this regard, you must consult with the F-REC.

We wish you success with your research and trust that it will make a positive contribution to the questfor knowledge at NUST.

Any ethical issues that need to be highlighted?	Why are these issues important?	What must/could be done to minimize the ethical risk?
No	N/A	N/A

Recommendation: The application is approved.

Sincerely,

Prof. Suama L Hamunyela
Chairperson: Faculty Ethics Screening
CommitteeTel: +264-61-207-2922
CC: Co-supervisor: None

NAMIBIA UNIVERSITY OF SCIENCE AND TECHNOLOGY
Private Bag 13388 Windhoek NAMIBIA
2024 -04- 03
OFFICE OF THE DEAN
FACULTY OF COMPUTING & INFORMATICS

Appendix B: Confirmation Data Collection Letter

Sanlam

Enquires: Mrs. Zaskia Joubert
Tel: +264612947070

08 May 2025

SUBJECT: STUDENT CONFIRMATION ON DATA COLLECTION FOR ACADEMIC RESEARCH PROJECT

This letter serves to confirm that Mr. Nelson Mitchell Kamati, student number 222143851, has successfully collected the required data from Sanlam Namibia for his academic project titled "**Predicting Lapse Rate in Life Insurance Using Machine Learning Algorithms**".

We commend Mr. Kamati for his diligence and professionalism in conducting this research and gathering valuable insights from Sanlam Namibia. His adherence to the ethical guidelines and maintaining of professional conduct during data collection process were exemplary.

In light of his effort, I am happy to support and clarify any data related queries he may have.

Kind regards

Mrs. Zaskia Joubert
Head of Actuarial Department
Sanlam Namibia

Sanlam Centre, 145 Independence Avenue
PO Box 317, Windhoek, Namibia

T +264 (0)61 294 7111
F +264 (0)61 294 7352
E namibia.client@sanlam.com.na

Appendix C: Implementation Details

```
using Flux ✓
using CSV ✓
using DataFrames ✓
using Flux ✓ :onehotbatch
using CategoricalArrays ✓
using StatsBase ✓
using StatsPlots ✓
using FreqTables ✓
using HypothesisTests ✓
using MLJBase ✓
using Distributions ✓
using MLJ ✓
using MLJModels ✓
using MLJDecisionTreeInterface ✓
using Plots ✓
using FeatureSelection ✓
using Statistics ✓
using EvalMetrics ✓
```

```
1 using LIBSVM
```

```
1 using DecisionTree
```

```
data =
```

	PRODUCT-TYPE	EARNING-BASIC-SALARY	PHYSICAL-TOWN	REGION	AGE	GENDER	
1	"RISK"	700	"ONDANGWA"	"Oshana"	60	"MALE"	"WAIT
2	"RISK"	600	"ONDANGWA"	"Oshana"	35	"FEMALE"	"UNEM
3	"RISK"	500	"Okahao"	"Omusati"	37	"FEMALE"	"CASH
4	"RISK"	300	"OSHAKATI"	"Oshana"	50	"FEMALE"	"DESI
5	"RISK"	300	"WALVISBAY"	"Erongo"	35	"FEMALE"	"DOME
6	"RISK"	300	"WALVISBAY"	"Erongo"	28	"FEMALE"	"THER
7	"RETIREMENT"	500	"Grootfontein"	"Otjozondjupa"	37	"MALE"	"CLEA
8	"RISK"	500	"WALVISBAY"	"Erongo"	30	"FEMALE"	"SELF
9	"RISK"	500	"OSHAKATI"	"Oshana"	47	"FEMALE"	"UNEM
10	"RISK"	500	"ONESI"	"Omusati"	33	"FEMALE"	"ACCO
⋮	more						
16975	"RISK"	13001131	"WALVISBAY"	"Erongo"	46	"MALE"	"MANA

```
data =DataFrame(CSV.File("Cleaned_Data_Process_2024_Final30.csv"))
```

```
16975×26 OneHotMatrix(::Vector{UInt32}) with eltype Bool:
 1  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅
 ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅
 ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅
 ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅
 ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅
 ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅
 ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅
 ⋮              ⋮              ⋮              ⋮              ⋮              ⋮
 ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅
 ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅
 ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅
 ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅
 ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅
 ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅
```

```julia
begin

PRODUCTTYPE = onehotbatch(unique(data."PRODUCT-TYPE"),data."PRODUCT-TYPE")

PHYSICALTOWN = onehotbatch(unique(data."PHYSICAL-TOWN"),data."PHYSICAL-TOWN")

REGION = onehotbatch(unique(data."REGION"),data."REGION")

GENDER = onehotbatch(unique(data."GENDER"),data."GENDER")

OCCUPATION = onehotbatch(unique(data."OCCUPATION"),data."OCCUPATION")

EDUCATIONLEVEL = onehotbatch(unique(data."EDUCATIONAL-LEVEL"),data."EDUCATIONAL-
LEVEL")
CONTRACTPAYMENTFREQUENCY = onehotbatch(unique(data."CONTRACT-PAYMENT-FREQUENCY"),
data."CONTRACT-PAYMENT-FREQUENCY")

CONTRACTSTATUSREASON = onehotbatch(unique(data."CONTRACT-STATUS-REASON"),
data."CONTRACT-STATUS-REASON")

PAYMENTMETHOD = onehotbatch(unique(data."PAYMENT-METHOD"),data."PAYMENT-METHOD")
CONTRACTPAYMENTSTATUSREASON = onehotbatch(unique(data."CONTRACT-PAYMENT-STATUS-
REASON"),
data."CONTRACT-PAYMENT-STATUS-REASON")

OCCUPATIONSECTOR = onehotbatch(unique(data."OCCUPATION-SECTOR"),data."OCCUPATION-
SECTOR")

end
```

```
Y = 2×16975 OneHotMatrix(::Vector{UInt32}) with eltype Bool:
 1  ⋅  ⋅  1  ⋅  ⋅  1  1  1  1  ⋅  ⋅  ⋅  ⋅  …  1  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  ⋅  1  1
 ⋅  1  1  ⋅  1  1  ⋅  ⋅  ⋅  ⋅  1  1  1  1     ⋅  1  1  1  1  1  1  1  1  1  1  ⋅  ⋅
```

```julia
Y = onehotbatch(data.LAPSE, unique(data.LAPSE))
```

405 ms

```
combinedEncoded =
16975×1456 Matrix{Int64}:
 1  0  0  0  0  0  0  0  0  0  0  0  0  0  …  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  …  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  1  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0
 ⋮              ⋮              ⋮           ⋱        ⋮              ⋮              ⋮
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  …  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0
```

```julia
combinedEncoded =hcat(PRODUCTTYPE,data."EARNING-BASIC-SALARY",
data."AGE",data."ACTIVE-PREMIUMS-RATE",PHYSICALTOWN,GENDER,OCCUPATION,
EDUCATIONLEVEL,CONTRACTPAYMENTFREQUENCY,CONTRACTSTATUSREASON,PAYMENTMETHOD,
CONTRACTPAYMENTSTATUSREASON,OCCUPATIONSECTOR)
```

```
X = 1456×16975 adjoint(::Matrix{Int64}) with eltype Int64:
 1  0  0  0  0  0  0  0  0  0  0  0  0  0  …  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  1  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  …  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0  0
 ⋮                 ⋮                 ⋮     ⋱        ⋮                 ⋮
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  …  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  …  0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

```
X=combinedEncoded'
```

28.3 μs

```
datasetSize = 16975
```

```
datasetSize=size(X)[2]
```

```
1456×3396 Matrix{Int64}:
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  …  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  …  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0  0
 ⋮                 ⋮                 ⋮     ⋱        ⋮                 ⋮
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  …  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0     0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  …  0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

```
begin
Xtrain=X[:,1:Int(round((datasetSize*0.80)))]
Xtest=X[:,Int(round((datasetSize*0.20))):end]
end
```

```
2×3396 OneHotMatrix(::Vector{UInt32}) with eltype Bool:
 1  1  1  1  1  1  1  1  1  1  1  •  1  1  …  1  •  •  •  •  •  •  •  •  •  •  1  1
 •  •  •  •  •  •  •  •  •  •  •  1  •  •     •  1  1  1  1  1  1  1  1  1  1  •  •
```

```
begin
Ytrain = Y[:,1:Int(round((datasetSize*0.80)))]
Ytest= Y[:,Int(round((datasetSize*0.20))):end]
end
```

```
flux_model = Chain(
                Dense(1456 => 13580),                  # 19_786_060 parameters
                Dense(13580 => 2),                     # 27_162 parameters
                NNlib.softmax,
              )                   # Total: 4 arrays, 19_813_222 parameters, 75.582 MiB.
```

```
flux_model = Chain(Dense(1456 => 13580),Dense(13580 => 2), softmax)
```

1.3 s

```
function flux_loss(flux_model, x, y)
            ŷ = flux_model(x)
            Flux.logitcrossentropy(ŷ, y)
        end;
```

```
0.66650134f0
```

```
flux_loss(flux_model, Xtrain, Ytrain)
```

```
function train_flux_model()
            dLdm, _, _ = gradient(flux_loss, flux_model, Xtrain, Ytrain)
            @. flux_model[1].weight = flux_model[1].weight - 0.1 * dLdm[:layers][1]
[:weight]
            @. flux_model[1].bias = flux_model[1].bias - 0.1 * dLdm[:layers][1][:bias]
        end;
```

```
const classes = ▸["ACTIVE", "LAPSE"]
```

```
const classes=unique(data.LAPSE)
```

```
flux_accuracy(x, y) = mean(Flux.onecold(flux_model(x),classes) .== y);
```

```
for i = 1:500
            train_flux_model();
            flux_accuracy(Xtrain, data.LAPSE[1:13580]) >= 0.98 && break
        end
```

```
@show flux_accuracy(Xtest, data.LAPSE[13580:end]);
```

```
flux_accuracy(Xtest, data.LAPSE[13580:end]) = 0.519434628975265
```

```
Enter cell code...
```

```
model = Ensemble of Decision Trees
        Trees:      10
        Avg Leaves: 7.0
        Avg Depth:  6.0
```

```
model = build_forest(Ytrain, Xtrain', 2, 10, 0.5, 6)
```

12.4 s

```
2
```

```
n_folds=3; n_subfeatures=2
```

```
accuracy = ▸[0.927751, 0.92753, 0.923995]
```

```
accuracy = nfoldCV_forest(Ytrain, Xtrain', n_folds, n_subfeatures)
```

```
Fold 2
Classes:  InlineStrings.String7["ACTIVE", "LAPSE"]
Matrix:   2×2 Matrix{Int64}:
 1330   262
   66  2868

Accuracy: 0.927529827662395
Kappa:    0.836484549100191 8

Fold 3
Classes:  InlineStrings.String7["ACTIVE", "LAPSE"]
Matrix:   2×2 Matrix{Int64}:
 1334   278
   66  2848

Accuracy: 0.9239946973044632
Kappa:    0.8292583426255814

Mean Accuracy: 0.9264250994255413
```

```
model =
  SVM(SVC, RadialBasis::KERNEL = 2, nothing, 1528, 13580, 2, [String7("LAPSE"), String7("AC
```

```
model= svmtrain(Xtrain,Ytrain)
```

```
ŷ, decision_values = svmpredict(model, Xtest);
```

```
# # Compute accuracy
@info "Accuracy: %.2f%%\n" mean(ŷ .== Ytest) * 100
```

```
Accuracy: %.2f%%
mean(ŷ .== Ytest) * 100: 86.77856301531213
```

```
accuracy = nfoldCV_stumps(Ytrain, Xtrain',
                          n_folds,
                          n_iterations;
                          verbose = true)
```

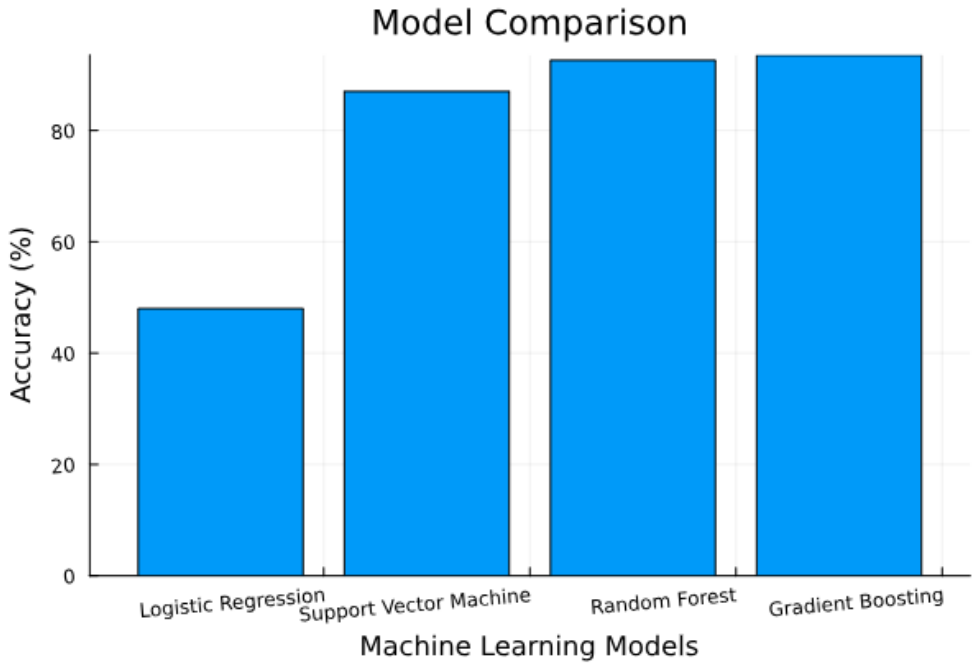
```
Fold 1
Classes:  InlineStrings.String7["ACTIVE", "LAPSE"]
Matrix:   2×2 Matrix{Int64}:
 1362   270
   32  2862

Accuracy: 0.9332744144940345
Kappa:    0.8505463793755726

Fold 2
Classes:  InlineStrings.String7["ACTIVE", "LAPSE"]
Matrix:   2×2 Matrix{Int64}:
 1304   253
   35  2934

Accuracy: 0.936367653557225
Kappa:    0.8541576365151914
```

30.8 s



```
RandomForestRegressor = MLJDecisionTreeInterface.RandomForestRegressor
```

```
RandomForestRegressor = @load RandomForestRegressor pkg=DecisionTree
```

```
For silent loading, specify `verbosity=0`.
```

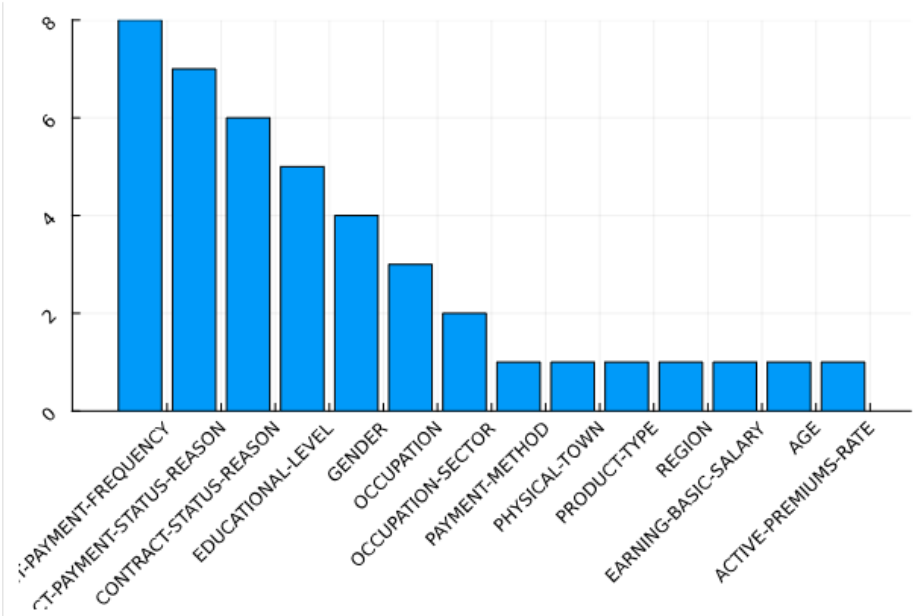
```
import MLJDecisionTreeInterface ✓
```

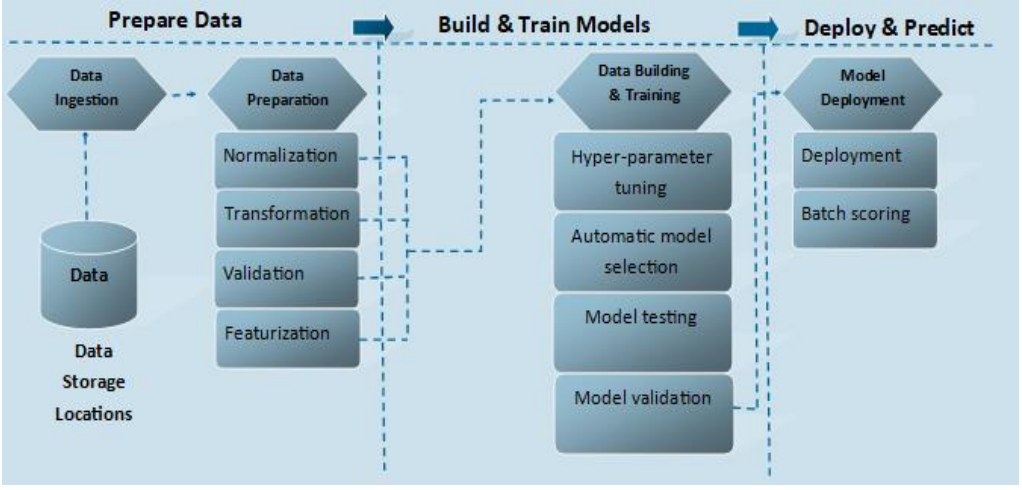
```
rf = RandomForestRegressor(
        max_depth = -1,
        min_samples_leaf = 1,
        min_samples_split = 2,
        min_purity_increase = 0.0,
        n_subfeatures = -1,
        n_trees = 100,
        sampling_fraction = 0.7,
        feature_importance = :impurity,
        rng = Random._GLOBAL_RNG())
```

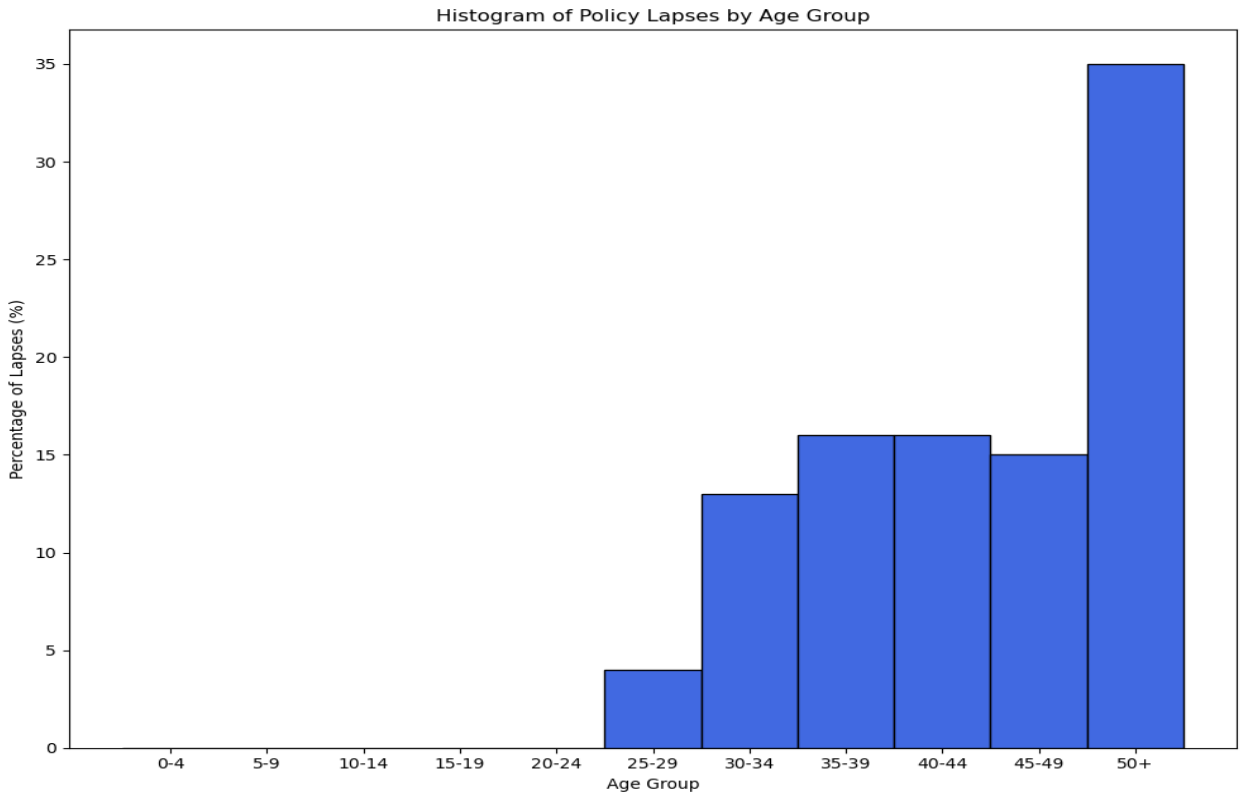
```
rf = RandomForestRegressor()
```

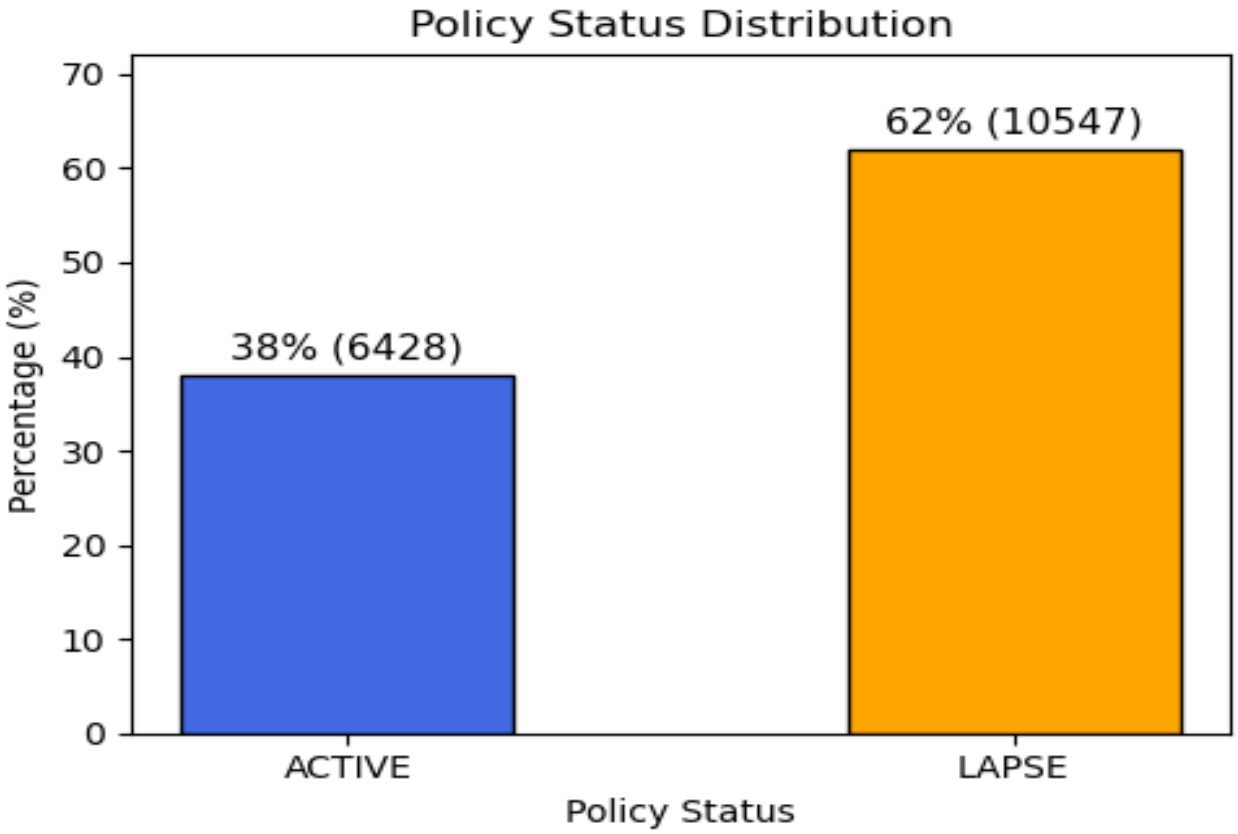
```
selector = DeterministicRecursiveFeatureElimination(
            model = RandomForestRegressor(
                  max_depth = -1,
                  min_samples_leaf = 1,
                  min_samples_split = 2,
                  min_purity_increase = 0.0,
                  n_subfeatures = -1,
                  n_trees = 100,
                  sampling_fraction = 0.7,
                  feature_importance = :impurity,
                  rng = Random._GLOBAL_RNG()),
            n_features = 0.0,
            step = 1.0)
```

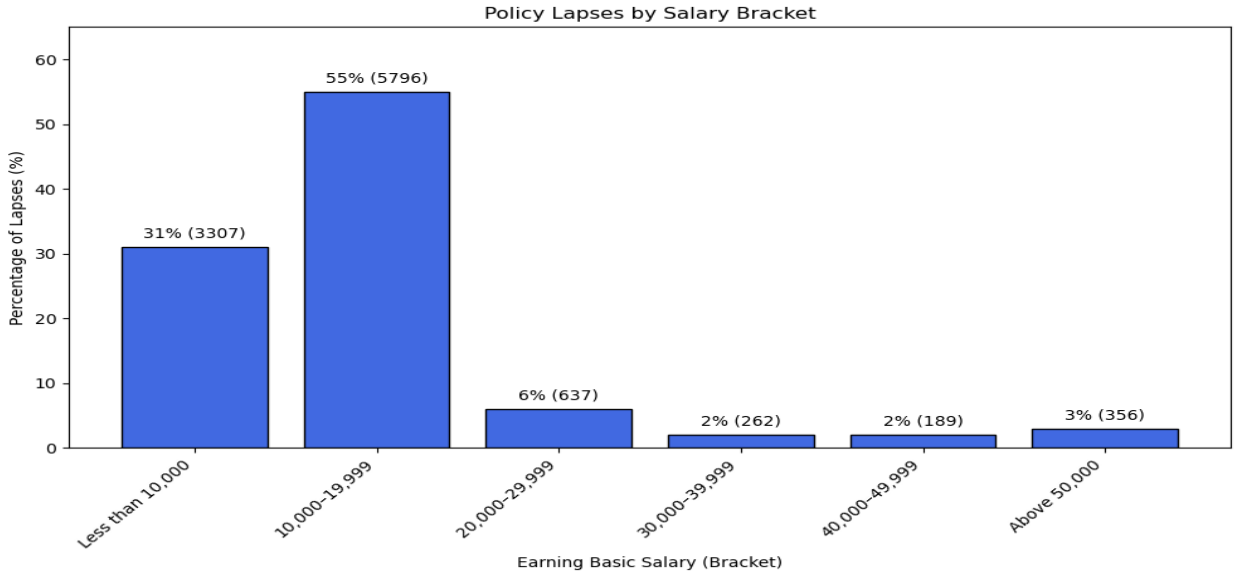
```
selector = RecursiveFeatureElimination(model = rf)
```

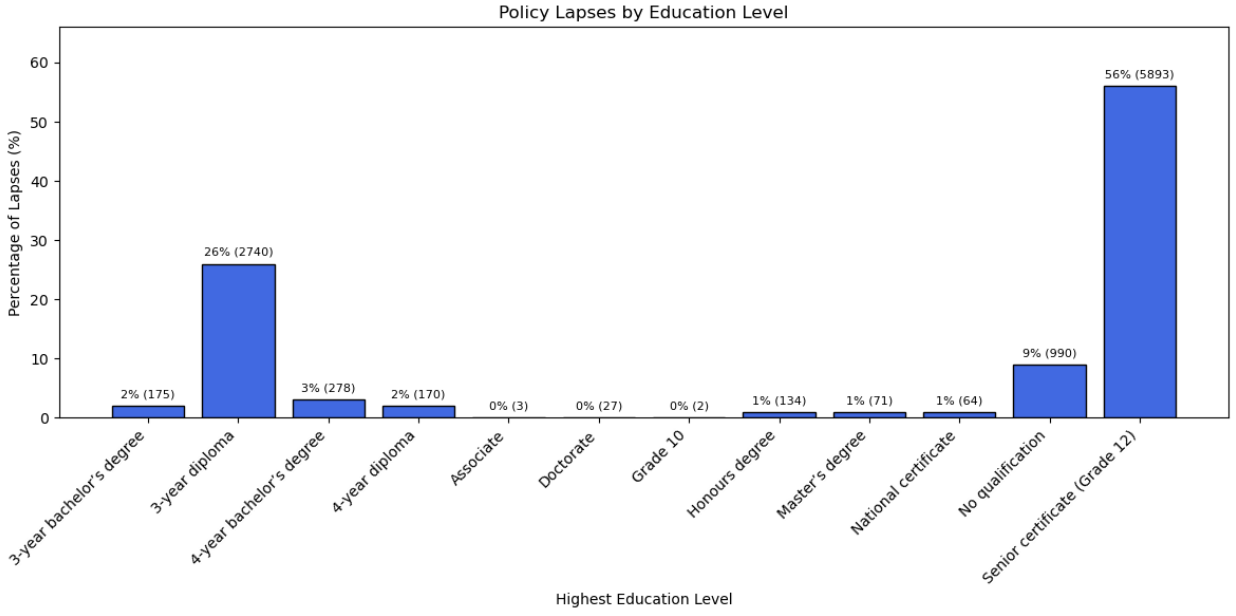


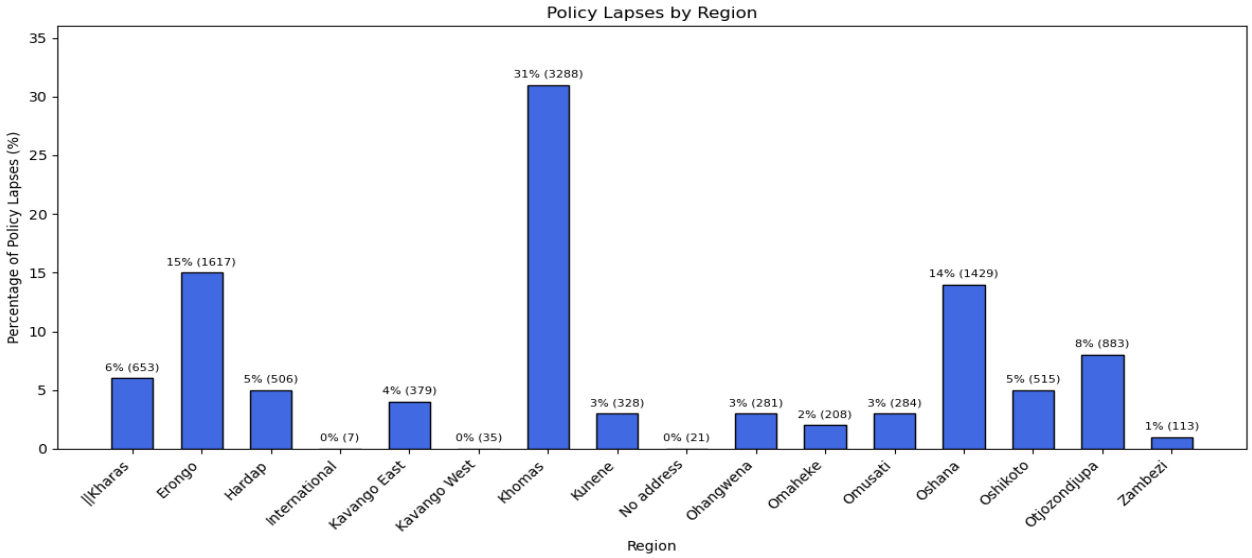


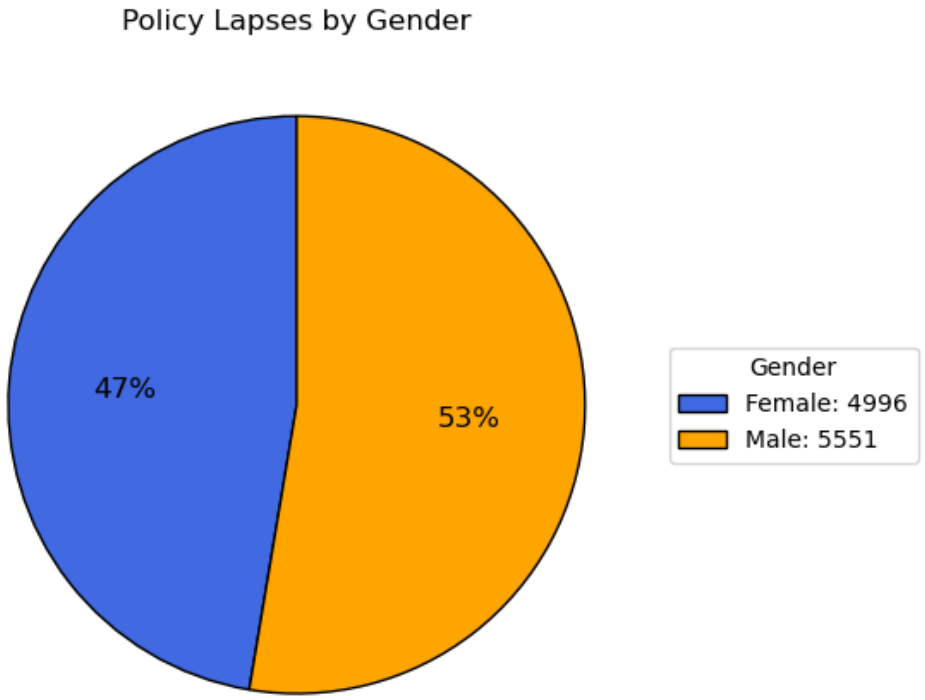

Histogram of Policy Lapses by Age Group
Percentage of Lapses (%)
0
5
10
15
20
25
30
35
0-4
5-9
10-14
15-19
20-24
25-29
30-34
35-39
40-44
45-49
50+
Age Group

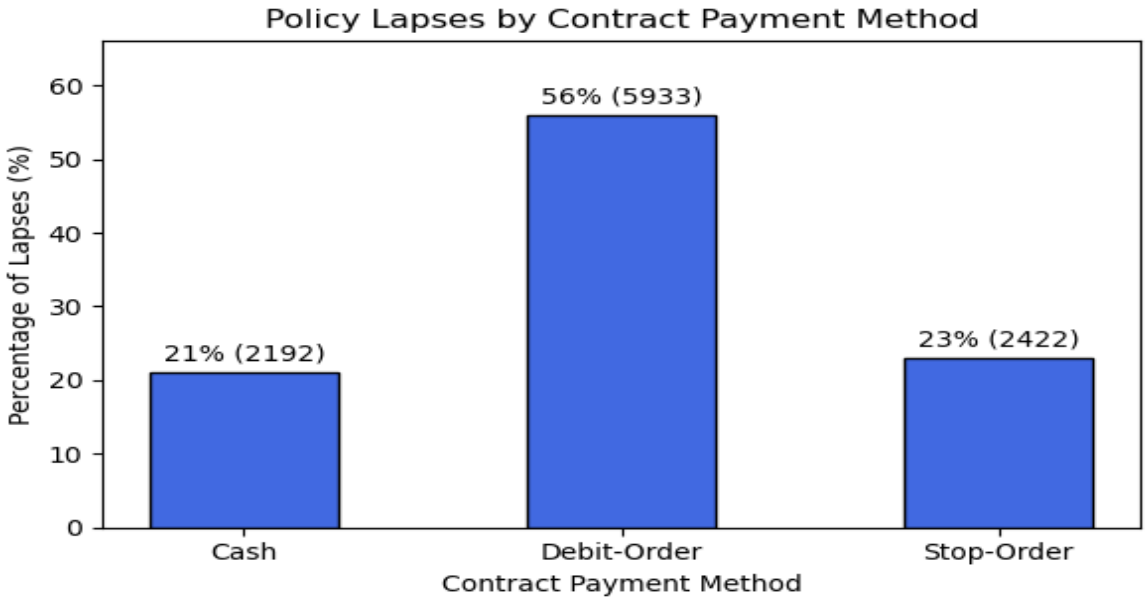

Policy Status Distribution
Percentage (%)
0
10
20
30
40
50
60
70
38% (6428)
62% (10547)
ACTIVE
LAPSE
Policy Status


Policy Lapses by Salary Bracket
Percentage of Lapses (%)
31% (3307)
55% (5796)
6% (637)
2% (262)
2% (189)
3% (356)
Less than 10,000
10,000–19,999
20,000–29,999
30,000–39,999
40,000–49,999
Above 50,000
Earning Basic Salary (Bracket)


Policy Lapses by Education Level
Percentage of Lapses (%)
2% (175)
26% (2740)
3% (278)
2% (170)
0% (3)
0% (27)
0% (2)
1% (134)
1% (71)
1% (64)
9% (990)
56% (5893)
3-year bachelor's degree
3-year diploma
4-year bachelor's degree
4-year diploma
Associate
Doctorate
Grade 10
Honours degree
Master's degree
National certificate
No qualification
Senior certificate (Grade 12)
Highest Education Level


Policy Lapses by Region
Percentage of Policy Lapses (%)
6% (653)
15% (1617)
5% (506)
0% (7)
4% (379)
0% (35)
31% (3288)
3% (328)
0% (21)
3% (281)
2% (208)
3% (284)
14% (1429)
5% (515)
8% (883)
1% (113)
||Kharas
Erongo
Hardap
International
Kavango East
Kavango West
Khomas
Kunene
No address
Ohangwena
Omaheke
Omusati
Oshana
Oshikoto
Otjozondjupa
Zambezi
Region


Policy Lapses by Gender
47%
53%
Gender
Female: 4996
Male: 5551


Policy Lapses by Contract Payment Method
Percentage of Lapses (%)
0
10
20
30
40
50
60
21% (2192)
56% (5933)
23% (2422)
Cash
Debit-Order
Stop-Order
Contract Payment Method