APPLICATION OF SURVIVAL ANALYSIS AND MACHINE LEARNING MODELS TO AGE AT FIRST MARRIAGE AMONG WOMEN IN SOUTH AFRICA.

by

MALAHLANE MAGDALINE KOMANE

DISSERTATION

Submitted in fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

STATISTICS

in the

FACULTY OF SCIENCE AND AGRICULTURE
(School of Mathematical and Computer Sciences)

at the

UNIVERSITY OF LIMPOPO

SUPERVISOR: DR TB. DARIKWA

CO-SUPERVISOR: DR A. BERE (UNIVEN)

OCTOBER 2024

Declaration

I, **Malahlane Magdaline Komane**, affirm that the research entitled: statistical and Machine Learning Models with application to the age at first marriage among women in South Africa has been conducted with the utmost academic integrity. All resources utilised, regardless of origin or authorship, have been meticulously cited and referenced. Moreover, I assert that this research has not been published before and has not been presented to any other institution for review or consideration.

Signature:.....*MMK* Date:...16 October 2024.....

Komane, M.M.

Abstract

Understanding factors influencing the age at first marriage is crucial for addressing social issues, promoting gender equality, and ensuring women's well-being. This research aims to identify key determinants of age at first marriage for South African women. The discrete survival tree approach is applied to analyze data and identify factors influencing women's age at first marriage. Key individual variables, such as birth year, ethnicity, education level, age at first marriage, and province, are used in the analysis. The performance of this model is compared with that of Random Forests and Classification and Regression Trees using the C-index to determine the best-performing model." All three models provided valuable insights, but Random Forest emerged as the most accurate age predictor at first marriage. Key determinants identified were province of residence, birth year, and educational level. These findings can contribute to policy-making aimed at improving the well-being of women in South Africa through targeted interventions.

Keywords: Age at first marriage, South Africa, women, survival analysis, recursive partitioning, Random Forest

Dedication

This research is a small token of my appreciation for the two remarkable women who have shaped my life: my grandmother Maria Komane and my mother Millicent Komane. Their unwavering support, love, and guidance have been the bedrock of my journey, and I dedicate this work to their enduring inspiration and influence.

Acknowledgments

I want to express my immense gratitude to Dr. Darikwa for his expertise and the invaluable guidance and support he extended to me during my research. Dr. Bere, my co-supervisor, deserves special recognition for his expert advice and assistance in completing this work. Your collective encouragement and mentoring have been invaluable in my academic journey.

I am grateful to NEPTTP for funding this research.

Contents

List of Figures

List of Tables

List of Abbreviations

CART = Classification And Regression Trees

DHS = Demographic and Health Survey

EC = Eastern Cape

FS = Free State

GLMs = Generalised Linear Models

GP = Gauteng

KZN = KwaZulu-Natal

LP = Limpopo

MP = Mpumalanga

NC = Northern Cape

NW = North West

RF = Random Forest

SA = South Africa

SAMRC = South African Medical Research Council

StatsSA = Statistics South Africa

WC = Western Cape

Chapter 1

Introduction and background

1.1 Introduction

Survival analysis is a powerful set of tools used to delve into the intricate dynamics of time until specific events transpire (Hosmer Jr et al., 2011). At its core, it excels in handling scenarios involving censoring, a situation where we possess knowledge of an event's time for some subjects but not all (David and Mitchel, 2012). The primary distinguishing feature of survival analysis lies in its central variable, which extends beyond time and encompasses the fundamental question of whether or not the anticipated event has taken place (Collett, 2023).

This analytical framework not only allows us to scrutinise the survival times of distinct groups but also zooms in to examine the survival experiences of individuals within those groups (Aalen et al., 2022). Beyond mere description, survival analysis is a robust method for comparing these survival times, revealing disparities, and deriving valuable insights.

Moreover, survival analysis not only looks at time but also explains how different types of data, like categories or numbers, affect survival (Bewick et al., 2004). By unraveling the intricate interplay between these factors and the likelihood of event occurrence, survival analysis empowers researchers to unearth critical patterns, ultimately enhancing our understanding of various phenomena (Cox, 1984).

Machine learning is a powerful concept for data analysis, helping us understand information from vast digital sources. Ensembling methods are a key part of this, where multiple models work together for better predictions (Kuhn et al., 2013). Similar to survival analysis, which studies events over time, ensembling methods deal with complex data, looking for meaningful patterns (Klein et al., 2003). These two methods complement each other because survival analysis focuses on identifying the timing of key events, while ensembling methods excel at improving prediction accuracy and handling data with missing information. This makes them crucial for research that investigates time-based events and seeks to draw meaningful conclusions from complex datasets (Hastie et al., 2009).

The strength of ensembling methods lies in their ability to reduce error through the collaboration of multiple models, while survival analysis offers a deep understanding of the impact of different variables on the timing of life events (James et al., 2013). Together, these tools empower researchers to uncover crucial patterns, providing a more comprehensive and accurate understanding of data.

In this research, we embark on a journey to explore the timing of significant life events, specifically focusing on the age at which women in South Africa enter their first marriage. This topic is particularly important within South

Africa's unique social, cultural, and legal landscape, where early marriage remains a complex and pressing issue. South Africa is home to diverse cultural traditions and legal frameworks that shape family structures and marriage practices (UNICEF, 2022). Early marriage, in particular, intersects with questions of gender equality, educational attainment, economic empowerment, and health outcomes. By studying the age at first marriage, we can better understand how these societal factors interact and influence the lives of South African women.

This research is timely and socially significant because it addresses the broader issue of early marriage, which can have profound implications for women's health, education, and economic opportunities. Early marriage often disrupts girls' education, limits career prospects, and heightens the risk of domestic violence and poor maternal health outcomes (Shukla et al., 2023). Understanding the age at which South African women marry for the first time can provide valuable insights into the broader socio-economic challenges facing women in the country, particularly in rural and underserved communities. It also helps policymakers, educators, and advocacy groups develop targeted interventions aimed at delaying marriage and improving women's life outcomes.

1.2 Background

When there is information about the time of an event for a subject but not the current iteration time, it leads to a phenomenon known as censoring (Leung et al., 1997). Censoring is a common occurrence in survival analysis, attributed to various factors. One of the primary reasons for censoring is the inability to observe the time to the event due to circumstances like the termination of research before all participants have encountered the event of interest or indi-

viduals leaving the study before the event occurs.

In many instances, it is presumed that the duration until an event occurs is continuous, allowing the event to happen at any time interval. Continuous-time models are extensively utilised in survival analysis as noted by Tutz et al. (2016). However, there are scenarios where the time-to-event is quantified in substantial units like hours, days, months, or years. In these cases, discrete survival models are better suited. For instance, in this study, the duration until the first marriage is calculated in full years, making it a discrete measurement.

Continuous and discrete survival analysis both focus on understanding the probability of an event (failure) occurring over time. In both scenarios, the survival function conveys the likelihood of surviving beyond a particular time point. However, the hazard function plays a key role in continuous analysis, expressing the instantaneous risk of failure at any given moment, provided that the event has not yet taken place. This concept is crucial for studies involving reliability or survival. A common model for continuous analysis is the Cox proportional hazards model, which expresses the hazard function as a function of explanatory variables.

A challenge with parametric models arises when combining all main and interaction effects, particularly with categorical predictors, leading to sparse cells for precise parameter estimation (Ree and Sanmartín, 2009). Moreover, many parametric models impose limitations by restricting associations to specific functional forms like linear relationships (Ree and Sanmartín, 2009). To address these issues, recursive partitioning methods, commonly known as ”tree-based” approaches, are employed.

Breiman’s random forest method, a prominent recursive partitioning technique

(Breiman, 2001), involves creating a collection of classification or regression trees (CART). Each tree divides the feature space into regions with similar response values, providing a solution to these limitations. In this study, a recursive partitioning approach is utilised to identify factors influencing the age at first marriage for South African women.

After applying recursive partitioning methods, we will apply ensemble methods to further improve the accuracy. Ensemble methods refer to machine learning techniques that combine multiple models intending to enhance accuracy and mitigate the risk of overfitting (Hastie et al., 2009). These methods work by building a diverse set of models and combining their predictions to arrive at a final output. Ensemble methods have become increasingly popular in recent years due to their ability to enhance the effectiveness of machine learning for broader benefit.

In the context of this study that uses recursive partitioning to determine the factors of age at first marriage for South African women, ensemble methods can be applied after the initial modeling to further improve the accuracy and generalization of the model. Accuracy is a metric that signifies the ratio of correctly classified instances to the total instances within a dataset. This is because recursive partitioning is a powerful technique for identifying complex interactions and relationships among variables, but it may not capture all the patterns and nuances in the data (James et al., 2013).

Ensemble methods, such as bagging and boosting, can be applied to the initial model to improve the performance (James et al., 2013). Bagging builds diverse models from subsets of data and combines their predictions to improve accuracy. Boosting, on the other hand, involves building a series of models sequentially, with each subsequent model attempting to correct the errors of the

previous one.

By applying ensemble methods after recursive partitioning, we can improve the accuracy and robustness of the model and reduce the risk of overfitting (Seni and Elder, 2010). Ensemble methods can also help to identify important variables that may have been missed by the initial model and provide a more comprehensive understanding of the factors that influence age at first marriage for South African women (Altman and Krzywinski, 2017).

In summary, ensemble methods can be utilised following recursive partitioning to enhance the model's accuracy and generalization. This approach offers a more thorough understanding of the factors that influence the age at first marriage for South African women.

1.3 Problem statement

In South Africa, the legal marriage age for both men and women is 18 years. However, there are some exceptions to this rule (Department of Home Affairs, 2022). For instance, individuals under 18 can marry with the consent of their parents or guardians. Furthermore, the Minister of Home Affairs may authorise a marriage involving a child under 18 if it is deemed to be in the child's best interests (Department of Home Affairs, 2022).

Various factors contribute to the high rate of early marriage in South Africa, which is commonly defined as individuals getting married before reaching the age of 18. The practice involves marrying young individuals under the legal age (Department of Social Development, 2021). Early marriage has severe negative effects that cannot be overlooked. These detrimental consequences encompass a range of outcomes, such as the exploitation of young brides as domestic

slaves by their in-laws, complications during pregnancy, the discontinuation of education, maternal mortality, social isolation, domestic violence, and limited personal autonomy (Mnisi, 2020). Moreover, early marriage has been found to curtail educational opportunities and hinder self-development, impeding the individual's intellectual, psychological, and emotional well-being (Omoeva and Hatch, 2022; John et al., 2019). The imposition of a lifetime of domestic and sexual subservience often results in early and unplanned pregnancies, undermining the empowerment of women and their autonomy (Mwambene, 2018).

In order to address and prevent early marriage effectively, gaining a thorough understanding of the factors affecting the age at which individuals marry for the first time is of the utmost significance. Extensive research has demonstrated that several key determinants influencing early marriage include religion, education, geographic location, place of residence, and the age at which individuals engage in their first sexual intercourse (Elengemoke and Susuman, 2021). Therefore, it is essential to analyse and examine these factors to comprehend their specific impacts and influences on the age at which individuals enter their first marriage.

1.4 Rationale

The age at first marriage for women in South Africa is a critical demographic and social issue that has significant implications for health, education, economic development, and gender equality (Le Roux, 2020). Early marriage is linked to several unfavorable consequences, such as less prospects for higher education, a higher risk of domestic violence, and worse health outcomes for both the mother and the child (Fatima et al., 2023). Therefore, understanding the factors that influence the age at first marriage is crucial for policy development and implementation aimed at addressing these issues. Through the

utilisation of statistical and machine learning models, it is possible to identify the crucial elements that impact the age at which women in South Africa enter their first marriage. These findings can be employed to create predictive models that aid in the development of policies and interventions aimed at postponing marriage and fostering gender equality.

1.5 Aim and objectives

1.5.1 Aim

The aim of the study is to identify the determinants of age at first marriage for South African women using survival trees and Random Forest.

1.5.2 Objectives

The objectives of the study are to:

1. Identify the determinants of age at first marriage using survival trees, and random forest.
2. Compare the models to determine the best-performing model.
3. Analyze the characteristics and impact of the factors identified by the model that influence age at first marriage.

1.6 Significance of the study

This study is significant because it is the first to use a recursive partitioning approach to identify the determinants of age at first marriage in South Africa. Recursive partitioning methods, such as random forests, are well-suited for identifying complex relationships between predictor variables and outcomes. By using a recursive partitioning approach, this study will be able to identify

the most important factors that influence age at first marriage in South Africa, even if these factors interact in complex ways.

1.7 Dissertation structure

The dissertation is organised into five main chapters, each with a specific purpose. Chapter one lays the groundwork, encompassing background information, the research problem, justification, research objectives, and the study's significance. Chapter two then examines the extant literature surrounding the subject. In Chapter 3, we delve into the methodology, focusing on recursive partitioning and survival analysis. We also touch upon their relevance to both discrete-time data and continuous-time survival analysis. Moving on to Chapter 4, we present and interpret the findings of this study. This includes the application of classification and regression trees to the data, as well as an analysis and interpretation of random forest results. We also present and discuss descriptive statistics. Finally, Chapter 5 offers conclusions and recommendations, along with a discussion of potential future research directions.

Chapter 2

Literature review

2.1 Introduction

This chapter delves into the application of recursive partitioning techniques and ensembling methods in the context of early marriage in South Africa. It starts by contextualizing the issue of early marriage in the country and introduces the theoretical foundations underpinning research in this area. The chapter then critically reviews existing literature, emphasizing the gaps and limitations in the literature. Specifically, it explores the lack of research utilizing recursive partitioning techniques and ensembling methods to understand the factors influencing the age at which women in South Africa marry for the first time. This chapter sets the stage for the present study, highlighting the need to employ innovative methodologies to address this research gap.

2.2 Review of literature in RSA

Ayiga and Rampagane (2013) explored how ethnicity shapes age at first marriage in Uganda and South Africa, where early and near-universal marriage prevails. Analyzing national survey data with a Cox proportional hazard model, they found stark contrasts: Uganda's median marriage age was 19, while South Africa's was 29. Ethnicity significantly influenced age in both countries, potentially mediated by factors like region, education, and sexual debut. The authors conclude that Uganda embodies an early marriage regime, while South Africa exhibits delayed nuptiality, likely due in part to differing female education and empowerment.

However, this study lacks a comprehensive exploration of alternative predictive methods such as recursive partitioning or ensemble learning. While Cox proportional hazard models provide valuable insights into individual determinants, they may miss complex, non-linear interactions between factors that more advanced machine learning methods could reveal. By employing these newer approaches, our research can potentially uncover patterns that traditional survival analysis overlooks, thus contributing novel insights into the determinants of marriage timing.

A 2013 study in South Africa conducted by Mpolokeng found women marrying much later, on average above 30. Analyzing survey data with various methods (univariate, bivariate, and Cox regression), the study revealed factors influencing marriage timing. Age, education, ethnicity, sexual debut, and childbirth all played significant roles. Notably, higher education slashed the risk of early marriage by 58%, while wealth and rural childhoods were linked to delayed nuptials. The study suggests promoting family life education across diverse groups to empower informed marriage choices.

While the study by Mpolokeng (2013). provides important insights into the role of education and wealth, the application of machine learning models like Random Forests could help in identifying variable interactions and hierarchical influences that are difficult to detect using regression models alone. Traditional statistical approaches, while powerful, may oversimplify the relationships between variables. Ensemble methods, on the other hand, could highlight more nuanced predictors, such as how different socioeconomic factors interact to influence marriage timing across different ethnic groups.

Mathabatha's 2023 doctoral dissertation titled Determinants of early marriage among women in South Africa: a multilevel analysis addresses the understudied issue of early marriage in the context of South Africa's cultural nuances and provincial variations. Utilizing data from the 2016 Demographic and Health Survey, the study employed a multi-pronged approach, including descriptive statistics, chi-square tests, and multilevel logistic regression. Key findings revealed that individual factors like education level, early sexual debut, and household wealth, alongside community-level elements such as poverty and province, significantly impacted the odds of early marriage (defined as before 18 years old). Notably, women with lower education, poorer households, and residing in specific provinces like Limpopo exhibited higher risks. Recognizing the persistence of early marriage despite existing legal frameworks, Mathabatha recommends tailored interventions, including raising awareness, strengthening child protection laws, and revisiting harmful traditional practices, to empower South African women and curb this concerning social issue (Mathabatha, 2023).

This study provides critical insights into how marriage patterns vary across provinces, with notable differences between urban and rural regions. However, a deeper examination of these regional and provincial differences, using

machine learning approaches, could identify interactions between poverty, education, and cultural norms in more detail. Ensemble methods, in particular, can handle a broader range of predictors and their interactions, making them ideal for studying the multifaceted influences on marriage timing across South Africa's highly diverse provinces (Gabrikova et al., 2023).

2.3 Review of literature elsewhere

Gobena and Berelie (2022) delved into the understudied determinants of marriage timing in Ethiopia, where early marriage prevails. Employing innovative Cox models with mixed effects, they analysed data from the 2016 Demography and Health Survey. This approach revealed not only individual factors like education influencing marriage timing, but also significant regional variations. Their findings pave the way for targeted interventions addressing women's health and marriage-related challenges in Ethiopia.The use of machine learning models could offer a more robust understanding of how these factors contribute to early marriage, providing a more comprehensive predictive framework that extends beyond the limitations of mixed-effect models.

Sah et al. (2010) found early marriage persists in Terai regions. Analyzing diverse datasets, she revealed declining prepubertal marriage, ongoing early unions among specific Terai groups, and near-universal marriage by age 24. Factors like age, location, education, caste/ethnicity, and dowry significantly influenced timing. Notably, Terai groups faced higher risks due to prevalent dowry and distinct cultural norms. Sah argues the "Terai" category masks internal diversity, urging further research to understand specific drivers of early marriage in these communities.

In a bid to address the knowledge gap surrounding early marriage in sub-Saharan Africa, a study by Belachew et al. (2022) analysed data from nine high-fertility countries. Employing a multilevel logistic regression model, they revealed a concerning prevalence of early marriage, averaging 55.11%, with significant variations across nations. Key factors associated with higher risks included lower education levels, unemployment, large family size, community poverty, and rural residence. Highlighting the need for targeted interventions, the study calls for improved access to education, particularly in rural areas, and empowering women to actively participate in their marriage decisions. This comprehensive analysis sheds light on the complex interplay of individual and contextual factors driving early marriage in the region, paving the way for tailored policy initiatives to address this persistent social issue.

Singh et al. (2023) examined trends and determinants of age at first marriage for Indian women (1992-2021) using the first five rounds of NFHS data and advanced statistical methods. Early marriage significantly declined (65.9% in 1992-1993 to 23.2% in 2019-2021), with higher education, wealth, and media exposure reducing the risk. Region, caste, and religion also played a role, with higher marriage ages in southern states, certain castes, and some religious groups. The study concludes that socioeconomic factors and regional disparities influence marriage timing, suggesting policy interventions focusing on girls' education, economic empowerment, and media access for further decline in early marriage and its associated negative consequences.

2.4 Summary of the chapter

In summary, the literature review reveals a multifaceted picture of marriage timing across diverse contexts. While South Africa exhibits later marriage compared to countries like Uganda, significant regional and cultural differences exist within its borders. Studies by Ayiga and Rampagane (2013) andMpolokeng (2013) highlight the influence of factors like education, wealth, and rural upbringing, while Mathabatha's (2023) research delves deeper, revealing how individual and community-level elements like poverty and province impact the odds of early marriage within specific South African provinces.

Beyond South Africa, the issue remains complex. Research in Ethiopia (Gobena and Berelie, 2022) and Nepal (Sah et al., 2010) emphasises the persistence of early marriage in certain regions and communities, influenced by factors like location and cultural norms. Belachew et al. (2022) offer a broader perspective, analyzing data from nine sub-Saharan African countries and revealing a concerning prevalence of early marriage, highlighting the need for interventions focused on education, economic empowerment, and rural development. In India, however, Singh et al. (2023) demonstrate a declining trend in early marriage, attributing it to factors like improved education and media exposure.

However, a research gap exists within South Africa itself. No prior study has employed advanced techniques like Recursive Partitioning Techniques and ensembling methods to explore the factors influencing the age at which women marry. This void underscores the importance of the proposed study, which aims to bridge this gap and contribute valuable insights into this crucial societal issue.

Chapter 3

Methodology

3.1 Introduction

This chapter outlines the methodology employed to achieve the objectives stated in Chapter 1. It delves into the techniques used to analyse data, specifically focusing on Survival analysis method. The section explains the rationale behind choosing specific analytical tools. Recursive partitional approaches, essential for generating classification trees, are elaborated upon. This chapter serves as a crucial framework for understanding the research's validity and autonomy.

3.2 Survival analysis

3.2.1 Introduction

Survival analysis encompasses a range of techniques known by different names. It is commonly referred to as survival analysis in biostatistics, where the focus

is often on analyzing the time until death. In the social sciences, event history data is a prevalent term, and reliability methods are commonly used in technical applications (Jiang and Guterman, 2024). Regardless of the terminology, all these fields aim to model time-to-event data, predicting the duration until a specific event happens.

Survival data is unique because time(time at first marriage) serves as the response variable in regression models, studying how predictors influence a specific survival period T. As $T \geq 0$ must be satisfied, one encounters an input feature with limited support. Nevertheless, applying a transformation like $\log(T)$ makes it possible for all observations to occur, resembling a typical log regression problem: $\log(T) = x^T\gamma + \epsilon$, where x represents predictors, γ is the coefficients vector, and ϵ is the noise variable. However, while such models are useful in simple situations, they prove ineffective in more complex scenarios.

When working with survival data, there are two crucial aspects to consider: understanding the underlying processes often represented as risk or hazard functions, and dealing with censorship, which means that precise timing of events may not always be available (Collett, 2023).

3.2.2 Censoring

Censoring refers to Situations where the exact timing or outcome of an event is unknown. In this study, some participant's marital status was "censored" because the survey ended before they reached an age typically associated with marriage. Since we can't be sure if they were single, married, or even divorced, their data can't be definitively included in certain analyses focused on specific marital status. Therefore, censoring simply indicates incomplete information about the timing or outcome of an event for certain observations (Turkson et al., 2021).

Censorship in data analysis can take two forms: right censoring and left censoring. Left censoring happens when it is certain that the event occurred before the observation period started, yet the exact time of the event remains unknown. On the other hand, right censoring occurs when data is cut off after a specific duration, meaning events that happen after a fixed time are not included in the data (Ranganathan and Pramesh, 2012).

Moving on, we examine a fundamental distinction within survival models: discrete versus continuous approaches. By distinguishing between these two methodologies, researchers can more accurately and insightfully navigate the complexities of time-to-event data.

3.2.3 Discrete versus Continuous Survival models

Discrete and continuous survival models differ in their treatment of time intervals. Discrete survival models deal with data where the event times are measured in discrete intervals or as intrinsically discrete measurements. In discrete survival analysis, event times are recorded as integers representing specific time periods (e.g., days, months, years). Continuous survival models handle data where event times are measured on a continuous scale. In this context, time is considered to be a continuous variable, and events can occur at any point within a time interval.

In discrete time-to-event models, hazards can be expressed as conditional probabilities, making them easier to comprehend compared to continuous hazard functions. Many events are naturally discrete, and their occurrence times are often observed discretely. Therefore, discrete time-to-event models are more precise and suitable for analyzing observable data compared to continuous survival models (Kvamme and Borgan, 2021).

Unlike continuous survival models, which struggle with tied data, discrete survival models effectively handle tied events. Discrete survival models can be formulated as generalised linear models (GLMs), allowing the application of estimation methods designed for GLMs in discrete survival analysis (Fiorentin et al., 2020). This methodology's versatility extends to advanced models such as frailty models, which include subject-specific parameters. Considering that age at first marriage is measured in whole years, the discrete survival approach emerges as the ideal choice due to its ability to directly capture the discrete nature of the event and avoid assumptions about continuous time intervals.

3.2.4 Continuous Survival models

In the context of continuous-time (T), our focus typically revolves around the survival function denoted as:

$$S(t) = P_r\{T \geq t\} = 1 - F(t) \tag{3.1}$$

Here, the survival function ($S(t)$) measures the probability that the event of interest has not yet transpired by time t. In simpler terms, for T representing time until death, $S(t)$ signifies the probability of surviving beyond time t. $F(t)$ represents the cumulative probability of the event happening up to a specific time t.

The hazard function, often denoted as $\lambda(t)$, is a fundamental concept in survival analysis. It defines the instantaneous rate at which events occur, given that the individual has survived up to time t. Mathematically, the hazard function can be expressed as:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} \tag{3.2}$$

where $f(t)$ is the probability density function of T.

Cox's proportional hazard model, introduced by Cox (1984), is widely employed in continuous survival models. In this model, an individual's hazard at time t ($\lambda_i(t|X_i)$) is determined by covariates x_i, including a constant, through the formula:

$$\lambda_i(t|X_i) = \lambda_0(t) \exp\{X_i'\beta\} \tag{3.3}$$

Here, $\lambda_0(t)$ signifies the default hazard function for individuals with $x_i = 0$, while $\exp\{X_i'\beta\}$ represents the relative risk. This term illustrates the proportionate increase or decrease in risk associated with the characteristics x_i.

The Cox proportional hazards model is a semi-parametric technique used in survival analysis to explore the association between the survival time of subjects and one or more predictor variables (covariates). Unlike parametric models that specify the entire distribution of survival times, the Cox model does not assume a specific baseline hazard function, $\lambda_0(t)$. Instead, it focuses on the ratio of hazards for different levels of covariates. This model assumes that the effects of covariates on the hazard function are multiplicative and can vary over time, allowing for flexibility in analyzing survival data (Deo et al., 2021).

The assumptions underlying the Cox proportional hazards model are crucial for its validity. The primary assumption is the proportional hazards assumption, which states that the ratio of hazards for any two individuals is constant over time. This implies that the effects of covariates are independent of time and that the hazard functions can be expressed as a product of a baseline hazard and a function of covariates. Additionally, the model assumes that there is no unmeasured confounding, meaning all relevant covariates must be included in the model to estimate the hazard ratios accurately. Violations of these as-

sumptions can lead to biased estimates and incorrect conclusions, making it essential for researchers to validate these conditions before applying the model (Deo et al., 2021).

3.2.5 Discrete survival models

Discrete survival analysis primarily centers on the hazard function, expressed as:

$$\lambda(t|x) = P(T = t|T \geq t, x), t = 1, ..., q \tag{3.4}$$

In Equation 3.4, the numerator signifies the conditional probability that the event will happen within the interval $[t, t + dt)$ given it has not occurred before, while the denominator represents the width of the interval. The hazard function ($\lambda(t)$) characterises the instantaneous risk of the event occurring at a specific time point, provided it has never occurred before.

Consider time denoted by T. The discrete survival function, denoted as $S(t|X)$, is defined as:

$$S(t|X) = P(T > t|x) = \prod_{i=1}^{t}(1 - \lambda(i|x)) \tag{3.5}$$

This function represents the probability of failure occurring after time t. In practical terms, considering the underlying intervals, it signifies the likelihood of surviving the interval $[a_{t-1}, a_t)$.

In discrete survival modeling, the hazard function $\lambda(t|X)$ is defined as $P(T = t|T \geq t, x)$ for $t = 1, \ldots, q$. In the context of discrete survival models, the hazard function, a central element in time-to-event analysis, is formalised as a mathematical function that delineates the conditional probability of an event's occurrence at a specific time point, given a set of covariates. This equation is

represented as:

$$\lambda(t|X_i) = F(\gamma_0 t + X_i^T \beta) \tag{3.6}$$

In Equation 3.6, F represents the inverse link function, and $\gamma_0 t$ represents the baseline hazard function, capturing the influence of time on the hazard. X_i^T represents the transpose of the vector of covariate values for the individual i and β is a vector of coefficients associated with the covariates.

Estimating the parameters $X_i^T \beta$ is typically done using techniques like maximum likelihood estimation. Interpretation of parameter estimates involves understanding how changes in covariates impact the hazard function. The magnitude and sign of covariate coefficients in a discrete survival model indicate the direction and degree of their association with the hazard function. Positive coefficients signify an increased risk of the event occurring at any given time point, while negative coefficients denote a decreased hazard and hence a lower probability of the event at that specific time (Suresh et al., 2022).

Model evaluation in discrete survival analysis focuses on two key aspects: calibration and discrimination. Calibration assesses how well the predicted probabilities of the event align with the actual observed event rates. Discrimination evaluates the model's ability to differentiate between individuals with high and low risks of the event. Techniques like likelihood ratio tests or information criteria can be used to assess the overall model fit, but calibration and discrimination provide more specific insights into the model's performance (Simino, 2009).

3.3 Recursive partitioning

3.3.1 Introduction

Recursive partitioning divides the data space into rectangular regions, where a simple model with no additional variables is applied to each region. In cases where the outcome is time-to-event, a covariate-free estimate of the survival function is obtained using the Kaplan-Meier estimator. A popular implementation of this method is the CART algorithm, which iteratively splits the data into smaller and smaller regions based on specific criteria (Zhang and Singer, 2010).

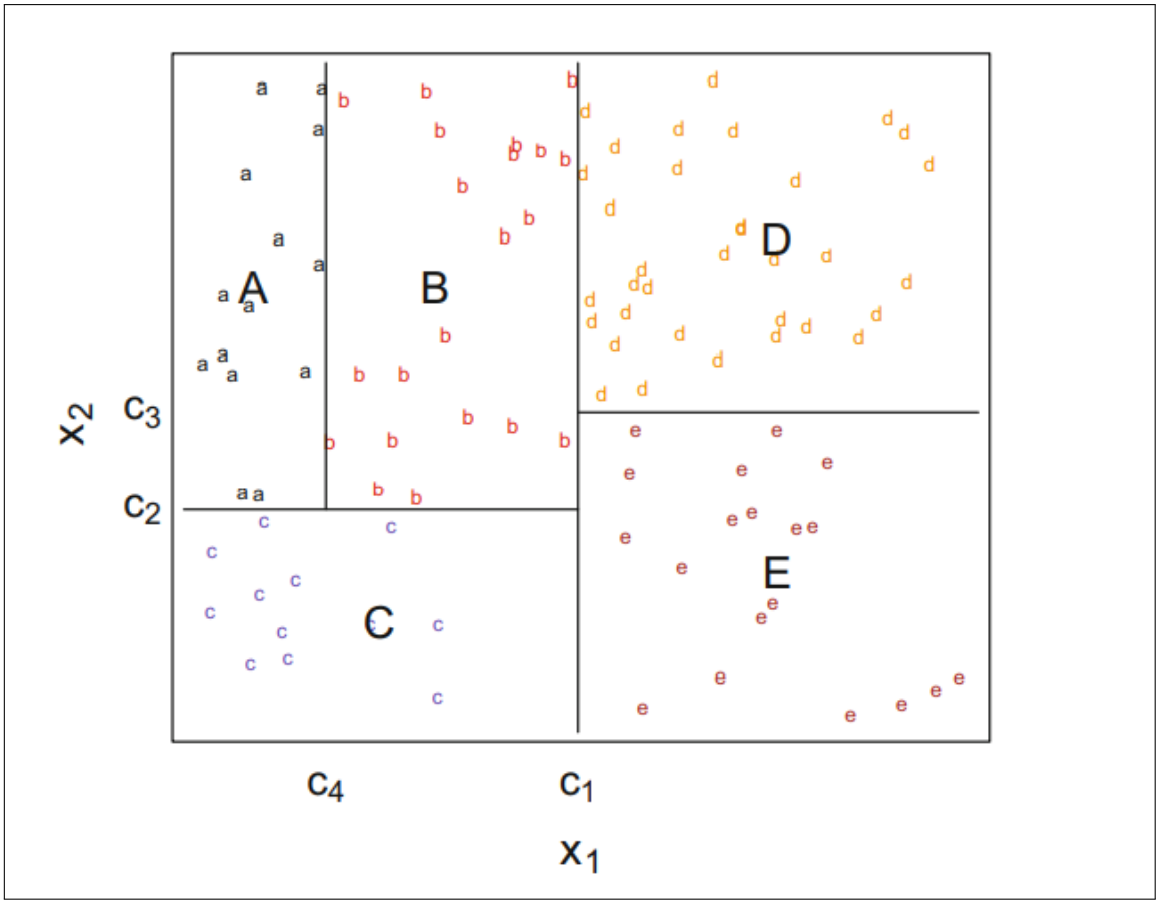

Figure 3.1: A recursively partitioned tree

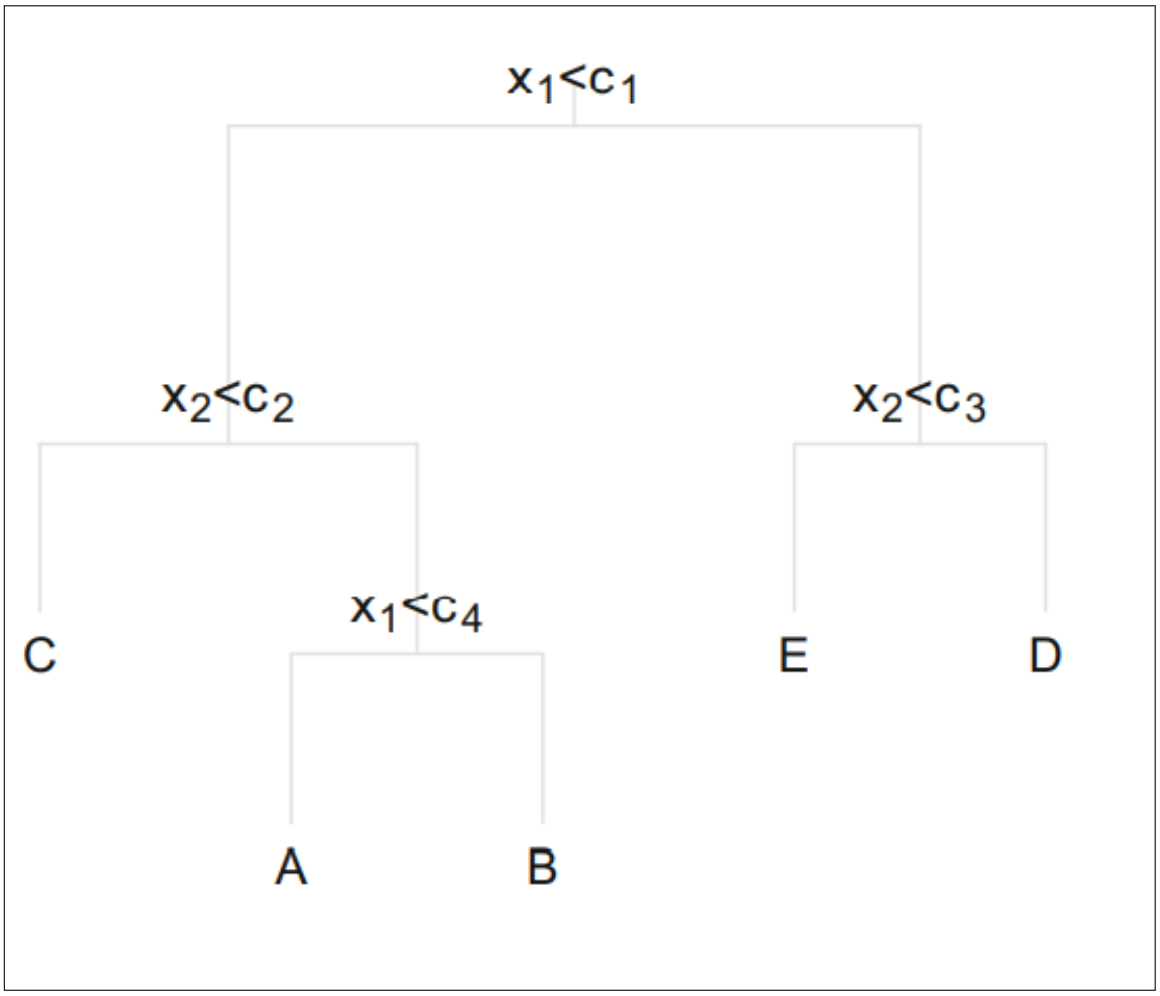

Figure 3.2: The tree generated by the recursive partitioning process

Recursive partitioning is depicted in Figure 3.1. In an example partition of a covariate space illustrated in the upper figure (Figure 3.2), x_1 and x_2 represent two continuous predictor variables. Initially, at threshold c_1, x_1 is used to divide the covariate space into two rectangles in the first phase. The left rectangle is further split into two more rectangles in the second step, based on the conditions $x_2 \leq c_2$ and $x_2 > c_2$. Subsequent splits at thresholds c_3 and c_4 create two additional divisions, resulting in a total of five rectangles covering the entire covariate space (Tutz et al., 2016). The resulting tree, displayed in Figure 3.2, has five terminal nodes, corresponding to the five rectangles.

Within each rectangle or node, the approach involves selecting a single covariate $X_k, k = 1, \ldots, p$ and an arbitrary splitting rule $R(x_k)$ optimised based on a predefined splitting criterion. This criterion is determined by the outcome variable Y. For instance, if Y is a binary variable, a common strategy is to construct a classification tree. The Gini impurity measure is a well-known criterion for splitting binary outcome variable Y. The child nodes are represented as $N_m(x_k, R(x_k)), m = 1, 2$. Let $p_m(x_k, R(x_k))$ denote the proportions of ones in the child nodes. The Gini impurity measure is defined as

$$G_m(x_k, R(x_k)) = 2 \cdot p_m(x_k, R(x_k))(1 - p_m(x_k, R(x_k)))$$

so that it achieves its minimum value of zero in regions where nodes are pure ($p_m \in \{0, 1\}$).

The best split is the one that results in child nodes with the lowest combined impurity, as measured by the weighted Gini coefficients:

$$\min_{x_k, R(x_k)} \{|N_1(x_k, R(x_k))| \cdot G_1(x_k, R(x_k)) + |N_2(x_k, R(x_k))| \cdot G_2(x_k, R(x_k))\}$$

where $|N_1|$ and $|N_2|$ represent the numbers of elements in the sets of observations contained in the child nodes. In addition to classification and regression trees, alternative recursive partitioning methods exist. The key differentiation between these methods lies in the algorithms used to identify the optimal covariate x_k and select the best tree size. Various splitting criteria and techniques can be employed for determining tree size. Besides CART and Quinlan's C4.5, two other popular approaches are discussed in detail elsewhere (Hothorn et al., 2006; Quinlan and Cameron-Jones, 1993). Recursive partitioning techniques have been extensively reviewed in the literature (e.g., Strobl et al. (2009); Hastie et al. (2009); Tutz et al. (2016)). In the following section, we present a specialised method developed for the analysis of discrete survival

times.

3.3.2 Recursive partitioning for discrete survival methods

This method involves representing survival time as binary data, indicating whether the event occurred or not at specific time points, and then fitting a survival tree to this transformed dataset. The approach is credited to Schmid (Schmid et al., 2016). Below is a summary of the method.

[**Input:**] A set of learning data points (t_i, δ_i, x_i), where i ranges from i=1,...,n.
[**First Step**] **(Data expansion):** Generate an augmented dataset by constructing matrices based on the ordinal representations of the binary outcome values y_{it} and time variable T.
[**Second Step**]**(Tree construction):** Employ the CART algorithm on the augmented dataset, incorporating the Gini impurity measure for splitting decisions.
[**Third Step**] **(Tuning):** Implement cardinality pruning, utilizing information criteria or cross-validation techniques, to determine the ideal minimum node size.
[**Fourth step**]**(Conditional survival function estimation):** Predict hazard and survival functions by incorporating new sets of observations at potential time points. Navigate down the generated tree using the resulting vectors.
This method transforms survival time into binary representations, facilitating the application of CART algorithms and enabling the prediction of hazard and survival functions.

This method converts the survival time into a series of binary indicators, recording whether the event of interest happened or not at each specific time point. Then, it uses these indicators to build a survival tree model, as originally pro-

posed by (Schmid et al., 2016).

3.3.3 Ensembling methods

Ensemble methods are a popular strategy to improve the prediction accuracy of recursive partitioning methods, such as decision trees, by reducing the variance of individual estimators. Due to the huge variance that recursive partitioning algorithms' resulting tree estimators typically experience, even little changes in the data can produce completely different trees. As a result, we use ensembling techniques to lower the variance. Bagging and random forests are two widely used ensemble methods.

Bagging

Bagging stands for bootstrap aggregation which was proposed by Breiman et al. (1984) and involves creating multiple bootstrapped samples from the original data set and fitting a separate tree to each of these samples. The resulting predictions from all individual trees are then aggregated to form the final prediction. The more samples generated, the greater the improvement in prediction accuracy. The randomised training process employed in random forests, where each tree learns from a distinct data subset, inherently reduces the susceptibility to overfitting and results in predictions characterised by lower variance compared to those generated by a single, non-diversified model.

In the context of discrete failure time outcomes, bagging can be implemented via two distinct pathways: 1) the direct application of Method 1 (see description below) to bootstrap samples drawn from the original dataset, or 2) the initial augmentation of the data, followed by the application of Method 2 (see description below) to bootstrap samples extracted from the augmented dataset.

Method 1: This involves directly applying bagging to bootstrap samples drawn from the original dataset. Each bootstrap sample is used to fit a model, and the resulting predictions are aggregated to obtain final estimates of hazard and survival functions.

Method 2: This method involves augmenting the original dataset by adding pseudo-observations representing censored failure times. Bootstrap samples are then generated from this augmented dataset, and hazard and survival function estimates are calculated for each sample. The final estimates are obtained by averaging across these models.

In both cases, bagging can improve the accuracy of hazard and survival function estimates by reducing the variance of the individual models. By averaging the estimates from multiple models, bagging can also help to reduce overfitting and improve the generalizability of the model to new data.

In summary bagging can be done in three steps which are:

[**First Step**] Create B copies of the original dataset by randomly sampling data points with replacement. This process is repeated B times, resulting in B ”bootstrap” datasets.
[**Second Step**] For each of the B bootstrap datasets, build a separate survival tree model using the recursive partitioning algorithms described in Sections 3.2.3. If using Method 2, the minimum node size can be optimised based on the entire augmented dataset (not just the individual bootstrap datasets). This optimised minimum node size can then be applied to all B trees.
[**Third Step**] To obtain the final ”bagging estimate,” average the hazard function predictions from all B individual trees.

3.3.4 Random Forest

A Random Forest is a supervised learning algorithm that builds an ensemble of decision trees by selecting a random subset of predictors at each split. The algorithm uses bootstrap sampling to generate multiple samples from the data set and trains a decision tree on each sample. The predictions from these individual trees are then aggregated to make the final prediction, with each tree having an equal say in the output. This approach helps to reduce overfitting and improve the accuracy of the overall prediction. Additionally, Random Forest provides a measure of variable importance, which can be used for feature selection and gaining insights into the relationship between predictors and the response variable.

In summary Random forests can be done in three steps which are:

[**First Step**] Create B copies of the original dataset by randomly sampling data points with replacement. This process is repeated B times, resulting in B ”bootstrap” datasets.
[**Second Step**] For each of the B bootstrap datasets, build a separate survival tree model using recursive partitioning techniques. However, when making split decisions within the tree, only a randomly selected subset of mtry p predictor variables (out of the total p) is considered. mtry stands for ”number of features tried at each split This adds an element of randomness to the model. For Method 2, the minimum node size can be optimised based on the entire augmented dataset, not just the individual bootstrap datasets. This optimised minimum node size can then be applied to all B trees.
[**Third Step**] To obtain the final ”random forest estimate,” average the predicted hazard functions from all B individual trees.

3.4 Model evaluation methods

3.4.1 Measures of discrimination

Discrimination assesses the ability of a survival model to differentiate between individuals with different survival outcomes. In the context of survival analysis, it is about how well the model can distinguish between individuals who experience an event (e.g., death) and those who do not. The Concordance Index (C-index) is a commonly used metric for evaluating discrimination in survival analysis. It quantifies the proportion of subject pairs where the predicted survival times align with the observed survival times. A C-index of 0.5 suggests no discrimination (equivalent to random guessing), whereas a value of 1.0 signifies perfect discrimination.

3.4.2 Measures of Calibration

Calibration assesses how well the predicted probabilities of survival from the model align with the observed survival rates. It evaluates whether the predicted probabilities accurately reflect the true likelihood of survival. A common Method used is the Calibration Plot which is a graphical representation of the predicted probabilities versus the observed probabilities. It helps visualise how well the model's predictions align with the actual outcomes across different predicted probability ranges. Ideally, points on the plot should fall along the 45-degree line, indicating perfect calibration. The Brier score constitutes a widely employed metric for evaluating the calibration of probabilistic forecasts. It calculates the average squared difference between the predicted probabilities and the binary realizations of the event of interest, with lower scores indicating greater accuracy. A lower Brier score indicates better calibration. Calibration-in-the-large and Slope are additional calibration metrics that

assess overall calibration and the slope of the calibration curve, respectively.

3.5 The data

3.5.1 Data source and study area

This research utilises secondary data from the 2016 Demographic and Health Survey (DHS) for South Africa. This data is open source and is readily available on the DHS website. It's an anonymous data which covers the whole of South Africa. The data was collected jointly by South African Medical Research Council (SAMRC) and StatsSA in the year 2016. This study will focus on the dependent variable, "Age at first marriage," and explore its relationship with various independent variables, including educational level, ethnicity, province, birth years, and religion.

3.5.2 Data manipulation

Below we describe the data preprocessing steps undertaken to prepare a dataset for survival analysis. The data originates from the DHS and the file name is ZAIR71FLR.

The analysis utilises the haven Wickham et al. (2023b) library in R to read data from an SPSS file named "ZAIR71FLR.sav". The use.value.labels argument is set to TRUE to ensure value labels are used for categorical variables, and the to.data.frame argument is set to TRUE to convert the data into a data frame named "datF". Exploratory data analysis is performed using the View function to examine the initial structure and content of "datF".

Several data cleaning and transformation steps are applied to the data:
Subsetting: A new data frame "datF" was created by subsetting "ZAIR71FLR",

retaining only relevant variables for the analysis.
Variable Renaming: Variable names in ”datF” were renamed for improved readability and clarity.
New Variable Generation: New variables were constructed based on existing ones to incorporate additional information:
coh: Birth year information was used to categorise individuals into cohorts (e.g., ”1966-1975”, ”1976-1985”). tim: Age was categorised into groups based on specific ranges (e.g., 14-17, 18-20). provin: Province names were converted to abbreviated codes for easier handling. Data Integration: The newly created variables were combined with the original data in ”datF” to create a new, expanded data frame.

Data Reshaping: The data was reshaped from wide to long format using the dataLong function from the discSurv Welchowski et al. (2022) library. This function originates from a user-defined package and creates a new data frame named ”DatFLong”. The reshaped data structure separates subject ID, time point, event indicator (Cen), and outcome variable into distinct columns. The outcome variable ”y” was coded with labels ”censored” and ”dead” for clarity in the survival analysis.

Factor Conversion: Categorical variables in ”DatFLong” were converted into factor variables using the as.factor function. This step is essential for appropriate handling of categorical data in survival models. Examples of variables converted to factors include cohort(coh), province, language, ethnicity, and education level.

The data is split into training and testing sets for model development and evaluation. The set.seed function is used to ensure reproducibility, and a random selection of 70% of the rows from ”DatFLong” was assigned to the training set

(”train data”). The remaining 30% of the data was allocated to the testing set (”test data”).

Create the CART model using the rpart() function from the rpart library. This function builds a classification tree by recursively splitting the data into subsets based on the Gini impurity measure. The model is fitted to the training data with specific parameters (minsplit=2, minbucket=1, cp=0.001). Prune the tree using the prune() function to avoid overfitting. The complexity parameter (cp) is selected to minimise the cross-validation error (xerror). Predict probabilities on the test data using the pruned CART model with the predict() function. Calculate the concordance index (C-index) using rcorr.cens() to evaluate the model's performance in predicting survival outcomes.

The pruning of the tree is based on minimizing the cross-validation error (xerror). During the construction of the CART model, the complexity parameter (cp) controls the trade-off between the tree's complexity and its predictive performance. Higher values of cp will lead to simpler trees by pruning branches that do not provide significant improvement in the model's accuracy. The selection of the optimal cp is based on minimizing the cross-validated error, which helps to avoid overfitting by ensuring that the tree generalizes well to unseen data. By using cross-validation, we aim to find the cp value that leads to the best model performance without overly fitting to the training data (Hastie et al., 2009).

The survival tree was constructed using the rpart() function from the rpart library. The formula specified was Surv(timeInt, y) coh + provin + PlaceType + Language + Ethnicity + EducationLevel, indicating that survival time and event occurrence were modeled as a function of the covariates. The control parameters for the tree construction were set as follows: minsplit = 2, minbucket

= 1, and cp = 0.00034. The fitted survival tree was evaluated by predicting probabilities on the test data using the predict() function. The concordance index (C-index), a measure of the model's discriminatory power, was calculated using the rcorr.cens() function from the Hmisc library.

The random forest model was built using the rfsrc() function from the randomForestSRC library (Start, 2022). The formula specified was Surv(timeInt, y) coh + provin + PlaceType + Language + Ethnicity + EducationLevel, indicating that survival time and event occurrence were modeled as a function of the covariates. The fitted random forest model was evaluated by predicting probabilities on the test data using the predict() function. The out-of-bag (OOB) probabilities were also obtained to assess the model's performance without requiring a separate validation set. The concordance index (C-index), a measure of the model's discriminatory power, was calculated using Harrell's concordance index.

Chapter 4

Results and discussion

4.1 Introduction

In this chapter, the data format is detailed, providing descriptive statistics insights. Women participants were asked to disclose their age at first marriage. Individuals who did not have their age at first marriage recorded at the time of the survey had their data censored at their age during the survey. Birth years were grouped and labeled as the year code (coh) the groups are 1966-1975, 1976-1985, 1986-1995, 1996-2001. The chosen models will incorporate these factors (Ethnicity, Region, EducationLevel, timeInt, Laguages, coh) for analysis, and the subsequent results will be presented and interpreted.

4.2 Descriptive statistics

The results from this section were created before reshaping the data from wide to long format. This was done because descriptive statistics provide insights

into the structure, distribution, and summary characteristics of the dataset. This understanding is essential before any transformation, as it helps in making informed decisions about how to handle the data.

Table 4.1: Summary of Dataset Variables

Variables	**Explanations**
Ethnicity	Ethnic background of female participants.
Region	Province of residence for the participant.
Education Level	Highest level of education attained.
timeInt	Age at first marriage for women categorised into intervals.
Languages	Languages spoken.
coh	Birth years of women segmented into groups.

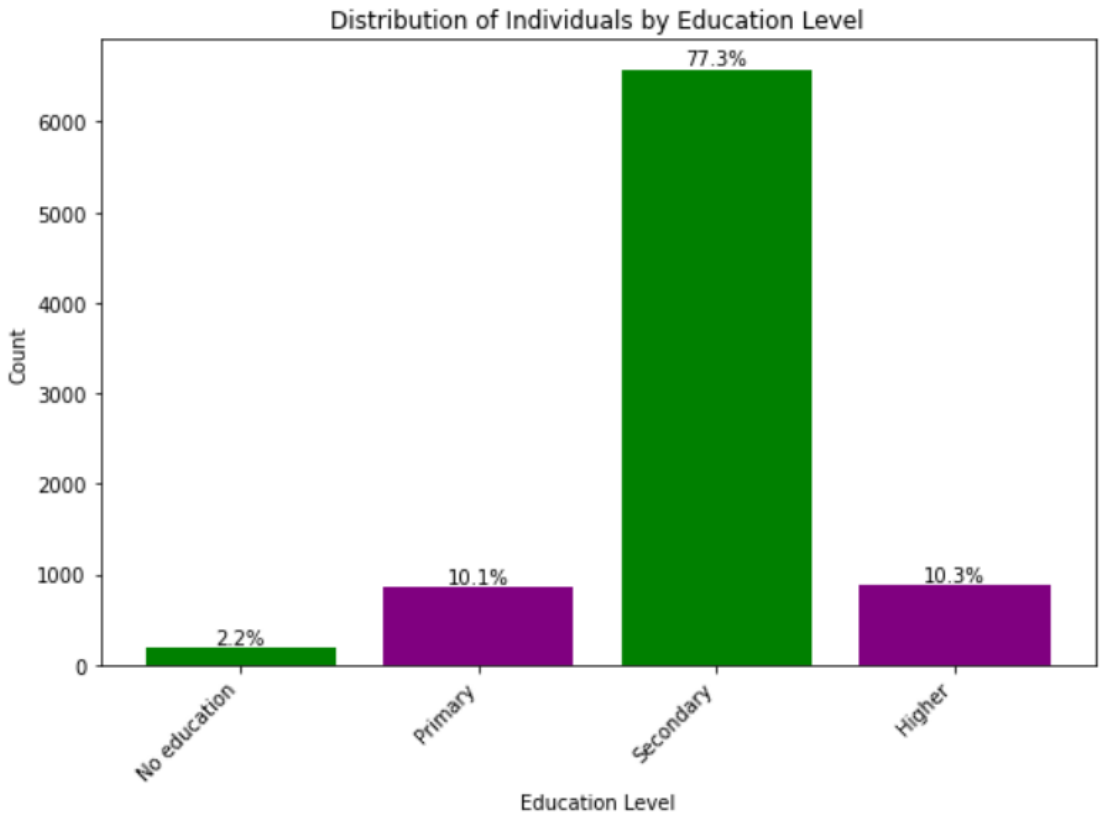

Figure 4.1: Educational level bar graph

Among a group of 8,514 women from Figure 4.1, a remarkable 77.3% have achieved a secondary education. This demonstrates a strong emphasis on liter-

acy and educational attainment within this population. While a small portion may not have formal schooling, the overall data suggests a well-educated and literate group of women.

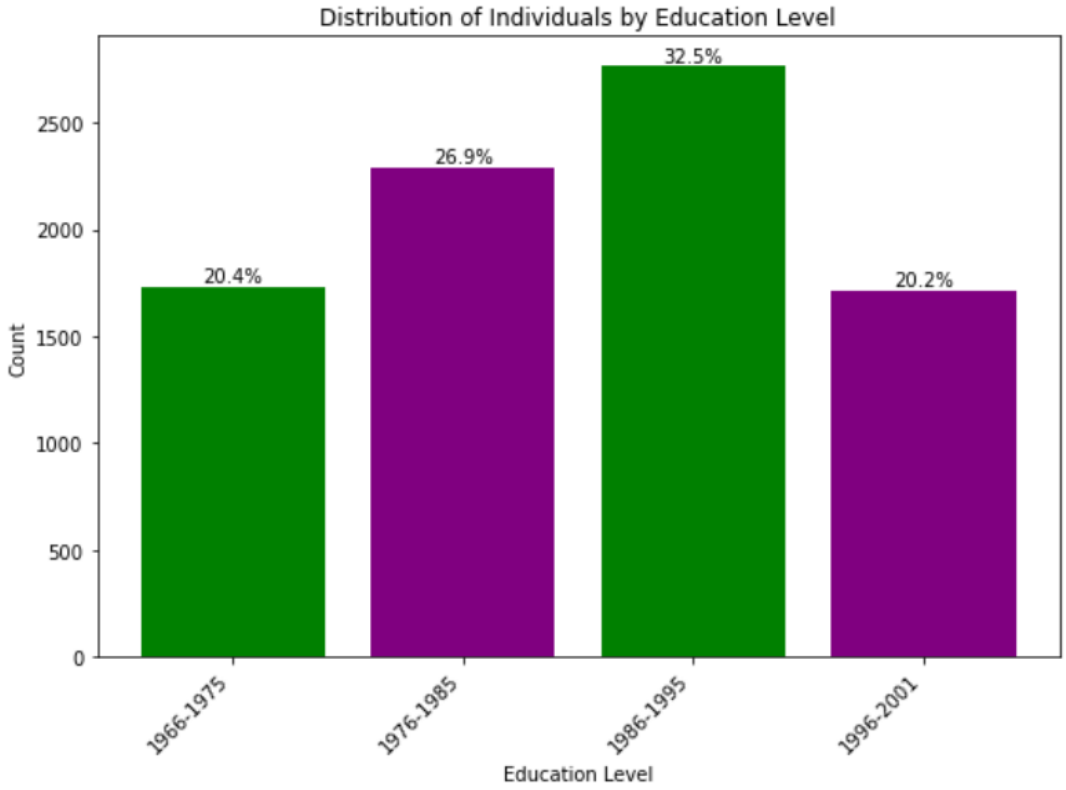

Figure 4.2: Year code bar graph

Figure 4.2 shows a significantly higher number of women (32.5%) born between 1986-1995 compared to other birth years. Notably, the 1996-2001 cohort has the lowest number of women, while the 1966-1975 cohort is slightly higher by 0.2%.

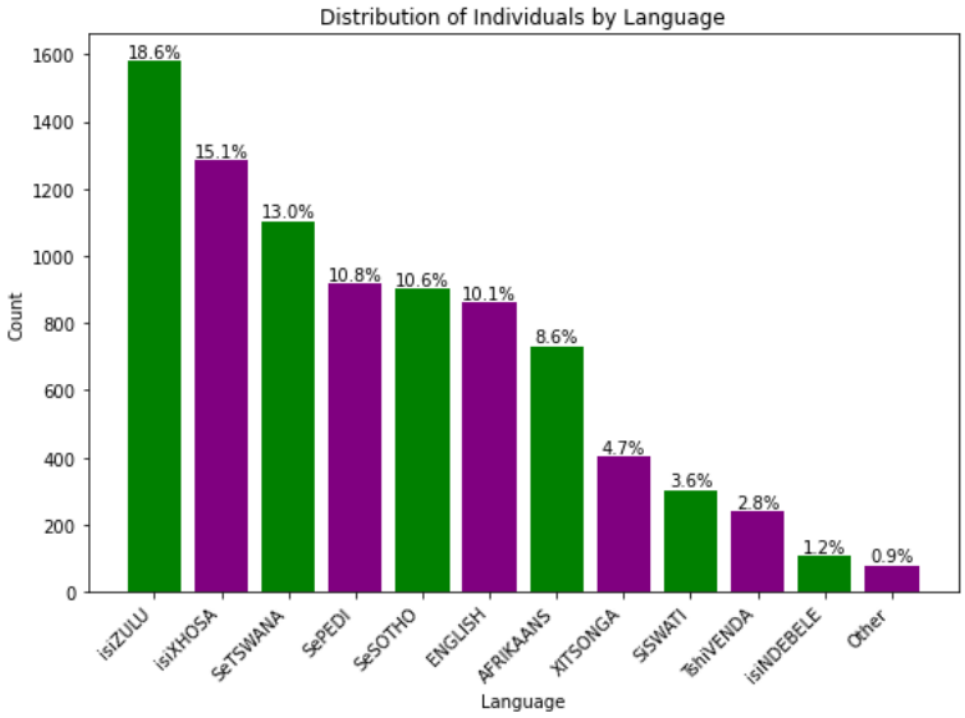

Figure 4.3: Language bar graph

In the study, Zulu was the most spoken language, with 18.6% of participants speaking it. Xhosa came in second, while isiNdebele was the least common language, spoken by only 0.9% of participants. This information is based on Figure 4.3 .

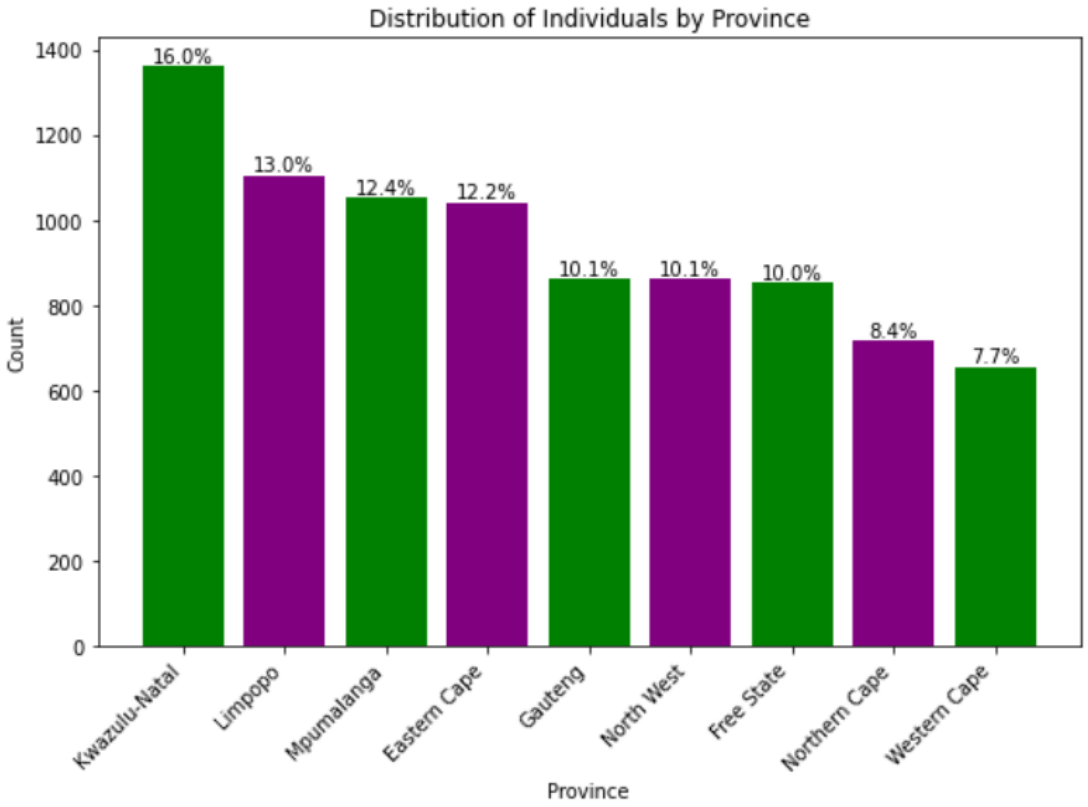

Figure 4.4: Province bar graph

The graph in Figure 4.4 complements the language analysis, confirming that KwaZulu-Natal boasts the highest participation in the study with 16%. Conversely, the Western Cape has the lowest participation rate, with women from this province comprising only 7.7% of the 8,514 participants.

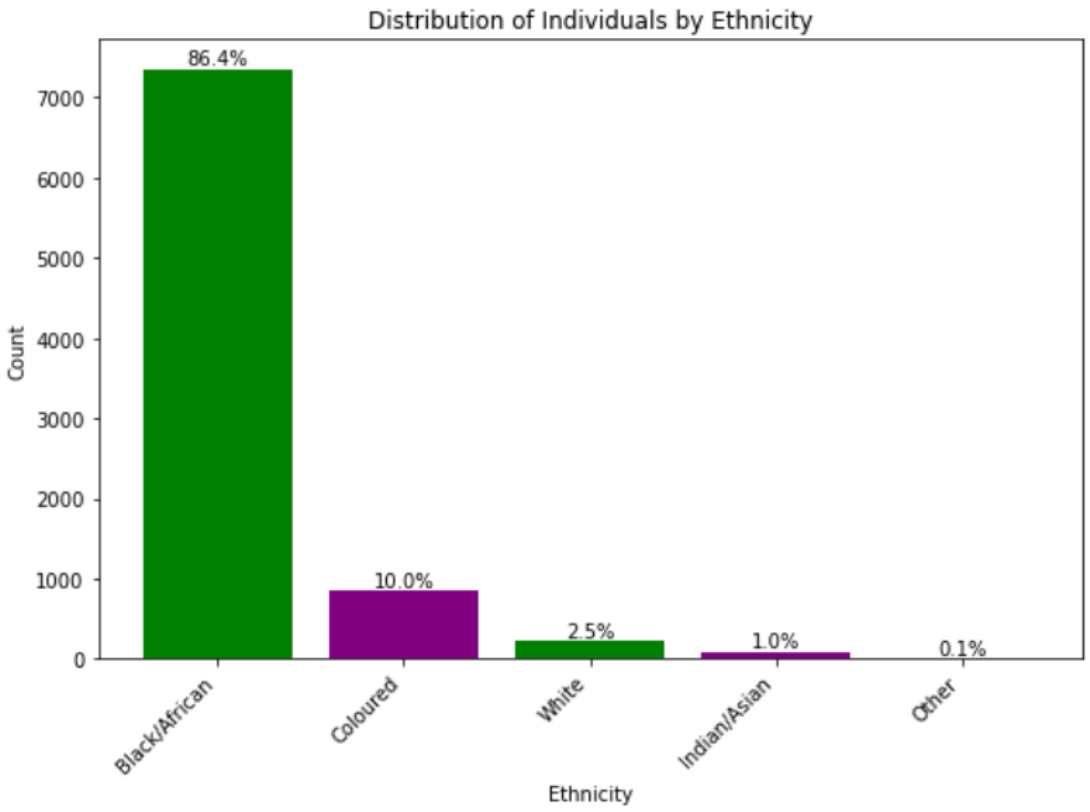

Figure 4.5: Ethnicity bar graph

The majority of participants in the study, representing 86.4%, identified as Black/African. This suggests a strong presence of Black/African women in the study. In contrast, the Indian/Asian demographic had the smallest representation, at only 1% of women. For a detailed breakdown of ethnicity among participants, see Figure 4.5.The section that follows will show and analyse the results of the models chosen.

4.2.1 Tree Models

Classification And Regression Tree

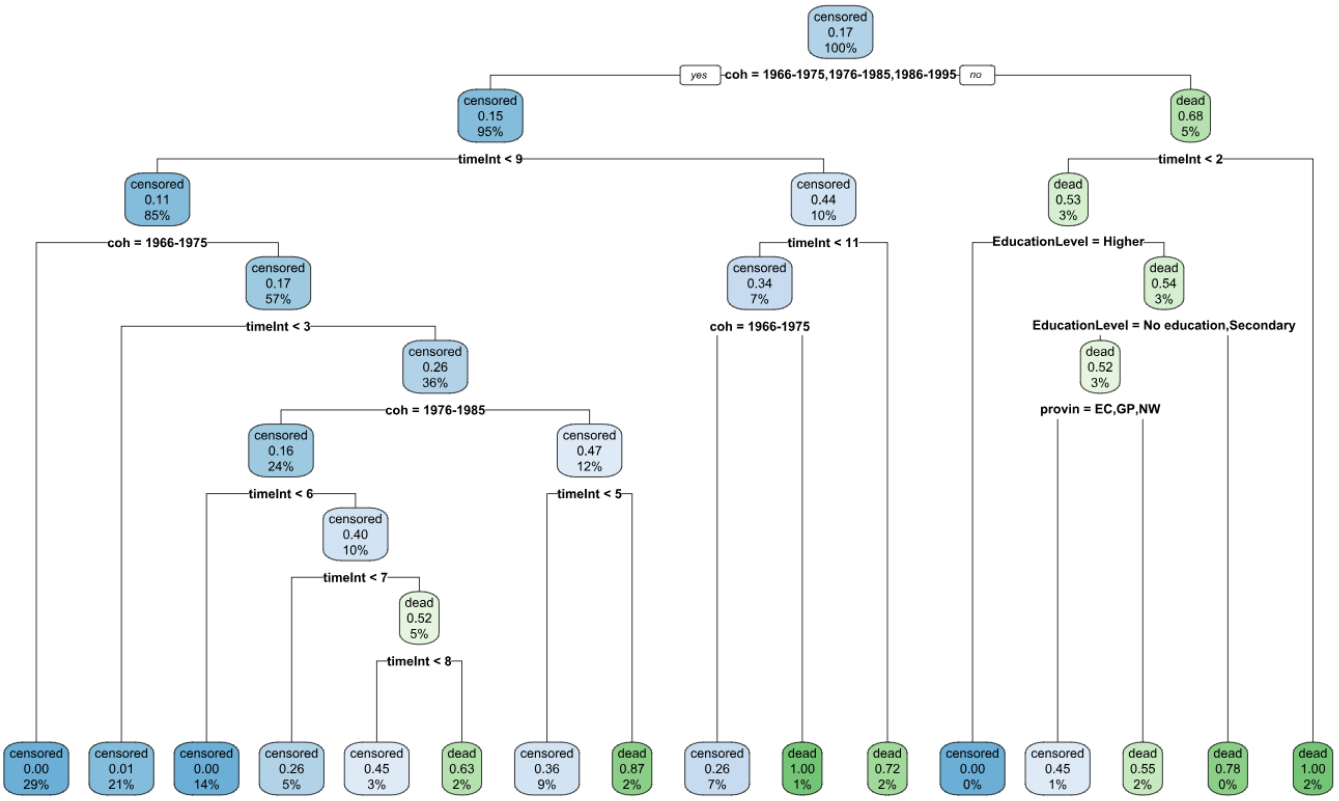

Figure 4.6: In a CART tree, each question is structured so that an affirmative ”yes” answer is positioned on the left, while a negative ”no” answer is positioned on the right.

Censored refers to having a low risk of marriage and dead refers to having a higher risk of getting married.

As seen from Figure 4.6 the first split is based on the cohort year. These cohorts 1966-1975, 1976-1985, and 1986-1995 fall into the ”censored” category, indicating a low risk of getting married. One possible reason for this could be social or economic shifts that occurred during these years. For instance, women in these cohorts may have had increased access to education and career opportunities, delaying or deprioritizing marriage. Furthermore, cultural shifts in attitudes toward marriage, where it may be seen as less of a necessity or a later-life

event, might also contribute to these results (Treas et al., 2014). However, this assumption requires further exploration of other factors like urbanization and changes in marriage legislation over time.

The second split is according to the time interval. For women aged between 36 and 38 (time interval less than 9), the next split is again based on the cohort. Those born between 1966-1975 have a low risk of getting married, while those born between 1976-1985 and 1986-1995 undergo further splitting based on the time interval. Women aged between 18 and 20 (time interval is less than 3) have a low risk of getting married; if they are aged between 21 and 23 (time interval is greater than or equal to 3), another split occurs based on the cohort. Women aged between 27 and 29 (time interval less than 6) have a low risk of getting married, while those aged between 30 and 32 with a time interval of 6 or more are split again. A time interval less than 7 (30-32 years) results in having a low risk of getting married.

These findings indicate that age plays a significant role in the likelihood of marriage, particularly among certain birth cohorts. This could reflect life-stage events like the pursuit of education or professional advancement in younger women, especially those in more recent cohorts. Conversely, women entering their late twenties or early thirties might feel greater societal or familial pressure to marry, possibly explaining the higher risks for marriage as age increases (Carmichael, 2011). However, it is important to consider that while age and time interval provide one dimension of understanding, other influences like personal preferences or the availability of suitable partners are not captured by this analysis.

For individuals with a time interval greater than or equal to 9 (39-41 years), the next split is based on a time interval of 11 (45-47 years). Those with a time

interval less than 11 (42-44 years) are further split by cohort: those born between 1966-1975 have a low risk of getting married, while those born between 1976-1985 have a high risk of getting married.

This later-life split introduces a unique consideration. It could indicate a pattern where older women either delay marriage due to life circumstances or opt for it once they reach a stage of stability (social, economic, or personal). There is an intriguing contrast in that the earlier cohort (1966-1975) has a low risk of marriage compared to their slightly younger counterparts (1976-1985), suggesting that generational values or economic conditions may vary significantly enough to influence marriage decisions (Ariho and Kabagenyi, 2020).

The third split is for the cohort 1996-2001, where the first split is based on the time interval of 2 . If the time interval is less than 2 (15-17 years), there is a further split based on education level. women with higher education have a low risk of getting married, while those with no education, primary, or secondary education undergo further splitting. For those with no education or secondary education, there is a regional split: individuals from EC, GP, and NW have a low risk of getting married, while those from FS, KZN, Limp, MP, NC, and WC have a higher risk of getting married. For individuals with primary education, they have a higher risk of getting married. If the time interval is greater than or equal to 2 (18-20 years), they have a higher risk of getting married.

Educational attainment emerges as a key determinant of marriage risk in this cohort. Higher education may lead to delayed marriage due to a greater focus on career and personal development (Maharaj and Shangase, 2020). Additionally, regional disparities reflect how local customs, economic opportunities, and societal expectations shape marriage decisions. Regions with higher marriage risks may be those where traditional values or limited educational and employ-

ment prospects push individuals toward early marriage.

Survival tree

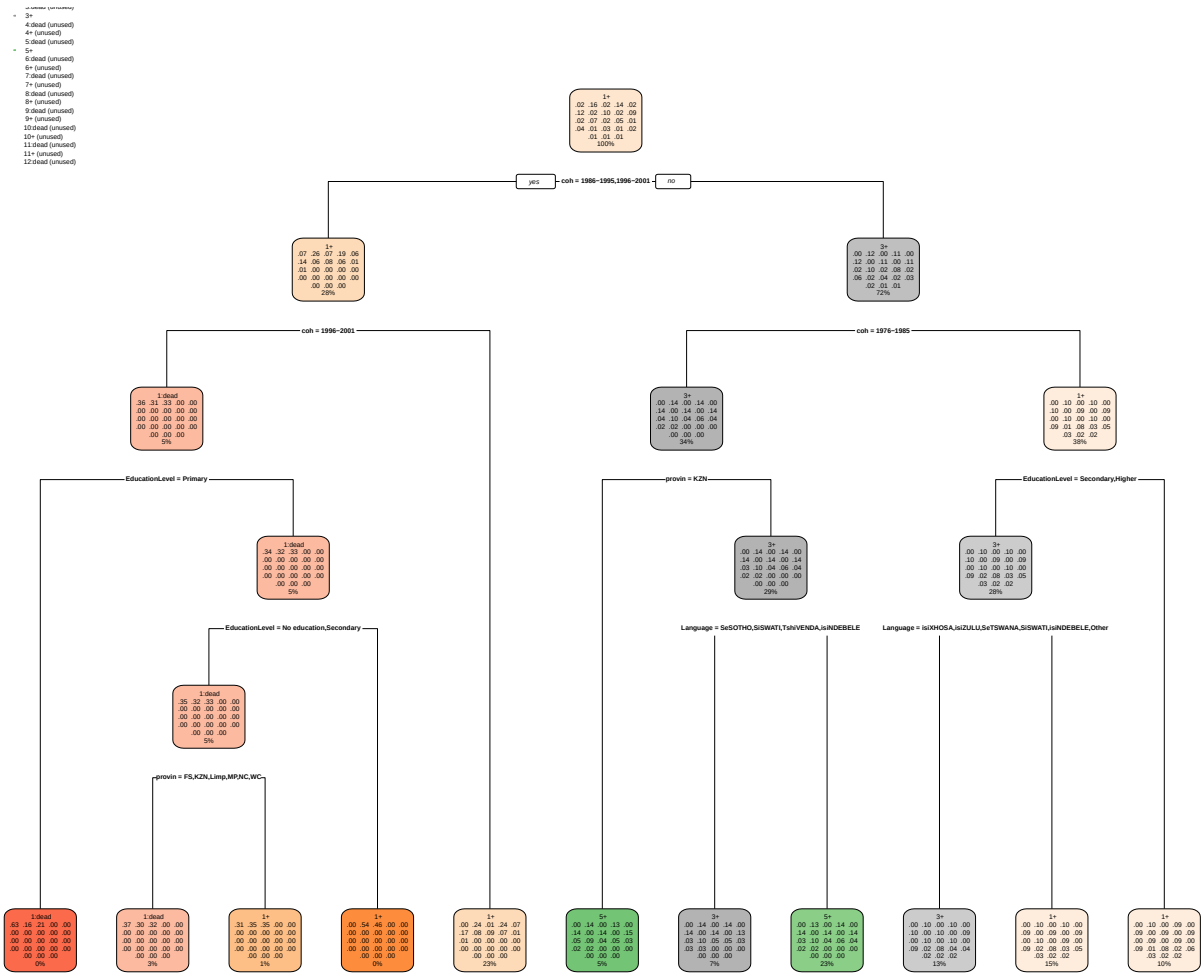

Figure 4.7: Discrete survival tree

As seen from Figure 4.7 Between the cohort years 1996-2001, women with No education or Secondary education level residing in provinces Freestate, KwaZulu-Natal, Limpopo, Mpumalanga, Northwest, or Western Cape had a 3% likelihood of a high risk of getting married. Women with Primary education level had a 61% likelihood of a high risk of getting married. These variations

may be influenced by socio-cultural norms in different provinces, where some regions may place greater emphasis on early marriage as a social expectation, regardless of the woman's education level. Traditional practices and economic conditions could also play a role in pushing women towards early marriage (Treas et al., 2014).

For the cohort years 1966-1975, women speaking English, Sesotho, Sepedi, Tshivenda, Xitsonga, or isiNdebele had a 15% likelihood of a low risk of getting married. Women born between 1986-1995 had a 23% likelihood of a low risk of getting married. The linguistic and cultural backgrounds of these groups may reflect different societal attitudes towards marriage, where some language communities may support or delay marriage due to cultural expectations or shifting norms. Additionally, economic shifts and urbanization could influence marriage decisions over time.

Women born in 1976-1985 with secondary or higher education who can speak isiXhosa, isiZulu, Setswana, siSwati, isiNdebele and other have a 13% likelihood of low risk for marriage. This could be tied to cultural differences, where certain language groups may have traditions or societal pressures that influence the timing of marriage. In some communities, marriage may be more closely tied to social status or family obligations, which could delay or hasten the decision.

Women born in cohort 1966-1975, speaking Afrikaans, isiXhosa, isiZulu, SeTswana, SiSwati, or Other, and residing in GP, Limp, or NW have a 5% chance of low risk for marriage. Women with born in cohort 1976-1985, speaking English, isiZulu, or SiSwati, or Other have a 11% chance of low risk for marriage. Geographic factors may also play a significant role in influencing marriage risk. In provinces such as Gauteng or Limpopo, economic opportunities and urban-

ization may contribute to a delay in marriage, as individuals may prioritize economic stability or career growth before settling down.

Random forest

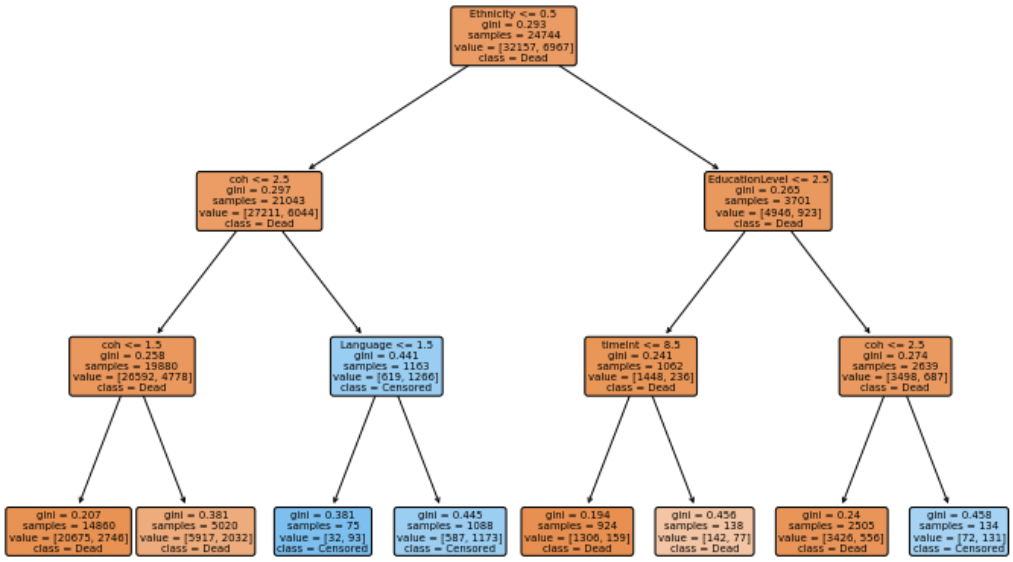

Figure 4.8: A representative Random forest tree

The data presented in Figure 4.8 indicates that women born in the periods 1966-1975 and 1976-1985, irrespective of ethnicity, exhibit a higher risk of entering into marriage and are consequently classified as having a higher risk of getting married. This trend may be influenced by social norms, economic factors, or cultural expectations during those periods, where marriage was often seen as a marker of social stability and success.

Furthermore, the analysis identifies that women born between 1986-1995 or 1996-2001 who speak Isixhosa, Sesotho, Isizulu, or English also fall into the higher risk category and are classified as having a higher risk of getting married. The linguistic or ethnic aspect might reflect cultural or community norms

where marriage is highly valued and encouraged at an early stage in life. Additionally, the role of socioeconomic status associated with specific language groups could potentially play a role in shaping marriage patterns. For instance, in certain communities, marriage might still be linked to economic security, especially where higher education or career opportunities are limited.

Additionally, women who are White, Indian/Asian, or of Other ethnicities, and who possess a Primary or Secondary educational level, are considered at higher risk of getting married. Women from these backgrounds may experience different levels of autonomy in choosing whether or not to delay marriage. This pattern could also reflect traditional expectations within these ethnic groups, where marriage is regarded as a key milestone of adulthood (Maharaj and Shangase, 2020).

Lastly, women with a Higher level of education and birth years falling between 1966-1975 or 1976-1985 are similarly deemed at higher risk and are classified as having a higher risk of getting married.

The below section will compare the c-index for all the models.

4.2.2 Comparison of models

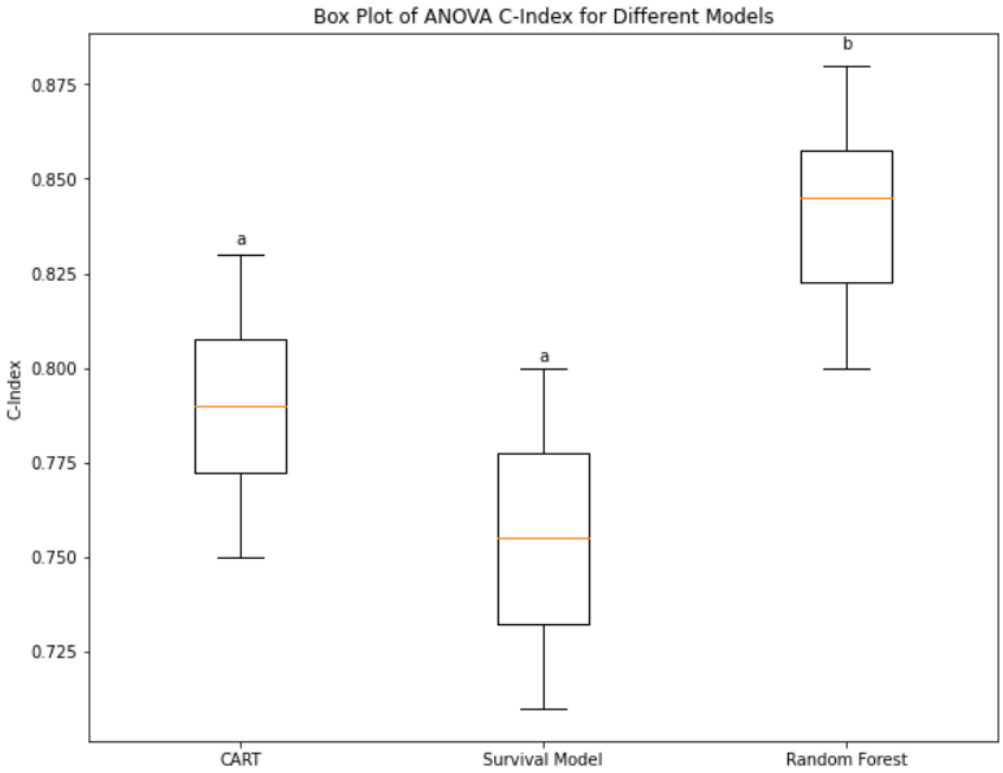

Figure 4.9: Box Plot of ANOVA C-Index for different models

Table 4.2: Model Performance Summary

Model	**C Index**	**Accuracy**
CART	0.79	0.80
Discrete Survival	0.75	0.72
Random Forest	0.84	0.83

Figure 4.9 was created by Repeatedly dividing the data set into training and testing samples and using each pair to get a C-index value for each model.Below is the interpretation of the graph.

Random Forest has the highest mean C-index (0.841), indicating its predictions tend to be closer to the true order of event times on average compared to the other models. CART and Survival Model have lower mean C-indices (around

0.790 and 0.755, respectively), suggesting their predictions might be less accurate in ordering event times.

Random forest also demonstrates the least variability in C-Index values, indicating more consistent performance across different datasets or evaluation runs. CART and survival models exhibit slightly wider ranges of C-Index scores, suggesting their performance might be more sensitive to dataset characteristics or model configurations.

There are no significant outliers in the C-Index values for any of the models, suggesting that the results are generally consistent and reliable. Based on these results, random forest would likely be the preferred choice for this particular task, as it offers the strongest predictive performance and consistency.

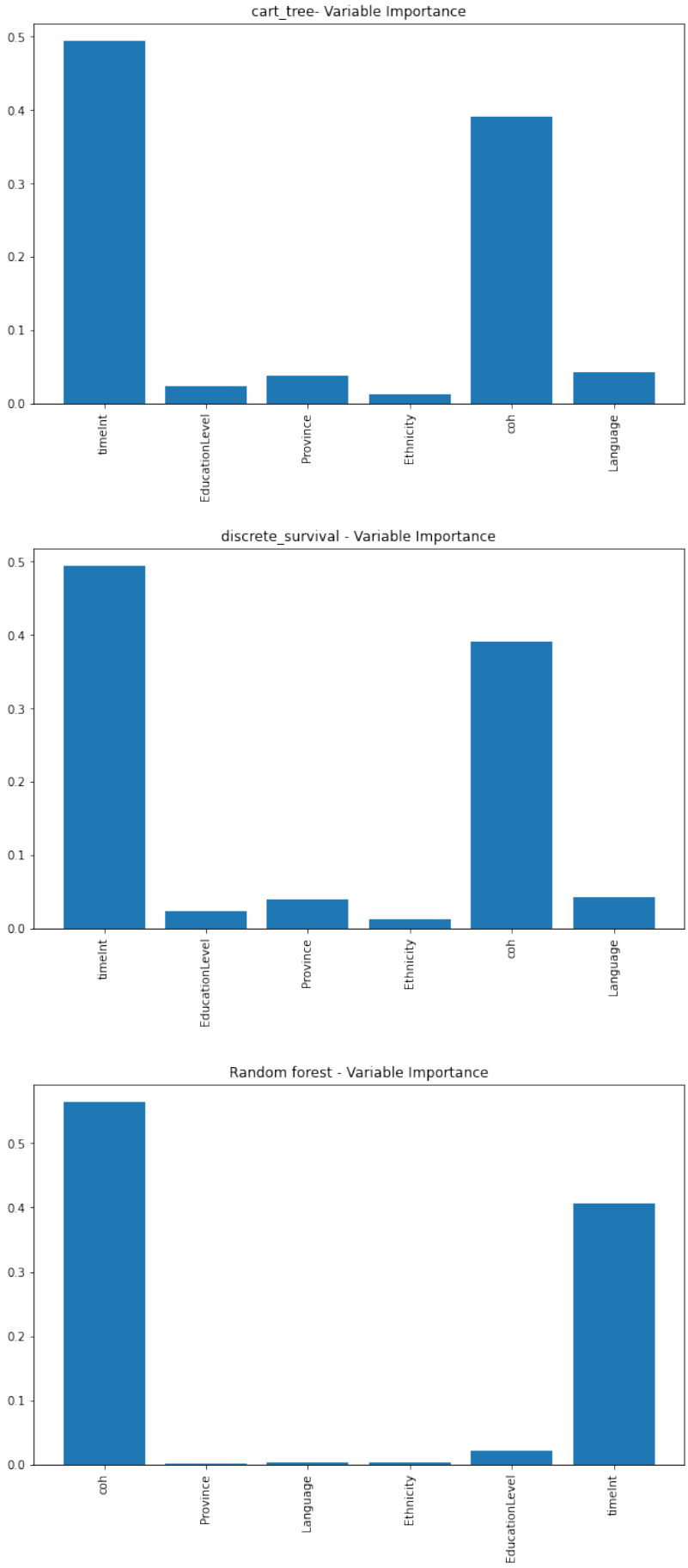

Figure 4.10: Variable importance graphs

According to Figure 4.10 all the models show that the most important variables affecting the age at first marriage for women are the birth years and age. They are followed by province and language for Cart and discrete survival model. For random forests, education level also affects the age at first marriage.

4.2.3 Discussion

The results of the three models – CART (Classification And Regression Trees), Survival Analysis, and Random Forest – highlight several factors that influence the age at first marriage for women in South Africa. The chosen model was random forest as it performed better. The superior performance of the random forest model may be attributed to its ability to capture complex interactions between variables, such as regional, educational, and ethnic factors, which influence marriage risks.

The models suggest that marriage patterns vary across different age groups. For example, the random forest model identifies cohorts born between 1966-1975 or 1976-1985 as potentially having higher marriage prevalence compared to other cohorts.This could be due to sociocultural factors during these periods, where marriage was considered an essential societal institution. Additionally, economic stability in certain regions may have influenced earlier marriages during these years. Studies like those of Brannen and Collard (2023) discuss how economic opportunities and societal pressures have historically encouraged marriage at a younger age.

The models indicate that regional variations exist in marriage patterns. For instance, the CART model highlights distinctions between provinces like Eastern Cape, Northwest, and others regarding marriage prevalence. The differences could be explained by variations in economic development and cultural norms

across regions. For instance, rural areas may have stronger traditional marriage customs, while urban regions might see delayed marriages due to better access to education and employment. Furthermore, economic hardships or job opportunities may drive marriage decisions in certain provinces, as noted in the work of Posel et al. (2011), which connects economic pressures such as lobola costs with marriage timing.

There is association between education level and marriage patterns. The CART and survival tree models indicate that women with higher education might have lower marriage prevalence. This could potentially be linked to increased opportunities for pursuing careers, financial independence, or personal aspirations beyond marriage. Education can alter societal roles for women, leading to delayed marriage as they prioritize career development and personal goals. According to Becker's (1981) theory of marriage, women's increased access to education shifts the opportunity cost of marriage, making them more likely to postpone marriage (Grossbard, 2006). However, socioeconomic status and urban versus rural residency could confound this relationship, as more educated women may reside in urban areas with access to better resources.

The models suggest potential variations in marriage patterns across ethnic and language groups. However, it's crucial to interpret these findings with extreme caution to avoid perpetuating harmful stereotypes. It's important to remember that these models are correlative, not causal, and cannot establish direct links between ethnicity, language, and marriage choices. For instance, language groups may align with distinct cultural practices regarding marriage, but this does not imply a deterministic relationship. Literature such as Elmira et al. (2024) has pointed out that cultural norms often intersect with socioeconomic conditions, meaning that regional disparities in wealth or access to education could be driving the observed patterns rather than ethnicity alone.

The analysis revealed a significant regional disparity in the timing of first marriage from Cart and survival trees. Females residing in Eastern Cape, Gauteng, and North West provinces demonstrated a higher propensity for early marriage compared to their counterparts in Free State, KwaZulu-Natal, Limpopo, Mpumalanga, Northern Cape, and Western Cape. These regional disparities may stem from a combination of socioeconomic conditions, such as higher levels of poverty or traditional practices in the former regions, compared to provinces with more urbanization and modernized economies. For instance, as Posel et al. (2011) discusses, increasing lobola costs could explain delayed marriage in KwaZulu-Natal. Additionally, regions with more educational and employment opportunities may see a lower prevalence of early marriage as women delay marriage to pursue other life goals.

The language spoken also influences the age at first marriage. Women born in 1966-1975 with secondary or higher education and language in Isixhosa, Isizulu, Setswana, Siswati, Isidebele, or Other have a higher risk of getting married compared to those with No education or Primary education. This may reflect the fact that women from specific language groups may face cultural or familial expectations to marry early despite having higher education levels, or their higher education could be aligned with economic stability, enabling marriage at a younger age. However, this relationship might also be confounded by socioeconomic status and regional factors that overlap with language groups.

Overall, the results suggest that a woman's age at first marriage is influenced by a complex interplay of factors, including her birth cohort, education level, province of residence, and language spoken. These factors likely reflect underlying social, cultural, and economic conditions that shape women's experiences and expectations around marriage. The role of socioeconomic status and the urban versus rural divide might serve as potential confounders in these find-

ings. Women in urban areas generally have greater access to education and career opportunities, which can delay marriage, while those in rural regions might face economic pressures or traditional expectations leading to earlier marriages, as evidenced by the work of Elder et al. (2020). Therefore, any analysis must consider these contextual factors when interpreting marriage patterns.

The results of the comparism revealed that random forest boasts the highest median C-Index (0.85), indicating better overall performance in predicting the correct order of marriage events compared to CART (0.78) and the survival model (0.75). Random forest also exhibits the least variability in C-Index values, suggesting its performance is more consistent across different datasets and less sensitive to specific data characteristics or model configurations. These findings align with research highlighting the strengths of Random Forest by Breiman (2001). The study demonstrated that Random Forest generally outperforms other decision tree algorithms like CART due to its ability to reduce variance and improve prediction accuracy. The greater predictive power of random forest further underscores its strength in analyzing complex social patterns where multiple interacting variables are at play, such as regional, educational, and linguistic factors.

4.3 Summary of the chapter

This chapter examines the factors affecting the age at first marriage for women in South Africa using three models: CART, Discrete Survival Analysis, and Random Forest. The findings reveal a complex interplay of factors influencing women's marriage timing.
Key findings:

1. Women born in cohorts 1966-1975 and 1986-1995 were more likely to

marry than those born in 1976-1985.

2. Shorter time interval since leaving school increases the risk of getting married for women born in 1976-1985.

3. Education plays a complex role: higher education increases the likelihood of marrying for women with a short time interval , while lower education increases the likelihood for women with a longer time interval.

4. Women from certain provinces (Eastern Cape, Gauteng, North West) were more likely to marry.

5. Language also influences marriage timing: women with specific languages (Isixhosa, Isizulu, Setswana, Siswati, Isidebele) and higher education were more likely to marry.

These findings suggest that a woman's age at first marriage is influenced by a combination of personal, social, and cultural factors. Understanding these factors can inform interventions and policies aimed at promoting healthy and informed marriage decisions. Random forest outperforms CART and Survival tree.

Chapter 5

Conclusion

We will conclude the study in this chapter by summarizing its main elements and outlining potential methods for strengthening future research efforts.

5.1 Conclusion

The goal of this study was to uncover the determinants of age at first marriage among South African women through the application of three modeling techniques: CART, Discrete Survival Analysis, and Random Forest. Comparing the models, Random Forest emerges as the most accurate model. The results of the three models – CART (Classification And Regression Trees), Discrete Survival Analysis, and Random Forest – underscore the significance of personal, social, and cultural factors in shaping women's marriage timing.

Birth cohort, education level, province of residence, and language spoken all emerged as key determinants of age at first marriage. These factors likely reflect underlying social norms, cultural expectations, and economic conditions that influence women's experiences and choices related to marriage. Under-

standing these factors can inform interventions and policies aimed at promoting healthy and informed marriage decisions among women in South Africa.

5.2 Limitations of the study

While ensemble methods like survival trees and Random Forest offer advantages over single-tree models in terms of stability and prediction accuracy, they also have their limitations. One of the key limitations of ensemble methods is that they can underestimate standard errors, leading to an inflated sense of statistical significance. This underestimation of standard errors can occur due to the correlation between the individual models in the ensemble.

To address this limitation, researchers can consider using multilevel models, which are specifically designed to handle data with a hierarchical structure. In the context of analyzing the determinants of age at first marriage, multilevel models could account for the nesting of individuals within communities or provinces. By incorporating this hierarchical structure, multilevel models can provide more accurate estimates of standard errors, leading to more reliable hypothesis testing. Due to time, this research will not look at multilevel models.

While the study doesn't directly suggest measures to delay marriage, it highlights the importance of addressing the underlying factors that contribute to early marriage in South Africa. This could involve initiatives that empower women through education, economic opportunities, and access to reproductive health services.

To ensure successful interventions, it's crucial to understand and respect the cultural context of South African communities. Programs should be designed

in collaboration with local leaders, educators, and healthcare providers who have a deep understanding of the community's specific needs and challenges. This collaborative approach will foster culturally sensitive interventions that resonate with the target population and have a greater chance of achieving positive outcomes.

It is essential to recognise that the results of this study are derived from a sample of women in South Africa and may not be applicable to other populations. Additionally, the study is limited by the use of secondary data and the lack of information on certain variables, such as religion and socioeconomic status.

5.3 Future studies

In future studies, we can apply the concept of multilevel models to further improve the results of this study. When the assumption that the outcome is independent breaks, we can use multilevel models to control for unobserved heterogeneity among groups or clusters. This is particularly useful when the data exhibits a hierarchical structure, such as individuals nested within communities or families. Multilevel models can effectively account for the correlation between outcomes within the same group, providing more accurate estimates of the effects of individual-level factors and reducing the risk of spurious associations.

For instance, when analyzing the determinants of age at first marriage for South African women, incorporating multilevel models could address potential dependencies within communities or provinces. If women within the same community tend to marry at similar ages, ignoring this clustering effect could lead to an overestimation of the impact of individual-level factors. Multilevel models would allow us to separate the effects of individual characteristics from

the influence of shared community-level factors, providing a more precise understanding of the determinants of age at first marriage.

In conclusion, multilevel models offer a powerful tool for analyzing data with a hierarchical structure, particularly when the assumption of independence is violated. By accounting for unobserved heterogeneity and correlation within groups, multilevel models can enhance the accuracy and interpretability of statistical analyses, leading to more robust and meaningful insights into complex social phenomena.

References

AALEN, O. O., ANDERSEN, P. K., BORGAN, Ø., GILL, R. D., AND KEIDING, N. (2022). Martingales in survival analysis. *In The Splendors and Miseries of Martingales: Their History from the Casino to Mathematics*. Springer, pp. 295–320.

ALDRICH, C., AURET, L., ALDRICH, C., AND AURET, L. (2013). Tree-based methods. *Unsupervised Process Monitoring and Fault Diagnosis with Machine Learning Methods*, 183–220.

ALTMAN, N. AND KRZYWINSKI, M. (2017). Ensemble methods: bagging and random forests. *Nature Methods*, **14** (10), 933–935.

ARIHO, P. AND KABAGENYI, A. (2020). Age at first marriage, age at first sex, family size preferences, contraception and change in fertility among women in uganda: analysis of the 2006–2016 period. *BMC women's health*, **20**, 1–13.

AYIGA, N. AND RAMPAGANE, V. (2013). Determinants of age at first marriage in Sub-Saharan Africa: A comparative study of Uganda and South Africa. *Journal of Social Development in Africa*, **28** (1), 9.

BELACHEW, T. B., NEGASH, W. D., KEFALE, G. T., TAFERE, T. Z., AND ASMAMAW, D. B. (2022). Determinants of early marriage among married women in nine high fertility Sub-Saharan African countries: a multilevel analysis of recent demographic and health surveys. *BMC Public Health*, **22**.
URL: *https://api.semanticscholar.org/CorpusID:254768844*

BEWICK, V., CHEEK, L., AND BALL, J. (2004). Statistics review 12: survival analysis. *Critical Care*, **8**, 1–6.

BOU-HAMAD, I., LAROCQUE, D., BEN-AMEUR, H., MÂSSE, L. C., VITARO, F., AND TREMBLAY, R. E. (2009). Discrete-time survival trees. *Canadian Journal of Statistics*, **37** (1), 17–32.

BRANNEN, J. AND COLLARD, J. (2023). *Marriages in trouble: The process of seeking help*. Taylor & Francis.

BREIMAN, L. (2001). Random forests. *Machine Learning*, **45**, 5–32.

BREIMAN, L., FRIEDMAN, J., OLSHEN, R., AND STONE, C. (1984). Cart. *Classification and Regression Trees*.

BURRELL, T. J., DOUGHTY, C., CERNEAZ, N., ET AL. (2020). Developing a predictive model for cancer survival: A case study in data-driven prognostics. *Reliability Engineering System Safety*, **201**, 107096. doi:10.1016/j.ress.2020.107096.

CARMICHAEL, S. (2011). Marriage and power: Age at first marriage and spousal age gap in lesser developed countries. *The History of the Family*, **16** (4), 416–436.

CARUANA, R. AND NICULESCU-MIZIL, A. (2006). An empirical comparison of supervised learning algorithms. *In Proceedings of the 23rd International Conference on Machine Learning*. pp. 161–168.

COLLETT, D. (2023). *Modelling survival data in medical research*. Chapman and Hall/CRC, Boca Raton, Florida, USA. doi:10.1201/9781003282525.

COURONNÉ, R., PROBST, P., AND BOULESTEIX, A.-L. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*, **19**, 1–14.

COX, D. (1984). *Analysis of Survival Data*. Chapman and Hall/CRC, Boca Raton, Florida, USA. doi:10.1201/9781315137438.

DAVID, G. K. AND MITCHEL, K. (2012). *Survival analysis: a Self-Learning Text*. Spinger, New York.

DEO, S. V., DEO, V., AND SUNDARAM, V. (2021). Survival analysis—part 2: Cox proportional hazards model. *Indian journal of thoracic and cardiovascular surgery*, **37**, 229–233.

DEPARTMENT OF HOME AFFAIRS (2022). Marriage certificates. Government Gazette No. 48914. https://www.dha.gov.za/marriage-certificates. Accessed: 23 June 2023.

DEPARTMENT OF SOCIAL DEVELOPMENT (2021). National child care and protection policy. Government Gazette No. 44636. https://www.gov.za/national-child-care-and-protection-policy. Accessed: 25 June 2023.

DUBE, T., MUTANGA, O., ABDEL-RAHMAN, E. M., ISMAIL, R., AND SLOTOW, R. (2015). Predicting eucalyptus spp. stand volume in Zululand, South Africa: an analysis using a stochastic gradient boosting regression ensemble with multi-source data sets. *International Journal of Remote Sensing*, **36** (14), 3751–3772.

ELDER, G. H., ROBERTSON, E. B., AND ARDELT, M. (2020). Families under economic pressure. *In Families in troubled times*. Routledge, pp. 79–103.

ELENGEMOKE, J. M. AND SUSUMAN, A. S. (2021). Early marriage and correlates among young women in Sub-Saharan African countries. *Journal of Asian and African Studies*, **56** (6), 1345–1368.

ELMIRA, E. S., CHICHAIBELU, B. B., AND QAIM, M. (2024). Marriage customs and nutritional status of men and women. *Food Policy*, **128**, 102734.

FATIMA, S. ET AL. (2023). Rural development and education: critical strategies for ending child marriages.

FIORENTIN, L. D., BONAT, W. H., PELISSARI, A. L., MACHADO, S. D. A., TÉO, S. J., AND ORSO, G. (2020). Generalized linear models for tree survival in loblolly pine plantations. *Cerne*, **25**, 347–356.

GABRIKOVA, B., SVABOVA, L., AND KRAMAROVA, K. (2023). Machine learning ensemble modelling for predicting unemployment duration. *Applied Sciences*, **13** (18), 10146.

GOBENA, M. G. AND BERELIE, Y. (2022). Modeling the determinant of time to age at first marriage among women in Ethiopia using cox models with mixed effects. *Reproductive Health*, **19** (1), 1–6.

GROSSBARD, S. (2006). Beckerâs theories of marriage and the shrinking role of demand and supply models. *San Diego State University, Department of Economics, Working Papers*.

HAFFEJEE, S., TREFFRY-GOATLEY, A., WIEBESIEK, L., AND MKHIZE, N. (2020). Negotiating girl-led advocacy: Addressing early and forced marriage in south africa. *Girlhood Studies*, **13** (2), 18–34.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. H., AND FRIEDMAN, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. 2. Springer, New York.

HOSMER JR, D. W., LEMESHOW, S., AND MAY, S. (2011). *Applied survival analysis: regression modeling of time-to-event data*. John Wiley & Sons, Hoboken, New Jersey, USA.

HOTHORN, T., HORNIK, K., AND ZEILEIS, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, **15** (3), 651–674.

JAMES, G., WITTEN, D., HASTIE, T., TIBSHIRANI, R., ET AL. (2013). *An introduction to statistical learning*. 112. Spinger, New York.

JIANG, F. AND GUTERMAN, E. (2024). Survival analysis. *Statistical Methods in Epilepsy*, 124–142.

JOHN, N. A., EDMEADES, J., AND MURITHI, L. (2019). Child marriage and psychological well-being in niger and ethiopia. *BMC Public Health*, **19**, 1–12.

KLEIN, J. P., MOESCHBERGER, M. L., ET AL. (2003). *Survival analysis: techniques for censored and truncated data*. 1230. Spinger, New York.

KUHN, M., JOHNSON, K., ET AL. (2013). *Applied predictive modeling*, volume 26. Spinger, New York.

KVAMME, H. AND BORGAN, Ø. (2021). Continuous and discrete-time survival prediction with neural networks. *Lifetime data analysis*, **27** (4), 710–736.

LANDMAN, S. (2012). *A multi-model ensemble system for short-range weather prediction in South Africa*. University of Pretoria, South Africa.

LE ROUX, E. (2020). Recognising and responding to complex dilemmas: Child marriage in south africa. *In* GROBBELAAR, J. AND JONES, C. (Editors) *Childhood vulnerabilities in South Africa: Some ethical perspectives*. SUN Press, Stellenbosch. doi:10.18820/9781928480952/07.

LEUNG, K.-M., ELASHOFF, R. M., AND AFIFI, A. A. (1997). Censoring issues in survival analysis. *Annual Review of Public Health*, **18** (1), 83–104.

LIGTHELM, A. (2011). Survival analysis of small informal businesses in south africa, 2007–2010. *Eurasian Business Review*, **1**, 160–179.

MAHARAJ, P. AND SHANGASE, T. (2020). Reasons for delaying marriage: Attitudes of young, educated women in south africa. *Journal of Comparative Family Studies*, **51** (1), 3–17.

MATHABATHA, S. M. L. M. (2023). *Determinants of early marriage among women in South Africa: a multilevel analysis*. Ph.D. thesis, North-West University, South Africa.

MCGRATH, N., NYIRENDA, M., HOSEGOOD, V., AND NEWELL, M.-L. (2009). Age at first sex in rural South Africa. *Sexually Transmitted Infections*, **85**, 49–50.

MNISI, N. (2020). *Forced child marriages as a form of child trafficking in the South African context*. Ph.D. thesis, North-West University, South Africa.

MPOLOKENG, E. (2013). *Socio-Economic and Demographic Factors Affecting Age at First Marriage among Females in South Africa*. Ph.D. thesis, University of Pretoria, Pretoria, South Africa.

MWAMBENE, L. (2018). Recent legal responses to child marriage in Southern Africa: The case of Zimbabwe, South Africa and Malawi. *African Human Rights Law Journal*, **18** (2), 527–550.

NAGY, Á., MUNKÁCSY, G., AND GYŐRFFY, B. (2021). Pancancer survival analysis of cancer hallmark genes. *Scientific Reports*, **11** (1), 6047.

OMOEVA, C. AND HATCH, R. (2022). Teenaged, married, and out of school: effects of early marriage and childbirth on school exit in Eastern Africa. *Prospects*, **52** (3-4), 299–324.

POSEL, D. AND CASALE, D. (2013). The relationship between sex ratios and marriage rates in South Africa. *Applied Economics*, **45** (5), 663–676.

POSEL, D., RUDWICK, S., AND CASALE, D. (2011). Is marriage a dying institution in South Africa? exploring changes in marriage in the context of ilobolo payments. *Agenda*, **25** (1), 102–111.

QUINLAN, J. R. AND CAMERON-JONES, R. M. (1993). Foil: A midterm report. *In Machine Learning: ECML-93: European Conference on Machine Learning Vienna, Austria, April 5–7, 1993 Proceedings 6*. Springer, New York, pp. 1–20.

RANGANATHAN, P. AND PRAMESH, C. (2012). Censoring in survival analysis: potential for bias. *Perspectives in Clinical Research*, **3** (1), 40.

REE, R. H. AND SANMARTÍN, I. (2009). Prospects and challenges for parametric models in historical biogeographical inference. *Journal of Biogeography*, **36** (7), 1211–1220.

SAH, N. ET AL. (2010). Patterns and determinants of age at first marriage of women in Nepal.

SCHMID, M., KÜCHENHOFF, H., HOERAUF, A., AND TUTZ, G. (2016). A survival tree method for the analysis of discrete event times in clinical and epidemiological studies. *Statistics in Medicine*, **35** (5), 734–751.

SCHOBER, P. AND VETTER, T. R. (2018). Survival analysis and interpretation of time-to-event data: the tortoise and the hare. *Anesthesia & Analgesia*, **127** (3), 792–798.

SENI, G. AND ELDER, J. (2010). *Ensemble methods in data mining: improving accuracy through combining predictions*. Morgan & Claypool Publishers, San Rafael, California, USA.

SHUKLA, S., EZEBUIHE, J. A., AND STEINERT, J. I. (2023). Association between public health emergencies and sexual and reproductive health, gender-based violence, and early marriage among adolescent girls: a rapid review. *BMC public health*, **23** (1), 117.

SIMINO, J. M. (2009). *Discrimination and calibration of prognostic survival models*. The Florida State University.

SINGH, M., SHEKHAR, C., AND SHRI, N. (2023). Patterns in age at first marriage and its determinants in India: A historical perspective of last 30 years (1992–2021). *SSM-Population Health*, **22**, 101363.

SMIT, C. (2016). *The use of recursive partitioning to build a financial distress prediction for JSE listed companies*. Master's thesis, University of Cape Town, Cape Town, South Africa.

START, Q. (2022). Getting starting with the randomforestsrc r-package for random forest analysis of regression, classification, survival and more.

STEYN, B. AND HAMMAN, W. (2006). Company failure in South Africa: classification and prediction by means of recursive partitioning. *South African Journal of Business Management*, **37** (4), 7–18.

STROBL, C., MALLEY, J., AND TUTZ, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, **14** (4), 323.

SURESH, K., SEVERN, C., AND GHOSH, D. (2022). Survival prediction models: an introduction to discrete-time modeling. *BMC medical research methodology*, **22** (1), 207.

SVETNIK, V., LIAW, A., TONG, C., AND WANG, T. (2004). Application of breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. *In Multiple Classifier Systems: 5th International Workshop, MCS 2004, Cagliari, Italy, June 9-11, 2004. Proceedings 5*. Spinger, New York, pp. 334–343.

THERNEAU, T. AND ATKINSON, B. (2022). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1.16.
URL: *https://CRAN.R-project.org/package=rpart. Accessed: 25 June 2024*

THERNEAU, T. M. (2022). *A Package for Survival Analysis in R*. R package version 3.3-1.
URL: *https://CRAN.R-project.org/package=survival. Accessed: 25 June 2024*

TREAS, J., LUI, J., AND GUBERNSKAYA, Z. (2014). Attitudes on marriage and new relationships: Cross-national evidence on the deinstitutionalization of marriage. *Demographic research*, **30**, 1495.

TURKSON, A. J., AYIAH-MENSAH, F., AND NIMOH, V. (2021). Handling censoring and censored data in survival analysis: a standalone systematic literature review. *International journal of mathematics and mathematical sciences*, **2021** (1), 9307475.

TUTZ, G., SCHMID, M., ET AL. (2016). *Modeling discrete time-to-event data*. Spinger, New York.

UNICEF (2022). Child marriage in eastern and southern africa. United Nations Children's Fund. https://data.unicef.org/wpcontent/uploads/2022/06/ChildMarriage-in-Eastern-and-Southern-Africa-June-2022-UNICEF-web.pdf. Accessed: 13 October 2023.

WELCHOWSKI, T., BERGER, M., KOEHLER, D., AND SCHMID, M. (2022). *discSurv: Discrete Time Survival Analysis*. R package version 2.0.0.
URL: *https://CRAN.R-project.org/package=discSurv. Accessed: 25 June 2024*

WICKHAM, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, **40** (1), 1–29.
URL: *https://www.jstatsoft.org/v40/i01/. Accessed: 25 June 2024*

WICKHAM, H., FRANÇOIS, R., HENRY, L., MÜLLER, K., AND VAUGHAN, D.

(2023a). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.3. **URL:** *https://CRAN.R-project.org/package=dplyr. Accessed: 25 June 2024*

WICKHAM, H., MILLER, E., AND SMITH, D. (2023b). *haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*. R package version 2.5.3. **URL:** *https://CRAN.R-project.org/package=haven. Accessed: 25 June 2024*

XU, W., CHE, J., AND KONG, Q. (2016). Recursive partitioning method on competing risk outcomes. *Cancer Informatics*, **15**, CIN–S39364.

YAO, W., FRYDMAN, H., LAROCQUE, D., AND SIMONOFF, J. S. (2022). Ensemble methods for survival function estimation with time-varying covariates. *Statistical Methods in Medical Research*, **31** (11), 2217–2236.

ZHANG, H. AND SINGER, B. H. (2010). *Recursive partitioning and applications*. Springer Science & Business Media.