



Analysis of Early Age Fertility in Women using Machine Learning Survival Estimation Approach

By

Maham Safdar

Registration Number 365367

MS Statistics

School of Natural Sciences(SNS)

National University of Sciences and Technology. Islamabad

A dissertation that has been presented in substantial
accomplishment of the graduation requirement **Master of**
Science
in
Statistics

Supervised by: Dr.Shakeel Ahmed

Department of Mathematics & Statistics

School of Natural Sciences
National University of Sciences and Technology
H-12, Islamabad, Pakistan

Year 2024

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS thesis written by Maham Safdar (Registration No. 00000365367), of School of Natural Sciences has been vetted by undersigned, found complete in all respects as per NUST statutes/regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS/M.Phil degree. It is further certified that necessary amendments as pointed out by GEC members and external examiner of the scholar have also been incorporated in the said thesis.


Signature: _____ 

Name of Supervisor: Dr. Shakeel Ahmed

Date: 9-08-2024

Signature (HoD): _____ 


Date: 13-8-2024

Signature (Dean/Principal): _____ 

Date: 15-08-2024

National University of Sciences & Technology**MS THESIS WORK**

We hereby recommend that the dissertation prepared under our supervision by: "**Maham Safdar**" Regn No. **00000365367** Titled: "**Analysis of Early Age Fertility in Women Using Machine Learning Survival Estimation Approach**" accepted in partial fulfillment of the requirements for the award of **MS** degree.

Examination Committee Members1. Name: PROF. TAHIR MAHMOODSignature: 2. Name: DR. MUHAMMAD ASIF FAROOQSignature: Supervisor's Name: DR. SHAKEEL AHMEDSignature: 


Head of Department

13-8-2024
Date

COUNTERSIGNEDDate: 15.08.2024


Dean/Principal

*This thesis is dedicated to my parents for their
love, endless support, and encouragement.*

Acknowledgment

All honor and praise should be given to "**Allah Almighty**", the most merciful and kind, who created the entire universe. He has showered upon me innumerable favors, including the courage and strength to successfully complete my thesis, for which I am incredibly appreciative and thankful to Him.

Unquestionably, I want to express my sincere gratitude to my Supervisor, **Dr.Shakeel Ahmed**, one of the best professors I've ever had. I owe them a heartfelt thanks for their encouragement, guidance, and above all, their constant patience during this process. May Allah bless them with His countless blessings. Without their expertise, support, and guidance, this research could not have been completed. Furthermore, because of their efforts and appreciative responses to my questions, I have gained a complete grasp and respect for this field. I also want to express my gratitude to my GEC members **Dr.Tahir Mehmood** and **Dr. Muhammad Asif Farooq** for their support and guidance in completing this thesis. And lastly, a million thanks to my siblings and friends, for their support throughout my studies. GOD bless you all for making this difficult journey a success.

Maham Safdar.

August, 2024.

Abstract

Early age fertility is a significant public health concern in Pakistan, with profound implications for women’s health, socio-economic development, and population dynamics. This study explores the socio-demographic factors influencing early age fertility among women in Pakistan using the Pakistan Demographic and Health Survey (PDHS) dataset from 2017-2018. The research aims to identify the key socio-demographic determinants of early age fertility and to evaluate the effectiveness of advanced survival analysis models in predicting these outcomes.

The study addresses two primary objectives: first, to identify and analyze socio-demographic factors associated with an increased risk of early age fertility, and second, to compare the predictive performance of three advanced survival models—the Cox Proportional Hazards Model (CPH), Random Survival Forest (RSF) Model, and Conditional Inference Forest (CIF) Model—in the context of early age fertility prediction.

To achieve the first objective, the study employs the CPH model to assess the impact of various socio-demographic factors on early age fertility. The results indicate that lower educational attainment, rural residence, and lower socio-economic status are significantly associated with an increased risk of early age fertility. Specifically, women with no education have a 4.858 times increased risk of early age fertility compared to those with higher education, and those living in rural areas face a 1.123 times greater risk compared to their urban counterparts. Additionally, women from poor socio-economic backgrounds are 1.229 times more likely to experience early age fertility than those from rich backgrounds.

For the second objective, the study compares the performance of the CPH, RSF, and CIF models in predicting early age fertility outcomes based on two variables: Age at Marriage (AAM) and Age at First Birth (RAAFB). The CPH model proves to be

the most effective for predicting early age fertility related to AAM, as it exhibits the lowest prediction error and Integrated Brier Score (IBS), along with a high C-index indicating reliable predictions. Conversely, for the RAAFB variable, the CIF model is identified as the best model due to its lowest prediction error rate and IBS score, despite RSF's higher C-index. The CIF model's superior performance in terms of error rate and IBS demonstrates its capability for precise prediction of early age fertility outcomes, making it the preferred choice for this aspect of the analysis.

This study contributes to the understanding of early age fertility in Pakistan by identifying critical socio-demographic factors and demonstrating the effectiveness of advanced statistical models for predicting early age fertility. The findings emphasize the need for targeted interventions addressing educational disparities, rural-urban differences, and socio-economic inequalities to mitigate the risks associated with early age fertility. Additionally, the study highlights the importance of using comprehensive performance metrics for model selection, with CIF emerging as the optimal model for precise prediction of early age fertility outcomes based on RAAFB.

Contents

Acknowledgment	v
Abstract	v
LIST OF TABLES	vi
LIST OF FIGURES	vii
1 Introduction	1
1.1 Survival Analysis	2
1.1.1 Survival Function (S(t))	3
1.1.2 Hazard Function (h(t))	4
1.2 Problem Statement	6
1.3 Research Questions	6
1.4 Research Objectives	7
1.5 Significance of the Study	7
1.6 Outline of thesis	8
2 Literature Review	9
2.1 Historical Perspectives on Early Age Fertility	9
2.2 Theoretical Frameworks for Understanding Early Age Fertility	10
2.3 Factors Influencing Early Age Fertility	11
2.4 Previous Research on Fertility Modeling	12
2.5 Survival Analysis Methods (Cox, RSF, Conditional Inference Forest)	13

3	Research Methodology	16
3.1	Data Collection	17
3.2	Variables in the Study	17
3.3	Data Preprocessing	20
3.4	Study Design	22
3.5	Cox Proportional Hazards Model	23
3.6	Random Survival Forest (RSF)	24
3.7	Conditional Inference forest	27
3.8	Model Evaluation Metrics	29
4	Results and Discussions	32
4.1	Descriptive Results	32
4.2	Cox Model	35
4.2.1	Age at Marriage as a Time Variable	35
4.2.2	Respondent Age at First Birth as a Time Variable	38
4.3	Random Survival Forest Model	42
4.3.1	AAM as a Time Variable	42
4.3.2	RAAFB as a Time Variable	46
4.4	Prediction Error Curve Comparisons	49
4.4.1	AAM as a Time Variable	49
4.4.2	RAAFB a Time Variable	50
4.5	Integrated Brier Scores	52
4.5.1	AAM	52
4.5.2	RAAFB	54
4.6	Concordance Index	55
4.6.1	AAM	55
4.6.2	RAAFB	57
5	Conclusion	59
5.1	Limitations	61
5.2	Future Recommendations	61

List of Tables

3.1	Variable Descriptions and Codes	19
4.1	Summary Statistics for RAAFB and AAM	32
4.2	Frequency and Percentage Frequency for Various Variables	33
4.3	Cox Model Results	36
4.4	Comprehensive Statistical Analysis Results	39

List of Figures

3.1	Overall Study Design	23
3.2	Working Algorithm of Random Survival Forest: The RSF algorithm uses bootstrap sampling to create multiple trees, calculates the CHF for each tree, and combines them to make robust survival predictions.	26
4.1	Forest Plot of Cox Proportional Hazards Model	37
4.2	Forest Plot of Cox Proportional Hazards Model	41
4.3	Number of tree and variable importance AAM	44
4.4	Number of tree and variable importance RAAFB	47
4.5	Comparison of Prediction Error Curves AAM	49
4.6	Comparison of Prediction Error Curves RAAFB	51
4.7	Box Plot of Integrated Brier Scores AAM	53
4.8	Box Plot of Integrated Brier Scores RAAFB	55
4.9	Comparison of Concordance Index AAM	56
4.10	Comparison of Concordance Index RAAFB	58

List of Abbreviation

CPH	Cox Proportional Hazards (Model)
RSF	Random Survival Forest (Model)
CIF	Conditional Inference Forest (Model)
PDHS	Pakistan Demographic and Health Survey
CHF	Cumulative Hazard Function
SF	Survival Function
VIM	Variable Importance Measure
OOB	Out-of-Bag (Estimates)
IBS	Integrated Brier Score
CI	Confidence Interval
Coef	Coefficient
C-Index	Concordance Index

Chapter 1

Introduction

Early age fertility refers to the occurrence of childbirth among girls and young women under the age of 18 [1]. Despite global efforts to improve maternal and child health, addressing early age fertility is complex due to its multifaceted nature. Over the past few decades, remarkable progress has been made globally in lowering early marriage, teenage pregnancy, and maternal death, yet teenage pregnancy remains a social and public health concern.[2, 3, 4, 5].About 12 million girls aged 15 to 19 years give birth annually in low- and middle-income countries, which shows how serious this problem is[6, 7].Teen pregnancy and childbirth are bad for the person, society,and the world[8, 9].Teenage motherhood affects a girl’s well-being and education and keeps her from reaching her full potential[10, 11, 12, 13, 14]. In Pakistan, teenage girls are frequently married and have children, especially in rural areas [15] Early-age fertility among Pakistani women is a multifaceted phenomenon with unique cultural and socioeconomic dimensions. In Pakistan, a significant number of women become mothers during their teenage years, and this early childbearing has far-reaching implications.[16]. Socio-economic disparities, educational limitations, cultural norms, and lack of awareness about reproductive health are intertwined factors contributing to early pregnancies[17]. These challenges are not isolated but form a complex web, necessitating advanced analytical methods to unravel their interconnections. Traditional statistical analyses often struggle to capture the intricate relationships within the data. Therefore, employing sophisticated models such as the Cox Proportional Hazards Model, Random Survival Forest (RSF), and Conditional Inference Forest becomes essential. These models excel in handling intricate survival data and can unveil

hidden patterns and predictors that might remain obscured in simpler analyses[18]. In this context, the Pakistan Demographic and Health Survey (PDHS) dataset collected in 2017-2018 is used. This dataset provides a comprehensive repository of socio-demographic information, healthcare utilization, and reproductive behavior, making it a rich source for understanding early age fertility patterns in Pakistan. By harnessing the power of this dataset in conjunction with advanced statistical models, this research endeavors to shed light on the specific determinants of early age fertility. Analyzing the PDHS data using the Cox Proportional Hazards Model, RSF, and Conditional Inference Forest offers a unique opportunity to explore the time-to-event nature of early pregnancies, providing a detailed understanding of the risk factors involved. The Cox Proportional Hazards Model allows for the analysis of survival data, considering the duration until an event[19], such as early childbirth, occurs. RSF and Conditional Inference Forest, on the other hand, harness the power of ensemble learning and decision trees, accommodating non-linear relationships and interactions between variables[20]. By combining these models with the robust data set like PDHS, this research aims not only to identify the predictors of early age fertility but also to quantify their impact, providing a comprehensive and nuanced understanding of this phenomenon. The objective is to employ a machine-learning survival model to estimate the fertility of such women. Utilizing the Pakistan Demographic and Health Survey (PDHS) dataset from 2017-18, the study seeks to calculate the proportion of women who married before 18 and experienced their first childbirth after the age of 20. In Pakistan, child marriage affects a significant portion of ever-married women, elevating their risk of experiencing high fertility rates and encountering poor fertility health outcomes. It underscores the urgent need to advocate for raising the legal marriage age for women in Pakistan. Effective measures to combat child marriage, including stringent law enforcement and the promotion of civil, sexual, and reproductive health rights for women, are essential steps toward eliminating this issue from the country.

1.1 Survival Analysis

Survival analysis is a statistical approach used to analyze the time until an event of interest occurs[21, 22]. This method is widely employed in various fields, including

medicine, biology, economics, engineering, and social sciences, to study the time until an event, such as death, relapse, failure, or occurrence of a specific outcome[23]. Survival analysis is particularly useful when dealing with data in which the outcome of interest is a time-to-event variable and some of the individuals in the study may not experience the event during the observation period. In the context of this research on early age fertility in women, survival analysis provides a robust framework to explore and understand the timing of important life events such as marriage and first childbirth[24]. In survival analysis, two key concepts are paramount: survival time and censoring[25]. Survival time is the time elapsed until the occurrence of the event of interest. It could be in any unit of time, such as days, months, or years. In this study, survival time refers to the duration between a woman's birth and the occurrence of specific events, such as marriage or the birth of her first child. It is often measured in years or months, representing the time Censoring occurs when the survival time of an individual is not observed completely. In longitudinal studies, not all participants may experience the events of interest within the study period. This situation is called censoring.

There are two types of censoring:

Right-censoring: The event of interest has not occurred by the end of the study period. For example, some women might not have married or given birth by the time the study concludes.

Left-censoring: The event of interest has occurred before the study started, but the exact time is unknown[26].

In survival analysis, the survival function and the hazard function are fundamental concepts used to describe the probability of survival over time and the instantaneous risk of an event occurring at a specific point in time, respectively.

1.1.1 Survival Function (S(t))

The survival function, denoted as $S(t)$, represents the probability that a subject will survive beyond a specified time t . In other words, it gives the likelihood that an event of interest (such as death, failure, or relapse) has not occurred by time t . Mathematically, the survival function is defined as: $S(t)$

$$S(t) = P(T > t)$$

Where: $S(t)$ is the survival function at time t . T is the random variable representing the time until the event of interest occurs. For censored data (where some individuals are not followed until the event occurs), the survival function is estimated using methods like the Kaplan-Meier estimator, which calculates the probability of surviving beyond each observed time point. The survival function ranges from 0 to 1, where $S(t)=1$ indicates that everyone has survived up to time t , and $S(t)=0$ indicates that no one has survived beyond time t .

1.1.2 Hazard Function ($h(t)$)

The hazard function, denoted as $h(t)$, represents the instantaneous risk or probability per unit of time that an event will occur at time t , given that the subject has survived up to time t . In other words, it describes the likelihood of the event happening in the next instant, provided the subject has survived up to time t . Mathematically, the hazard function is defined as:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \quad (1.1)$$

Where $h(t)$ is the hazard function at time t (1.1) [27]. T is the random variable representing the time until the event of interest occurs. Δt represents a small interval of time. The hazard function can vary over time, indicating how the risk of the event changes as time progresses. A constant hazard function means the event is equally likely to occur at any given time, while a changing hazard function signifies that the risk of the event occurring varies over time. Understanding the survival and hazard functions is crucial in survival analysis as they provide valuable insights into the probability of event occurrence and the instantaneous risk associated with a particular time point, respectively. Researchers use these functions to make predictions, compare different groups, and draw conclusions about the factors influencing the event of interest. In the field of bio-statistics, a highly active area of research revolves around survival analysis, which focuses on analyzing time-to-event outcomes, often with data that are censored, meaning the event of interest has not occurred for some individuals by the end of the study period. However, recent advancements in data learning technology have led to the availability of high-dimensional datasets for researchers. While this

influx of data has opened new avenues for research, it has also presented challenges in analyzing survival data effectively. In this context, traditional survival analysis methods like the Cox proportional hazard (CPH) regression, which have been valuable due to their straightforward interpretation of covariate effects and ease of inference, are no longer adequate. The complexity of high-dimensional data requires more sophisticated and nuanced approaches to extract meaningful insights. Researchers are now exploring innovative statistical techniques and machine learning algorithms tailored to handle the intricacies of these large and complex data sets. These modern methods not only account for the challenges posed by high-dimensional data but also enable researchers to uncover hidden patterns and associations in survival outcomes, ultimately enhancing our understanding of the underlying biological and clinical processes. As a result, the intersection of survival analysis and data learning technology continues to be a vibrant and evolving area of research in bio-statistics.

1.2 Problem Statement

Early-age fertility is a significant public health issue in Pakistan, contributing to adverse maternal and child health outcomes, perpetuating cycles of poverty, and hindering socio-economic development. Despite various interventions, early-age fertility remains prevalent, particularly among women in rural areas with lower education levels and limited access to contraceptive methods.

This study seeks to identify the determinants and predictors of early-age fertility among Pakistani women using the Pakistan Demographic and Health Survey (PDHS) dataset 2017-2018 and advanced statistical models. By employing the Cox Proportional Hazards Model (CPH), Random Survival Forest (RSF), and Conditional Inference Forest (CIF), the study aims to:

1. Determine the key socio-demographic factors influencing the age at marriage (AAM) and respondent age at first birth (RAAFB).
2. Compare the predictive accuracy of CPH, RSF, and CIF models in forecasting early-age fertility.
3. Provide evidence-based recommendations for policy interventions to reduce early-age fertility rates and improve reproductive health outcomes.

Understanding these determinants and predictors is crucial for developing effective policies and interventions to address early-age fertility, thereby improving the overall well-being of women and contributing to socio-economic progress in Pakistan.

1.3 Research Questions

1. What socio-demographic factors are associated with an increased risk of early age fertility among women in Pakistan?
2. How do the Cox Proportional Hazard Model, Random Survival Forest (RSF) Model, and Conditional Inference Forest Model differ in their ability to predict early age fertility in women based on the PDHS dataset?

1.4 Research Objectives

This study is conducted to analyze the prevalence and trends of early-age fertility in Pakistani women using the Pakistan Demographic and Health Survey (PDHS) dataset 2017-2018. The specific objectives are as follows:

- I. The objective of this study is to identify and analyze socio-demographic factors associated with an increased risk of early age fertility among women in Pakistan using the PDHS dataset.
- II. Compare and evaluate the predictive performance of the Cox Proportional Hazard Model, Random Survival Forest (RSF) Model, and Conditional Inference Forest Model in predicting early age fertility. These research objectives guide the comprehensive investigation into early-age fertility, emphasizing the need for advanced analytical approaches to unravel its complexity within the Pakistani context.

1.5 Significance of the Study

This study holds profound significance as it addresses a critical issue with far-reaching implications for healthcare, policy, and demographic research in Pakistan. By delving into the complex realm of early-age fertility in Pakistani women, it offers a unique opportunity to improve maternal and child healthcare services, inform evidence-based policies, enrich demographic research, and empower young women. The findings have the potential to enhance the healthcare landscape by reducing maternal and infant mortality rates associated with teenage pregnancies. This research contributes valuable insights to the field of demographic research, aiding in understanding fertility patterns and drivers in Pakistan. It also underscores the importance of educational empowerment for young mothers and advocates for gender equality and women's rights. In essence, this study has the power to catalyze positive change, fostering better health outcomes, gender equality, and socioeconomic development in Pakistan.

1.6 Outline of thesis

This thesis is organized into several chapters, each addressing specific aspects of the analysis. Chapter one presents a background on early age fertility, this section also delves into the significance of the study, articulates the problem statement we aim to address, outlines the objectives we seek to achieve, and formulates the research question that guides our investigation. Moving on to the second chapter, we delve into an extensive literature review of our research study. In this chapter, we meticulously explore the origins and practical applications of the specific methods we have chosen for our research. By delving into the details, we understand the context in which these methods are employed. Chapter three is dedicated to providing the data sources, preprocessing steps, and variable selection processes. Chapters four, five, and six will delve into the Cox Proportional Hazards model, RSF model, and Conditional Inference Forest model, respectively, explaining their methodologies and presenting the results of our analysis. Chapter seven encapsulates the conclusion of our entire study. In this concluding chapter, we summarize our research's key findings, insights, and contributions. We tie together the threads explored throughout the thesis and draw overarching conclusions that provide closure to the study. Overall, this thesis is structured in a manner that allows us to progressively build understanding, from establishing foundational concepts and investigating prior work to formulating innovative methods, evaluating results, and ultimately arriving at a comprehensive conclusion. In essence, this study endeavors to contribute valuable insights into early-age fertility in women through the application of advanced survival analysis models. By examining both traditional and novel factors influencing early-age fertility, we aim to provide a comprehensive understanding of this crucial aspect of reproductive health and equip stakeholders with data-driven tools for informed decision-making and policy formulation.

Chapter 2

Literature Review

Early-age fertility among women is a significant and complex issue in contemporary society. It specifically refers to pregnancies occurring in women under the age of 20. This phenomenon presents a multifaceted challenge with considerable implications for public health and social development [28]. Adolescents experiencing early-age fertility face heightened health risks, including complications during pregnancy and childbirth. This can also lead to adverse outcomes for their children [29]. Additionally, early motherhood can result in enduring educational and economic repercussions for young women, thereby limiting their future opportunities and perpetuating a cycle of disadvantage [30]. Addressing early-age fertility is multifaceted in its significance. It intersects with reproductive rights and gender equality, often entangled with social norms and gender-based disparities [31]. Tackling this issue is also vital for broader public health and development objectives, such as reducing maternal mortality and alleviating poverty [32].

2.1 Historical Perspectives on Early Age Fertility

Early age fertility, where young women have children, is a phenomenon deeply rooted in human history. In historical terms, early age fertility has been a constant feature, reflecting the societal norms and survival strategies of different eras [33]. In many early societies, having children at a younger

age was often the norm, driven by shorter life expectancies and the need for larger families to support agrarian or labor-intensive lifestyles [34]. This practice was further reinforced by social customs such as arranged marriages and the significant influence of religious and cultural beliefs on reproductive decisions [35]. In many cultures, these groups played a critical role in dictating the timing and frequency of childbirth, which in turn impacted early age fertility rates [36]. However, as societies evolved, so did attitudes towards early childbearing. The advent of modern healthcare, including improved prenatal care and birthing practices, has changed the landscape of early age fertility significantly [37]. Historical events like wars, economic upheavals, and pandemics have also played a crucial role in shaping fertility trends. These events often led to fluctuations in fertility rates, prompting societal and policy responses aimed at stabilizing population growth [38]. The introduction of family planning and educational initiatives has contributed to a shift in early age fertility patterns [39]. This historical journey through the landscape of early age fertility highlights the intricate interplay of various factors that have influenced it over time. By examining these elements, we gain a comprehensive understanding of the evolution of societal attitudes and practices surrounding early age fertility. This understanding is crucial for addressing the challenges associated with early age fertility in the modern world.

2.2 Theoretical Frameworks for Understanding Early Age Fertility

Examining early age fertility requires a complex approach, using different theories to understand its many aspects. This section looks at theories like the life course perspective, social determinants of health, and reproductive health theories, showing how they help us study early age fertility and its causes. The life course perspective is key to understanding early age fertility [40]. This idea sees people's lives as a series of connected events and changes. For early age fertility, it means looking at how having children

early can affect someone's whole life, including their education, job, and overall well-being. This perspective also helps us see how things that happen in childhood and as a teenager can lead to early childbearing. The social determinants of health framework highlights how social, economic, and environmental factors shape health [41]. Reproductive health theories focus on health during the reproductive years [42]. These theories help us look at how early childbearing affects the health of the mother and child, what choices they have, and how they access healthcare. They also cover things like birth control, family planning, and how health policies and programs work. Using these theories gives a deeper understanding of the health side of early age fertility and what can be done to improve reproductive health for young women. Together, these theories give a full picture of early age fertility.

2.3 Factors Influencing Early Age Fertility

In Pakistan, early-age fertility, which refers to teenage pregnancies and childbirth, is shaped by various socio-cultural and economic factors unique to the country. When exploring the reasons behind early-age fertility, several critical areas emerge:

Socioeconomic Status plays a significant role in teenage pregnancies in Pakistan. Studies show that income, occupation, and living conditions greatly influence early-age fertility [43]. Young women from economically disadvantaged backgrounds face a higher risk due to limited resources, poor healthcare, and lower living standards. There's a strong link between a girl's education level and her chances of becoming a teenage mother [44]. Education empowers young women with knowledge and skills and often delays marriage and childbirth as they focus on their studies. Thus, educational attainment is a crucial factor in determining early pregnancy rates in Pakistan. Pakistan's diverse ethnic and cultural communities have varying norms about early marriage and childbearing [45]. Demographic factors, like population density and urban-rural differences, also affect teenage pregnancy rates. The

availability and quality of healthcare, particularly in rural areas, are vital in understanding early-age fertility patterns in Pakistan [46]. Limited access to quality healthcare affects family planning and reproductive health services.

2.4 Previous Research on Fertility Modeling

Early age fertility has been a focus for researchers, leading to extensive studies aiming to understand its complexities. These studies have used various methods, including both quantitative and qualitative approaches, to investigate the multiple factors affecting young women’s decisions to have children. Methodologically, these studies have employed regression analysis, survival models, and qualitative interviews, contributing significantly to our understanding of early age fertility. Data from large-scale surveys, cohort studies, and demographic and health surveys have been crucial in supporting these findings [47]. However, previous research on early age fertility modeling faces limitations. Issues with data quality and availability, especially in low-resource settings, are common. Many studies are cross-sectional, limiting their ability to establish causality or explore temporal relationships [48]. While qualitative studies offer depth, their generalizability is often limited. Additionally, the diversity of contextual factors across different regions complicates the ability to draw broad conclusions, highlighting the need for context-specific research [49]. There remain gaps in the research on early age fertility modeling. These include the need for more comprehensive studies that include a wider range of influencing factors and the exploration of how these factors change over time. Also, there’s a need for deeper examination of regional and contextual disparities to understand geographical and cultural influences on early age fertility [50]. This study aims to address these gaps. Using advanced statistical modeling techniques like Cox proportional hazards models, Random Survival Forests (RSF), and Conditional Inference Forests, and leveraging data from the Pakistan Demographic and Health Survey (PDHS), this research seeks to offer a more detailed and context-sensitive analysis of early age fertility in Pakistan. This approach

aims to deepen our understanding of the dynamics and determinants of early age fertility, advancing the field's knowledge.

2.5 Survival Analysis Methods (Cox, RSF, Conditional Inference Forest)

Survival analysis is a crucial statistical method widely used in various research fields, especially for studying time-to-event data. This approach is particularly relevant for examining events that unfold over time, such as mortality, disease occurrence, or, in the context of this research, early age fertility. This model evaluates the hazard rate, which is the chance of an event happening at a specific time, assuming it hasn't occurred yet. The Cox model's strength lies in its capacity to assess how various factors or covariates influence the hazard rate. In the context of early age fertility, it allows for an in-depth examination of how elements like socioeconomic status, education, and healthcare access affect the timing of a woman's first childbirth [51]. Recent advancements have led to sophisticated techniques like the Random Survival Forest (RSF) and Conditional Inference Forest (CIF) models. These models, based on decision trees and ensemble methods, are particularly effective for complex, high-dimensional datasets typical in early age fertility research. They excel in capturing non-linear relationships and interactions among covariates, offering a comprehensive view of how various factors jointly influence childbirth timing [52, 53]. RSF and CIF are also adept at handling censored data, a common challenge in early age fertility studies where not every participant may experience their first childbirth during the study. These models integrate censored data while providing reliable predictions and insights, making them highly suitable for exploring the temporal aspects of early age fertility and its determinants [52, 53]. These advanced methods enable a thorough examination of complex relationships, manage censored data effectively, and contribute to a deeper understanding of the dynamics and influencing factors of early age fertility. The 2017 study by Nasejje et al. compares the performance of two random survival forest

models and a conditional inference forest model on time-to-event data, using both simulated and real datasets[54]. The simulated data incorporated various covariate attributes to robustly test model efficacy under different conditions. The real-world datasets included one on under-five child mortality in Uganda and another on extensively drug-resistant tuberculosis in South Africa. The study found that conditional inference forests outperformed random survival forests in simulation, particularly with covariates that had many split-points. In real-world application, the conditional inference forest model yielded the lowest prediction error for the child mortality dataset, though its performance was similar to that of the random survival forests for the tuberculosis dataset. [55] The study focused on Survival Trees, a key tool in survival analysis for managing time-to-event data, especially with right-censored data common in this field. Through a comprehensive review of literature and various models such as CART and its extensions, the research applied these models to real-world datasets to validate their efficacy. Findings indicated that Survival Trees excel at uncovering complex relationships in survival data where traditional models like Cox Proportional Hazards might not perform as well. This underscores the value of Survival Trees in medical and biological research, enhancing the understanding of survival data, risk factors, and their interactions, thereby enriching survival analysis methodologies. [18] aimed to compare the efficacy of survival forest models—specifically Random Survival Forest (RSF) and Conditional Inference Forest (CIF)—with the traditional Cox-proportional hazards (CPH) model in analyzing the first birth interval (FBI) among women. Utilizing a cross-sectional dataset of 610 married women aged 15-49 in Tehran, the study employed RSF and CIF models developed through bootstrap sampling using R-language packages, focusing on various influential covariates on FBI. The results demonstrated that the RSF model surpassed both CIF and CPH models in prediction accuracy, as measured by the out-of-bag (OOB) C-index and integrated Brier score (IBS), with a woman’s age being identified as the most significant predictor of FBI. The study concluded that

applying suitable methods like RSF in analyzing FBI is crucial for informed policy-making aimed at addressing fertility rate changes, highlighting the superior predictive performance of RSF in the context of fertility analysis. The literature on early age fertility examines the interplay of historical, sociocultural, and individual factors influencing young women's fertility patterns, highlighting the role of arranged marriages, religious beliefs, societal norms, and major historical events such as wars and economic changes. Changes in societal views and healthcare advancements have shifted narratives around the benefits and risks associated with early fertility. Key determinants such as socioeconomic status, education, healthcare access, and cultural beliefs play critical roles in the timing of a woman's first childbirth, with research showing varied and context-specific findings. To better understand these complexities, recent studies have utilized advanced survival analysis techniques, providing deeper insights into the factors affecting early age fertility and aiding in the development of targeted interventions. This sophisticated approach enhances our understanding of early age fertility and its diverse influencing factors.

Chapter 3

Research Methodology

In this chapter, the methodology employed to analyze early-age fertility among Pakistani women using the Pakistan Demographic and Health Survey (PDHS) dataset from 2017-2018 is outlined. The chapter begins by discussing the primary data source and its significance in providing comprehensive information on demographic and health-related indicators. Subsequently, preprocessing steps such as data cleaning, variable selection, sample selection, data transformation, and imputation are described to ensure the quality and compatibility of the dataset with the chosen modeling techniques. The variable selection process is then detailed, highlighting socio-demographic, cultural, and healthcare-related factors influencing early-age fertility. Following this, the chapter delves into the advanced survival analysis techniques employed, including the Cox Proportional Hazards model, Random Survival Forest (RSF) model, and Conditional Inference Forest model. Each method is explained in depth, elucidating its theoretical underpinnings and practical application in analyzing time-to-event outcomes. Additionally, the assessment of model performance through integrated Brier scores (IBS), incorporating Bootstrap cross-validated estimates to address overfitting, is discussed. By leveraging these sophisticated methodologies, the analysis aims to uncover significant predictors of early-age fertility and evaluate the predictive accuracy of the models, ultimately contributing valuable insights to the understanding of reproductive health dynamics in Pakistan.

3.1 Data Collection

The data for this study was obtained from the PDHS 2017-2018, which involves a complex, multistage sampling design. The survey collected data through interviews with eligible women of reproductive age (15-49 years) across various regions of Pakistan[56]. These interviews gathered key information on age at marriage, age at first birth, socioeconomic factors, and other relevant variables[57]. It covers a wide range of topics related to demographics, health, family planning, and socio-economic characteristics. The dataset usually consists of a large sample size, representing a cross-section of the Pakistani population. The sample is designed to be nationally representative, allowing for meaningful analysis at both the national and regional levels[58].

3.2 Variables in the Study

In this study, the primary focus is on understanding the age at which women experience their first childbirth, a significant milestone with implications for public health and social policies. Two key variables are central to this investigation:

Respondent Age at First Birth (RAAFB): This variable is recorded as numeric values, such as 12 or 42, representing the age at which respondents typically have their first child. RAAFB serves as a crucial demographic indicator, allowing researchers to examine the time from birth to the first childbirth. It helps assess how various socio-demographic factors influence the age at which women give birth for the first time and their risk of experiencing early-age fertility. The dataset categorizes women who give birth at age 20 or younger as 1, indicating an early fertility event, and those who give birth after 20 as 0, indicating censored observations.

Age at Marriage (AAM): Similarly recorded in years, such as 9 or 38, AAM captures the age at which respondents get married. This demographic measure is pivotal for understanding early-age fertility outcomes. AAM is

used to analyze the timing of marriage relative to birth events and investigate its influence on the risk of early-age fertility. By exploring socio-demographic factors affecting the timing of marriage, AAM sheds light on its association with early fertility outcomes.

These variables, "Age at Marriage" and "Age at First Birth," quantified in years, capture critical moments in a woman's life: marriage and the birth of her first child, respectively. Their analysis provides insights into the complex dynamics of early-age fertility, offering valuable information for shaping public health policies and social interventions aimed at supporting women's reproductive health.

[59, 60, 61]. Several predictors were considered in this study to analyze early age fertility in women; these are as follows:

The first variable, "Status," indicates whether an individual is experiencing early-age fertility, with a coding of 1 for Yes and 0 for No. This binary classification helps in distinguishing between those who have early-age fertility and those who do not.

"REL" stands for Respondent Education Level, which is categorized into four levels: 0 for no education, 1 for primary education, 2 for secondary education, and 3 for higher education. This variable is crucial in assessing the impact of educational attainment on various outcomes.

"REG" refers to the Region, which identifies the respondent's location within the country. The regions are coded as follows: 1 for Punjab, 2 for Sindh, 3 for Khyber-Pakhtunkhwa (KPK), 4 for Balochistan, 5 for Gilgit Baltistan (GB), 6 for Islamabad Capital Territory (ICT), 7 for Azad Jammu and Kashmir (AJK), and 8 for the Ex-Federally Administered Tribal Areas (FATA). This geographical categorization allows for regional comparisons in the analysis.

"HEL," or Husband Education Level, is coded similarly to the respondent's education level, with 1 indicating no education, 2 for primary, 3 for secondary, and 4 for higher education. This variable examines the educational background of the respondent's spouse.

Table 3.1: Variable Descriptions and Codes

Variable ID	Variable Name	Variable Description
Event	Experiencing Early-Age Fertility	1 = Yes, 0= No
REL	Respondent Education Level	0 = No education, 1 = Primary, 2 = Secondary, 3 = Higher
REG	Region	1 = Punjab, 2 = Sindh, 3 = Khyber-Pakhtunkhwa (KPK), 4 = Balochistan, 5 = Gilgit Baltistan (GB), 6 = Islamabad Capital Territory (ICT), 7 = Azad Jammu and Kashmir (AJK), 8 = Ex-Federally Administered Tribal Areas (FATA)
HEL	Husband Education Level	1 = No education, 2= Primary, 3 = Secondary, 4 = Higher
TOPOR	Type of Place of Residence	1 =Rural, 2 = Urban
SES	Socio-economic Status	1 = Poor, 2= Middle, 3= Rich
PO	Partner Occupation	0 = Unemployed, 1 = Employed
WAM	Work After Marriage	0 = No, 1 = Yes
CM	Contraceptive Method Usage	1 =Not Used, 2= Not Used
RAAFB	Respondent Age at First Birth	Numeric values (e.g., 12, 42)
AAM	Age at Marriage	Numeric values (e.g., 9, 38)

"TOPOR" represents the Type of Place of Residence, distinguishing between urban (coded as 2) and rural (coded as 1) settings. This differentiation is significant for understanding the influence of residential environment on the studied outcomes.

"SES," or Socio-economic Status, is categorized into three levels: 1 for poor, 2 for middle, and 3 for rich. This classification assesses the impact of socio-economic factors on various behaviors and outcomes.

"PO," or Partner Occupation, is a binary variable indicating whether the respondent's partner is unemployed (coded as 0) or employed (coded as 1). This variable helps in examining the economic contribution of the partner.

"WAM" stands for Work After Marriage, with 0 indicating no and 1 indicating yes. This variable investigates the respondent's employment status

post-marriage.

Use of Contraceptive:Indicates whether the women use contraceptives or family planning methods, providing insights into the family planning practices that might delay or space childbirth.

"CM," or Contraceptive Method Usage, is also a binary variable with 1 and 2 both indicating not used. There seems to be a discrepancy here that might require correction for clarity.

These explanatory variables were carefully selected to explore the determinant factors influencing the age at which women give birth for the first time[62, 63]. The study aims to analyze the relationships and patterns among these variables to understand the complex dynamics of early age fertility in the context of Pakistan.The coding scheme for the response variable allows for survival analysis techniques like the Cox proportional hazards model, Random Survival Forest, and Conditional Inference Forest to be applied effectively, providing a comprehensive understanding of the factors affecting early age fertility in women.

3.3 Data Preprocessing

Before conducting an analysis of early age fertility among women in Pakistan, a thorough data preprocessing procedure was performed to guarantee the dataset's quality and appropriateness for analytical purposes. The following series of steps outlines this process.

Data Cleaning: In the initial step, the focus was on identifying and addressing outliers within the "age at first birth" variable. These outliers, which represented unrealistic or erroneous data points, were subsequently removed to enhance the cleanliness of the dataset. Following this, the dataset underwent a meticulous examination to identify entries that were either invalid or implausible. Particular attention was given to instances of negative ages or ages surpassing predefined thresholds. Corrective measures were taken to rectify or eliminate such entries.

Data Transformation: In the data transformation phase, categorical variables such as Region, Work After Marriage, and Type of Place of Residence were transformed into numerical representations. This transformation involved assigning numerical codes or creating dummy variables to encode categorical data. For instance, Region was encoded with numeric values corresponding to different geographic regions within Pakistan. Additionally, age-related variables such as Age at First birth and Age at First marriage were derived from existing survey responses. These variables were computed to provide standardized numeric representations suitable for analytical techniques like survival analysis. This transformation was essential to ensure compatibility with machine learning algorithms and to include critical predictive factors in the analysis. These variables were important predictors for the analysis.

Handling Missing Data: Addressing missing data was a crucial step in preparing the dataset for analysis. Missing values were identified across key variables such as women Age at First birth, Age at marriage and Contraceptive Usage. Various imputation techniques were applied to handle missing data effectively. Mean imputation was utilized for numerical variables, where missing values were replaced with the mean of observed values for that variable. This approach helped to complete the dataset and render it suitable for analysis.

Data Splitting Upon completing the data cleaning, transformation, and handling of missing data, the dataset was split into two distinct subsets: a training set and a testing set. The training set comprised 70% of the total dataset, while the testing set comprised the remaining 30%.

Training Set The training set was utilized to train predictive models, including the Cox proportional hazards model, Random Survival Forest, and Conditional Inference Forest. Training models on a substantial portion of the data allowed for capturing underlying patterns and relationships within the dataset.

Testing Set The testing set served as an independent dataset used to eval-

uate the performance and generalizability of the trained models. By assessing model performance on unseen data, the testing set provided insights into how well the models could predict early age fertility outcomes in new observations.

By conducting these data preprocessing steps, the dataset was prepared for analysis using Cox proportional hazards, Random Survival Forest, and Conditional Inference Forest models. Meticulous execution of these steps ensured data accuracy, completeness, and suitability for the subsequent analysis of early age fertility among women in Pakistan.

3.4 Study Design

The diagram depicts an overall study design workflow for a predictive modeling task. It begins with the dataset, which undergoes preprocessing to read data, solve categorical data issues, and handle missing data. After preprocessing, the data is split into a training dataset with 10,469 samples and a testing dataset with 4,486 samples. The training dataset is used to train various models, including Cox, RSF (Random Survival Forest), and CIF (Cumulative Incidence Function). The trained model then makes predictions and evaluates the importance of each feature in determining the outcome, which classifies whether a subject will face early fertility or not. The final outcomes are used to make predictions and assess feature importance.

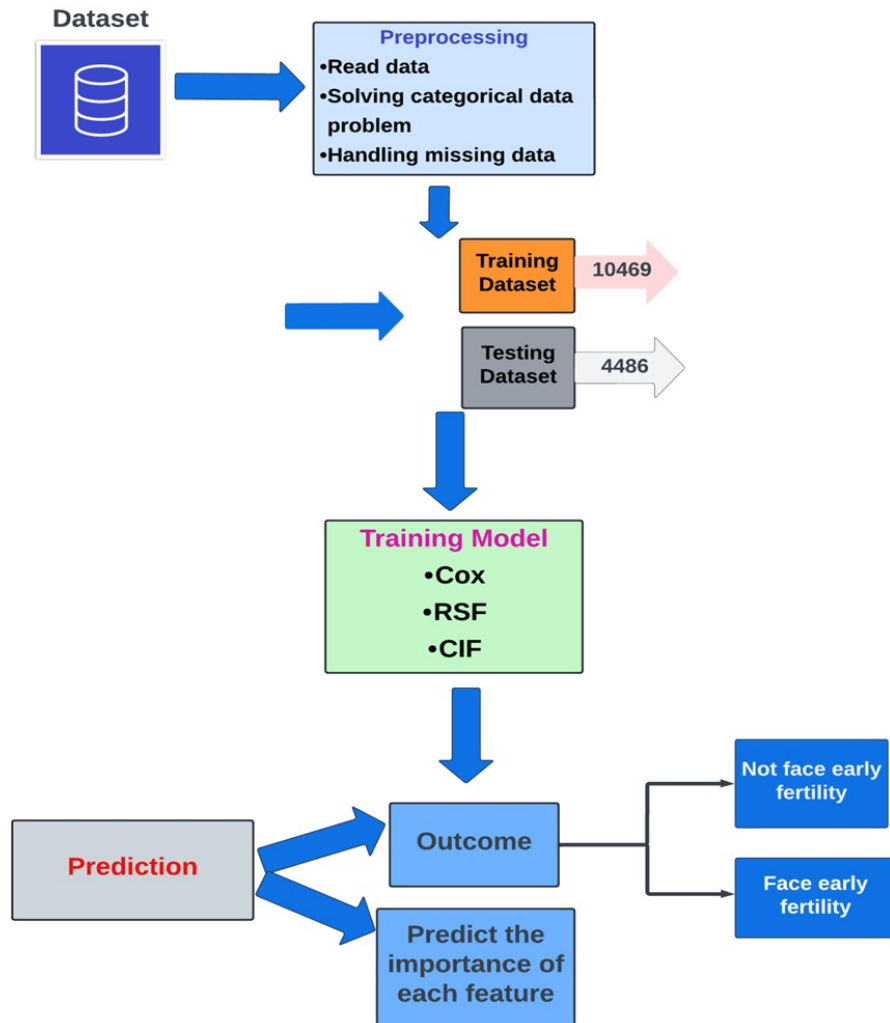


Figure 3.1: Overall Study Design

3.5 Cox Proportional Hazards Model

The Cox proportional hazards model, introduced by David R. Cox in 1972, is a valuable statistical technique for investigating the relationship between survival time[64, 65] (in this case, the age at marriage and the age at first birth for women) and a set of predictor variables[66].It is especially useful when dealing with time-to-event data and allows for the inclusion of censored

observations, making it an appropriate choice for studying early age fertility in women. In this study, the Cox model used to explore the factors influencing women's age at marriage and age at first birth [67]. The model assumes that the hazard function (the instantaneous rate of marriage or first birth) for an individual at any given time is a product of an unknown baseline hazard function and an exponential function of predictor variables. Mathematically, the model can be expressed as: For women's age at marriage:

$$h_1(t, \mathbf{X}) = h_0(t) \times \exp(\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p)$$

For women's age at first birth:

$$h_2(t, \mathbf{X}) = h_0(t) \times \exp(\gamma_1 X_1 + \gamma_2 X_2 + \dots + \gamma_q X_q)$$

$h_1(t, \mathbf{X})$ and $h_2(t, \mathbf{X})$ are the hazard functions for age at marriage and age at first birth, respectively. $h_0(t)$ represents the unspecified baseline hazard function. X_1, X_2, \dots, X_p and X_1, X_2, \dots, X_q are predictor variables for age at marriage and age at first birth, respectively. $\beta_1, \beta_2, \dots, \beta_p$ and $\gamma_1, \gamma_2, \dots, \gamma_q$ are the corresponding coefficients indicating the effect of predictor variables on the hazards. In this study, the relevant predictor variables were selected carefully such as socioeconomic factors, education, cultural norms, and geographical location, which are hypothesized to influence the age at marriage and age at first birth among women. The Cox model was applied to analyze the impact of these variables while accounting for censored data, providing valuable insights into the determinants of early age fertility in women.

3.6 Random Survival Forest (RSF)

The Random Survival Forests (RSF) method extends Breiman's Random Forest (RF) technique to handle right-censored time-to-event data [52]. In the context of the original dataset comprising N subjects and M features, the RSF algorithm operates as follows:

Bootstrapped Sampling: By drawing B bootstrap samples from the original data, each of size N . Bootstrapping involves random sampling with replacement, generating new training sets[68]. Each subject in the original data has a probability of $(1 - \frac{1}{N})^N$ of not being selected for a specific bootstrap sample, creating out-of-bag (OOB) data, which constitutes approximately 30 percent of the original data on average[69].

Binary Survival Tree Construction: For each bootstrapped sample, grow a binary survival tree. At each node of the tree, randomly select a subset of m (where $m \ll M$) features for splitting. Common choices for m include $m = \sqrt{M}$ or $m = \log_2 M$, ensuring a diverse selection of features. A split is made using the candidate feature and its cut-off point that maximizes the survival differences between daughter nodes under a predetermined split rule[70].

Calculation of Cumulate Hazard Function (CHF) and Survival Function (SF): Grow each tree to its full size under pre-specified constraints. Calculate a cumulate hazard function (CHF) and a survival function (SF) for each tree. These functions represent the probability of an event occurring and survival probability, respectively, for each individual in the dataset. Average over all trees to obtain the ensemble CHF, providing one estimate for each individual in the data.

Researchers have developed various splitting rules for Random Survival Forests (RSF), with four notable ones highlighted in reference[52]. These include the log-rank splitting rule, which divides nodes by maximizing the log-rank test statistic, and the log-rank score splitting rule, which splits nodes by maximizing a standardized log-rank score statistic. Additionally, there is the conservation-of-events splitting rule, which separates nodes by identifying daughters closest to the conservation-of-events principle, and the random log-rank splitting rule, where nodes are split based on the variable with the highest log-rank statistic at a predetermined random split point. Among these, the log-rank splitting rule and the log-rank score splitting rule are widely used in practical applications, indicating their popularity

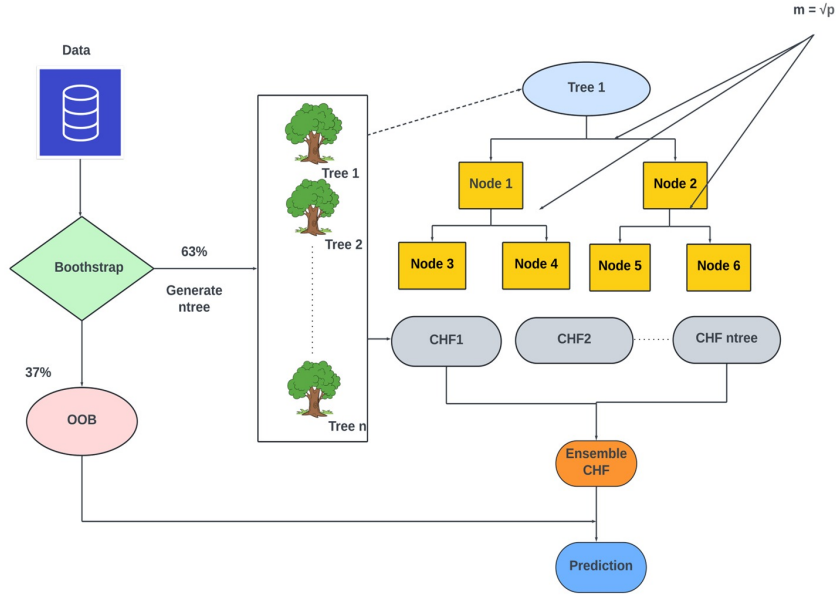


Figure 3.2: Working Algorithm of Random Survival Forest: The RSF algorithm uses bootstrap sampling to create multiple trees, calculates the CHF for each tree, and combines them to make robust survival predictions.

and effectiveness in the field of RSF.

Splitting Rules in RSF (Log-Rank Splitting Rule): The log-rank splitting rule is a representative method used in RSF. For a parent node splitting at the cut-off point value c for predictor X_j , the log-rank test statistic is computed. The log-rank test statistic for a parent node splitting at the cut-off point value c for predictor X_j is calculated as follows:

$$L(X_j, c) = \frac{\sum_{k=1}^K d_{k;l} - Y_{k;l}}{d_k} \times \prod_{k=1}^K \frac{Y_k}{Y_{k;l}} \times \sqrt{\frac{\sum_{k=1}^K Y_{k;l}}{Y_k \times (Y_k - d_k)}} \quad (3.1)$$

Where: $t_1 < t_2 < \dots < t_K$ are the distinct death times in the parent node, d_k and Y_k equal the number of deaths and individuals at risk at time t_k in the parent node respectively,

$$Y_k = Y_{k,l} + Y_{k,r}, \quad d_k = d_{k,l} + d_{k,r},$$

$Y_{k,l}$ represents those in the left daughter node ($Y_{k,l} = \{i : t_i \geq t_k, X_{ji} \leq c\}$). The log-rank statistic is employed to identify the predictor X_j^* and split value c^* with the maximum statistic value, determining the best split for

the Random Survival Forest (RSF) tree. RSF, outlined in reference, offers distinct advantages in the realm of genetic research. It stands out due to its non-parametric nature [71], flexibility, and ability to seamlessly handle high-dimensional co-variate data, essential elements in genetics studies. Notably, RSF excels in adaptability and freedom from stringent model assumptions, making it particularly valuable when dealing with intricate relationships between predictors and outcomes, such as nonlinear effects or high-order interactions. Additionally, RSF provides Variable Importance Measures (VIM) and Out-Of-Bag (OOB) estimates, enhancing its analytical depth. The implementation of RSF is facilitated through various packages, with the current study utilizing the `randomForestSRC` R package [72], which incorporates the effective Log-rank splitting rule, emphasizing its versatility and applicability in advanced genetic analyses.

The Random Survival Forest (RSF) algorithm, derived from bootstrapped sampling, binary survival tree construction, and specific splitting rules like the log-rank rule, provides a comprehensive and robust approach to analyzing right-censored time-to-event data. By integrating these steps, RSF offers a sophisticated method for understanding complex survival patterns, making it invaluable in fields where predicting time-to-event outcomes is critical, such as in the study of early-age fertility among women.

3.7 Conditional Inference forest

Conditional inference forests (CIF) method is a tree ensemble technique grounded in the principles of permutation tests [73]. Unlike traditional methods like CART, which can exhibit a bias towards variables with numerous split points, CIF addresses this issue by incorporating statistical significance considerations. In algorithms such as CART, the bias towards variables with many split points arises from maximizing a splitting criterion over all potential splits. This bias occurs because variables with more split points have a higher chance of finding a good split, even if the variable is uninformative. Consequently, uninformative variables can end up high in the

tree’s structure, leading to biased estimates[74].CIF mitigates this problem by factoring in the statistical significance, ensuring a more robust variable selection process and reducing the risk of biased estimates[75]. CIF builds forests using Conditional Inference Trees (CIT) as base learners[73], following a distinct approach from traditional methods. Instead of exploring all possible splits, CIT divides the process into two steps: first, it identifies the best split covariate through association tests, and then it performs the optimal binary split based on standardized linear statistics. In the context of the Conditional Inference Tree (CIT) splitting procedure, given a dataset consisting of N subjects and M features, and defining a new training set $L_n = \{\delta_i, X_{1i}(t), \dots, X_{Mi}(t)\}$ for $i = 1, \dots, n$, the first step involves determining if there is pertinent information about the response variable (δ_i) contained within a specific covariate ($X_j(t)$). This determination is made based on the partial hypothesis of independence $H_{j0} : D(\delta_i, X_j(t)) = D(\delta_i)$, with the overarching null hypothesis $H_0 = \sum_{j=1}^M H_{j0}$.

The strength of association between δ_i and $X_j(t)$ is quantified using linear statistics given by

$$T_j = \sum_{i=1}^n w_i g_j(X_{ji}(t)) h(T_i, \delta_i)$$

where w signifies case weights at each node, g_j denotes a non-random transformation of $X_{ji}(t)$, and h depends on responses $\delta_1, \dots, \delta_n$ in a symmetric permutation manner. Practical considerations may dictate variations in these functions; for instance, in time-to-event data, the choice of the influence function could be the log-rank score or Savage score. The assessment of T_j is grounded in the distribution of δ_i and X_{ji} , which is often unknown. However, under the null hypothesis, this dependency can be nullified by fixing the covariates and considering all possible permutations of responses, a methodology known as the theory of permutation tests. Subsequently, $T_j(t)$ is standardized to univariate test statistics $uT_j L_n; w$ for further comparison. If H_0 cannot be rejected at a predetermined significance level α , the recursion ceases. Alternatively, if rejected, X_j^* with the most compelling association (i.e., the smallest P-value) is chosen as the optimal split variable.

Once a covariate X_{j^*} is selected in step 1 of the algorithm, the subsequent step involves determining the optimal split point. The quality of the split is assessed using a two-sample linear statistic, a specialized form of the linear statistic employed in step 1. For all potential split points of X_{j^*} , the linear statistic is defined as

$$T_{c_{j^*}} = \sum_{i=1}^n w_i I(X_{j^*i} \geq c^*) h(T_i, \delta_i)$$

where

$$I(X_{i,j} \geq c^*) = \begin{cases} 1 & \text{if } X_{i,j} \geq c^* \\ 0 & \text{otherwise} \end{cases}$$

This two-sample statistic quantifies the difference between the two resulting daughter nodes. The split c^* with a standardized test statistic $uT_{c_{j^*}}L_n; w$ maximized across all feasible splits is established.

CIF distinguishes itself from RF and RSF not only in terms of the base learner but also in the applied aggregation scheme. Unlike RF, where predictions are directly averaged, CIF utilizes a unique approach. It averages observation weights extracted from each tree, providing a distinct method of aggregation. This specific aggregation technique sets CIF apart. The implementation of CIF can be found in the R package called "party"[76].

3.8 Model Evaluation Metrics

The **C-index (Concordance Index)** and the **Integrated Brier Score (IBS)** are two widely used metrics for evaluating the performance of survival models, each offering unique insights into different aspects of model performance.

C-index:

The cindex is a measure of the discriminatory power of a survival model, quantifying how well the model can distinguish between individuals with different event times. It is a generalization of the area under the receiver operating characteristic (ROC) curve (AUC) for binary outcomes to time-to-event outcomes. The C-index is calculated by considering all possible

pairs of individuals and checking if the predicted risk scores correctly rank the survival times. Specifically, for a pair of individuals, the model's prediction is considered concordant if the individual with the higher risk score has a shorter observed survival time. The C-index ranges from 0.5 to 1.0, where 0.5 indicates no discriminatory ability (equivalent to random guessing), and 1.0 indicates perfect discrimination. A higher C-index value signifies better predictive accuracy of the model. Mathematically, the C-index can be expressed as:

$$C = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^n I(h_i < h_j) \cdot I(t_i < t_j) \cdot \delta_i}{\sum_{i=1}^{n-1} \sum_{j=i+1}^n I(t_i < t_j) \cdot \delta_i}$$

where h_i and h_j are the predicted risk scores, t_i and t_j are the observed survival times, δ_i is the event indicator (1 if the event occurred, 0 if censored), and $I(\cdot)$ is the indicator function that returns 1 if the condition inside is true and 0 otherwise.

Integrated Brier Score (IBS)

The Integrated Brier Score (IBS), on the other hand, measures the accuracy of probabilistic predictions in survival analysis over time. The Brier score at a specific time point t evaluates the mean squared difference between the observed outcomes and the predicted survival probabilities. It incorporates both calibration (how well the predicted probabilities reflect the actual outcomes) and discrimination (the ability to distinguish between different outcomes). The IBS is obtained by integrating the Brier score over the entire follow-up period, providing a single summary measure of model performance. The Brier score at time t is defined as:

$$\text{Brier}(t) = \frac{1}{n} \sum_{i=1}^n \left(\hat{S}(t|X_i) - \delta_i \right)^2$$

where n is the number of subjects, δ_i is the event indicator at time t (1 if the event occurred by time t , 0 otherwise), and $\hat{S}(t|X_i)$ is the predicted survival probability at time t for subject i . The IBS then integrates this score over the follow-up period τ .

$$IBS = \int_0^{\tau} \text{Brier}(t) dt$$

Lower values of the IBS indicate better model performance, as they signify smaller deviations between predicted probabilities and actual outcomes.

In summary, the C-index provides a measure of how well the model can rank individuals by their risk of experiencing the event, while the IBS evaluates the overall accuracy of the predicted survival probabilities over time. Together, these metrics offer a comprehensive evaluation of a survival model's performance.

Chapter 4

Results and Discussions

4.1 Descriptive Results

In the context of early-age fertility among Pakistani women, the descriptive statistics for variables such as Respondent Age at First Birth (RAAFB) and Age at Marriage (AAM) provide crucial insights ((see Table 4.1)). The minimum RAAFB of 12 and AAM of 9 indicate that some women experience childbirth and marriage at extremely young ages, highlighting the severity of early-age fertility. Conversely, the maximum RAAFB of 42 and AAM of 38 suggest that women in the study have delayed childbirth and marriage to relatively older ages, showcasing the variability in reproductive timelines.

Table 4.1: Summary Statistics for RAAFB and AAM

Statistic	RAAFB	AAM
Mean	21.11	18.85
Median	21	18
Standard Deviation	3.91	3.75
Min	12	9
Max	42	38

The mean age at first birth (21.11 years) suggests that, on average, women in the study give birth relatively early after marriage. This aligns with the objective of understanding early-age fertility patterns. The median values of 21 for RAAFB and 18 for AAM indicate that the distribution is slightly skewed towards younger ages, highlighting a significant portion of women

experiencing early childbirth. The standard deviations of 3.91 for RAAFB and 3.75 for AAM signify a moderate amount of variation around the mean, indicating diverse age patterns.

The frequency table 4.2 provides a detailed overview of the distribution of various variables related to early-age fertility among Pakistani women. The "Status" variable, indicating whether women are experiencing early-age fertility, shows that 43.40% of women are currently experiencing early-age fertility, while 56.60% are not. This sheds light on the prevalence of early-age fertility in the study population.

Table 4.2: Frequency and Percentage Frequency for Various Variables

Variable	Category	Frequency	Percentage Frequency
Status	Yes	6491	43.40
	No	8464	56.60
REL	No education	7545	50.45
	Primary	2091	13.98
	Secondary	3118	20.85
	Higher	2201	14.72
HEL	No education	415	2.77
	Primary	3467	23.18
	Secondary	10136	67.78
	Higher	937	6.27
REG	Punjab	3392	22.68
	Sindh	2718	18.17
	KPK	2370	15.85
	Balochistan	1694	11.33
	GB	971	6.49
	ICT	1100	7.36
	AJK	1715	11.47
	FATA	995	6.65
TOPOR	Urban	7199	48.14
	Rural	7756	51.86
SES	Poor	6066	40.56
	Middle	5803	38.80

Variable	Category	Frequency	Percentage Frequency
	Rich	3086	20.64
PO	Unemployed	803	5.37
	Employed	14133	94.50
	Don't know	19	0.13
WAM	No	12624	84.41
	Yes	2331	15.59
CM	Used	4709	31.49
	Not Used	10246	68.51

Regarding education levels, a significant portion of women have no education (50.45%), followed by secondary education (20.85%). This highlights the educational disparities among women in the sample and the potential impact of education on early-age fertility decisions. Husband education level ("HEL") also shows a considerable proportion with secondary education (67.78%), suggesting a similar trend in educational attainment between spouses. Regional distribution ("REG") indicates that Punjab has the highest representation (22.68%), followed by Sindh (18.17%) and Khyber-Pakhtunkhwa (15.85%), which could reflect regional variations in early-age fertility determinants. The type of place of residence ("TOPOR") shows a relatively balanced distribution between urban (48.14%) and rural (51.86%) areas, indicating the need to consider both urban and rural contexts in understanding early-age fertility.

Socio-economic status ("SES") reveals that a substantial proportion of women are classified as poor (40.56%) or middle-class (38.80%), underscoring the economic factors at play in early-age fertility decisions. Partner occupation ("PO") highlights that the majority are employed (94.50%), suggesting that economic factors might influence fertility decisions. Work after marriage ("WAM") indicates that 15.59% of women work after marriage, which could impact their fertility decisions due to potential financial stability. Contraceptive method usage ("CM") shows that 31.49% of women use contraceptives, implying a level of family planning awareness and practices among the population. Overall, these findings provide a comprehensive

picture of the socio-demographic factors associated with early-age fertility among Pakistani women, highlighting the complex interplay of education, region, socio-economic status, and other factors in shaping fertility decisions.

4.2 Cox Model

4.2.1 Age at Marriage as a Time Variable

In this study, we examined the impact of various socio-economic and demographic factors on the likelihood of facing early age fertility among Pakistani women, defined as having the first birth at the age of 20 or before, using a Cox proportional hazards model. The event of interest is facing early age fertility, coded as 1 for those who experience it and 0 for those who do not, with the response variable (t) being the age at marriage. The results, as shown in Table 4.3, reveal several significant findings. The forest plot in Figure 4.1 visually represents the hazard ratios ($\exp(\text{coef})$) for each variable, providing a clear comparison of the impact of different factors on age at marriage.

Respondent Education Level (REL): Education level had significant effects on the likelihood of facing early age fertility. Respondents with primary education ($\text{coef} = -0.124783$, $\exp(\text{coef}) = 0.882688$, $p = 0.00547$) had a hazard ratio of 0.8827, indicating they were approximately 11.73% less likely to face early age fertility compared to those without education. Those with secondary education ($\text{coef} = -0.554170$, $\exp(\text{coef}) = 0.574549$, $p < 2e-16$) were significantly less likely to face early age fertility, with a hazard ratio of 0.5745, suggesting a 42.55% reduced likelihood. Higher education had an even stronger protective effect ($\text{coef} = -1.582577$, $\exp(\text{coef}) = 0.205445$, $p < 2e-16$), with a hazard ratio of 0.2054, indicating an approximately 79.46% reduced likelihood of facing early age fertility. This highlights the substantial protective effect of education on early age fertility.

Husband Education Level (HEL): Husband's education level did not show significant effects on the likelihood of facing early age fertility. The coeffi-

coefficients for primary (coef = 0.005802, exp(coef) = 1.005819, p = 0.94985), secondary (coef = -0.013995, exp(coef) = 0.986102, p = 0.87478), and higher education (coef = -0.236813, exp(coef) = 0.789139, p = 0.05138) were not statistically significant, indicating that husband’s education level did not significantly influence the likelihood of facing early age fertility.

Table 4.3: Cox Model Results

Variable	Factor	coef	exp(coef)	exp(-coef)	se(coef)	z	Pr(> z)	lower .95	upper .95
REL	Primary	-0.124783	0.882688	1.1329	0.044917	-2.778	0.00547 **	0.8083	0.9639
	Secondary	-0.554170	0.574549	1.7405	0.046008	-12.045	< 2e-16 ***	0.5250	0.6288
	Higher	-1.582577	0.205445	4.8675	0.077947	-20.303	< 2e-16 ***	0.1763	0.2394
HEL	Primary	0.005802	1.005819	0.9942	0.092243	0.063	0.94985	0.8395	1.2051
	Secondary	-0.013995	0.986102	1.0141	0.088810	-0.158	0.87478	0.8286	1.1736
	Higher	-0.236813	0.789139	1.2672	0.121549	-1.948	0.05138 .	0.6219	1.0014
REG	Sindh	0.224093	1.251188	0.7992	0.049744	4.505	6.64e-06 ***	1.1350	1.3793
	KPK	0.473395	1.605436	0.6229	0.049731	9.519	< 2e-16 ***	1.4563	1.7698
	Balochistan	0.330136	1.391158	0.6229	0.057400	5.752	8.84e-09 ***	1.2431	1.5568
	GB	0.495365	1.641097	0.6093	0.067403	7.349	1.99e-13 ***	1.4380	1.8729
	ICT	0.013404	1.013495	0.6093	0.074898	0.179	0.85796	0.8751	1.1737
	AJK	0.079390	1.082627	0.9237	0.060767	1.306	0.19139	0.9611	1.2196
FATA	FATA	0.526958	1.693773	0.5904	0.063932	8.242	< 2e-16 ***	1.4943	1.9199
	Urban	-0.116057	0.890424	1.1231	0.034890	-3.326	0.00088 ***	0.8316	0.9534
SES	Middle	-0.181893	0.833691	1.1995	0.038831	-4.684	2.81e-06 ***	0.7726	0.8996
	Rich	-0.206282	0.813603	1.2291	0.058054	-3.553	0.00038 ***	0.7261	0.9117
PO	Employed	-0.033113	1.0337	0.967430	0.064516	-0.513	0.60778	0.8525	1.0978
	Don't know	-0.130446	0.877704	1.1393	0.504355	-0.259	0.79591	0.3266	2.3586
WAM	Yes	0.164294	1.178561	0.8485	0.043871	3.745	0.00018 ***	1.0815	1.2844
CM	Not Used	-0.399511	0.670648	1.4911	0.032138	-12.431	< 2e-16 ***	0.6297	0.7143

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Region (REG): Regional differences were significant. Compared to the reference category (Punjab), respondents from Sindh (coef = 0.224093, exp(coef) = 1.251188, p = 6.64e-06), KPK (coef = 0.473395, exp(coef) = 1.605436, p < 2e-16), Balochistan (coef = 0.330136, exp(coef) = 1.391158, p = 8.84e-09), GB (coef = 0.495365, exp(coef) = 1.641097, p = 1.99e-13), and FATA (coef = 0.526958, exp(coef) = 1.693773, p < 2e-16) had significantly higher hazard ratios, suggesting higher risks of facing early age fertility. Specifically, individuals from Sindh were approximately 25.12% more likely, those from KPK were about 60.54% more likely, those from Balochistan were approximately 39.12% more likely, those from GB were about 64.11% more likely, and those from FATA were about 69.38% more likely to face early age

fertility compared to those from Punjab. ICT (coef = 0.013404, exp(coef) = 1.013495, p = 0.85796) and AJK (coef = 0.079390, exp(coef) = 1.082627, p = 0.19139) did not show significant differences from Punjab.

Type of Place of Residence (TOPOR): Living in a urban area was associated with a reduced risk of facing early age fertility (coef = -0.116057, exp(coef) = 0.890424, p = 0.00088), with a hazard ratio of 0.8904. This indicates that individuals living in urban areas were approximately 10.96% less likely to face early age fertility compared to those living in rural areas.

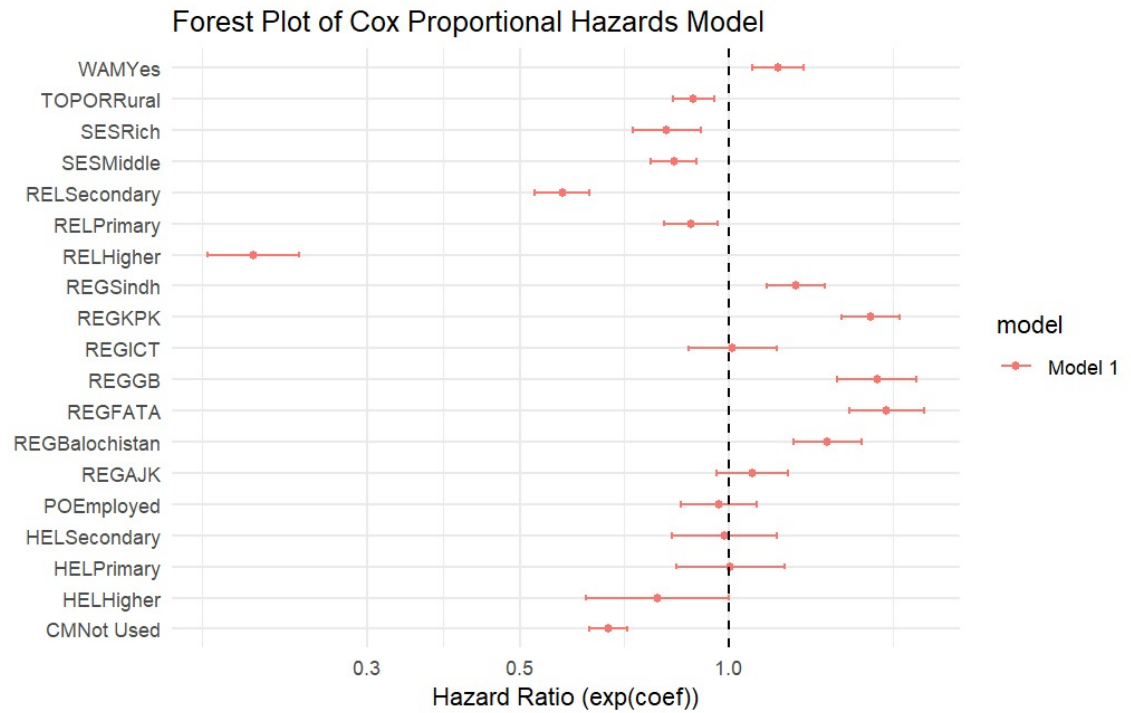


Figure 4.1: Forest Plot of Cox Proportional Hazards Model

Socio-economic Status (SES): Socio-economic status was a significant factor. Respondents from middle socio-economic status (coef = -0.181893, exp(coef) = 0.833691, p = 2.81e-06) were approximately 16.63% less likely, and those from rich status (coef = -0.206282, exp(coef) = 0.813603, p = 0.00038) were about 18.64% less likely to face early age fertility compared to those from poorer backgrounds. This highlights the protective effect of higher socio-economic status against early age fertility.

Partner Occupation (PO): Partner occupation was not a significant factor. The coefficients for employed partners (coef = -0.033113, exp(coef) = 0.967430, p = 0.60778) and those who don't know (coef = -0.130446, exp(coef) = 0.877704, p = 0.79591) were not statistically significant, indicating that partner occupation did not significantly influence the likelihood of facing early age fertility.

Work After Marriage (WAM): Respondents who worked after marriage had a significantly higher risk of facing early age fertility (coef = 0.164294, exp(coef) = 1.178561, p = 0.00018), with a hazard ratio of 1.1786. This suggests that individuals who worked after marriage were approximately 17.86% more likely to face early age fertility compared to those who did not work after marriage.

Contraceptive Method Usage (CM): Not using contraceptives was associated with a significantly higher risk of facing early age fertility (coef = -0.399511, exp(coef) = 0.670648, p < 2e-16), with a hazard ratio of 0.6706. This indicates that individuals who did not use contraceptives were approximately 32.94% less likely to face early age fertility compared to those who used contraceptives, highlighting the protective effect of contraceptive use.

In summary, the results indicate that higher levels of respondent education and better socio-economic status are associated with lower risks of facing early age fertility. Regional differences also play a significant role, with certain regions having higher risks. Living in rural areas and using contraceptives appear to provide protective effects, while working after marriage increases the risk. These findings underscore the importance of education, socio-economic status, regional context, and contraceptive use in understanding the factors influencing the likelihood of facing early age fertility.

4.2.2 Respondent Age at First Birth as a Time Variable

In this analysis, we explored the factors influencing the likelihood of facing early fertility, which is defined as giving birth for the first time at the age of 20 or younger. The response variable in our Cox proportional hazards model

is the age at which women experience their first birth, with a focus on identifying which socio-economic and demographic factors are associated with a higher or lower probability of early fertility. Table 4.3 presents the comprehensive statistical analysis results, highlighting several significant findings. The forest plot in Figure 4.1 visually represents the hazard ratios ($\exp(\text{coef})$) for each variable.

Age at Marriage (AAM): The age at which women marry is a significant predictor of early fertility. The negative coefficient for age at marriage ($\text{coef} = -0.508308$, $\exp(\text{coef}) = 0.601513$, $p < 2e-16$) indicates that for each additional year of age at marriage, the risk of experiencing early fertility decreases by approximately 39.85%. In other words, women who marry later are less likely to face early fertility. This substantial effect underscores the importance of delaying marriage as a strategy for reducing the likelihood of early childbearing.

Table 4.4: Comprehensive Statistical Analysis Results

Covariate	Factor	coef	exp(coef)	se(coef)	z	Pr(> z)	exp(-coef)	95% CI lower	95% CI upper
AAM		-0.508308	0.601513	0.006703	-75.836	< 2e-16	1.6625	0.5937	0.6095
REL	Primary	0.116390	1.123434	0.045782	2.542	0.011014	0.8901	1.0270	1.2289
	Secondary	0.024004	1.024294	0.047286	0.508	0.611715	0.9763	0.9336	1.1238
	Higher	-0.526924	0.590418	0.078643	-6.700	2.08e-11	1.6937	0.5061	0.6888
HEL	Primary	-0.035820	0.964814	0.092358	-0.388	0.698136	1.0365	0.8051	1.1563
	Secondary	-0.126820	0.880892	0.088974	-1.425	0.154053	1.1352	0.7399	1.0487
	Higher	-0.188237	0.828418	0.121715	-1.547	0.121975	1.2071	0.6526	1.0516
REG	Sindh	-0.129466	0.878564	0.050389	-2.569	0.010189	1.1382	0.7959	0.9698
	KPK	0.142208	1.152817	0.049948	2.847	0.004411	0.8674	1.0453	1.2714
	Balochistan	-0.016511	0.983625	0.058111	-0.284	0.776314	1.0166	0.8777	1.1023
	GGB	-0.269385	0.763849	0.069322	-3.886	0.000102	1.3092	0.6668	0.8750
	ICT	-0.018941	0.981237	0.074699	-0.254	0.799834	1.0191	0.8476	1.1359
	AJK	0.075215	1.078115	0.061050	1.232	0.217942	0.9275	0.9565	1.2152
	FATA	0.279045	1.321867	0.064172	4.348	1.37e-05	0.7565	1.1656	1.4990
TOPOR	Urban	-0.101925	0.903097	0.034613	-2.945	0.003233	1.1073	0.8439	0.9665
SES	Middle	0.074685	1.077545	0.039073	1.911	0.055951	0.9280	0.9981	1.1633
	Rich	0.083105	1.086656	0.058329	1.425	0.154224	0.9203	0.9693	1.2183
PO	Employed	0.209405	1.232944	0.064617	3.241	0.001192	0.8111	1.0863	1.3994
	Don't know	0.518873	1.680133	0.504627	1.028	0.303841	0.5952	0.6249	4.5173
WAM	Yes	0.043114	1.044057	0.044085	0.978	0.328092	0.9578	0.9576	1.1383
CMU	Not	-0.571616	0.564612	0.032194	-17.755	< 2e-16	1.7711	0.5301	0.6014

Respondent Education Level (REL): Education level plays a crucial role in determining the likelihood of early fertility. Women with primary education ($\text{coef} = 0.116390$, $\exp(\text{coef}) = 1.123434$, $p = 0.011014$) are about 12.34%

more likely to face early fertility compared to those with no education. In contrast, higher education (coef = -0.526924, $\exp(\text{coef}) = 0.590418$, $p = 2.08e-11$) significantly reduces the risk of early fertility by 40.96%. This suggests that higher education serves as a protective factor against early childbearing, highlighting the benefits of educational attainment for family planning.

Husband's Education Level (HEL): The educational attainment of the husband does not show a significant effect on the likelihood of early fertility. Coefficients for primary, secondary, and higher education (HELPrimary: coef = -0.035820, $\exp(\text{coef}) = 0.964814$, $p = 0.698136$; HELSecondary: coef = -0.126820, $\exp(\text{coef}) = 0.880892$, $p = 0.154053$; HELHigher: coef = -0.188237, $\exp(\text{coef}) = 0.828418$, $p = 0.121975$) are not statistically significant, suggesting that the husband's education level does not directly influence the timing of the first birth.

Regional Effects (REG): Geographic location reveals significant regional variations in the risk of early fertility. Women from Sindh (coef = -0.129466, $\exp(\text{coef}) = 0.878564$, $p = 0.010189$) are about 12.14% less likely to face early fertility compared to those from Punjab. In contrast, women from Khyber Pakhtunkhwa (KPK) (coef = 0.142208, $\exp(\text{coef}) = 1.152817$, $p = 0.004411$) have a 15.28% higher risk of early fertility, while those from Gilgit-Baltistan (GB) (coef = -0.269385, $\exp(\text{coef}) = 0.763849$, $p = 0.000102$) are 23.62% less likely to experience early fertility. Women from the Federally Administered Tribal Areas (FATA) (coef = 0.279045, $\exp(\text{coef}) = 1.321867$, $p = 1.37e-05$) face a 32.19% higher risk of early fertility. These regional differences highlight the influence of local cultural, economic, and healthcare factors on the likelihood of early childbearing.

Type of Place of Residence (TOPOR): Living in urban areas is associated with a lower risk of early fertility (coef = -0.101925, $\exp(\text{coef}) = 0.903097$, $p = 0.003233$), indicating that women in urban areas are about 9.70% less likely to face early fertility compared to those in rural areas. This finding reflects differences in access to family planning resources and educational

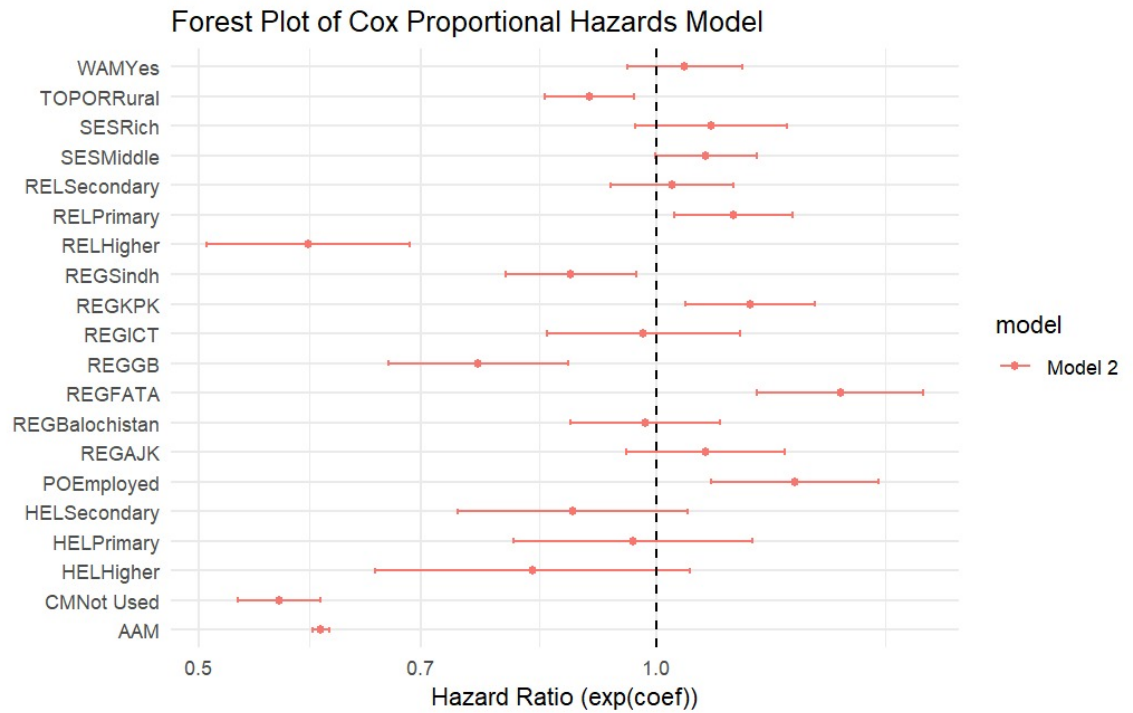


Figure 4.2: Forest Plot of Cox Proportional Hazards Model

opportunities between rural and urban settings.

Socio-economic Status (SES): Socio-economic status shows a near-significant effect on early fertility. Women from middle socio-economic backgrounds (coef = 0.074685, exp(coef) = 1.077545, p = 0.055951) have a 7.75% higher risk of early fertility, though this result is just above the conventional significance level. Women from rich backgrounds did not show a significant difference in the risk of early fertility compared to those from poorer backgrounds.

Partner Occupation (PO): Women whose partners are employed have a significantly higher risk of early fertility (coef = 0.209405, exp(coef) = 1.232944, p = 0.001192). The likelihood of facing early fertility increases by approximately 23.29% for women with employed partners. This might be due to increased financial stability and reduced pressures that might otherwise delay family planning.

Work After Marriage (WAM): Working after marriage does not have a sig-

nificant impact on the likelihood of early fertility (coef = 0.043114, exp(coef) = 1.044057, p = 0.328092), indicating that this factor alone does not substantially influence the timing of the first birth.

Contraceptive Method Usage (CM): Not using contraceptives is a strong predictor of early fertility (coef = -0.571616, exp(coef) = 0.564612, p < 2e-16). Women who do not use contraceptives are about 43.54% more likely to experience early fertility compared to those who do use contraceptives. This result emphasizes the critical role of contraceptive use in family planning and the prevention of early childbearing.

In summary, our analysis reveals that later marriage and higher levels of education are effective strategies for reducing the risk of early fertility. Specifically, every additional year of age at marriage decreases the likelihood of early fertility by nearly 40%. Women with higher educational attainment are significantly less likely to experience early fertility, with higher education reducing this risk by nearly 41%. In contrast, primary education increases the risk of early fertility by about 12%, while husband's education does not significantly affect early fertility. Regional variations also play a significant role, with women in certain regions like KPK and FATA facing higher risks of early fertility, while those in Sindh and GB have lower risks. Additionally, women in urban areas and those using contraceptives have lower risks of early fertility, whereas those with employed partners face a higher risk of early childbearing. These insights underscore the importance of education, family planning, and socio-economic factors in shaping fertility outcomes and highlight areas for targeted interventions to support delayed childbearing and improve reproductive health outcomes.

4.3 Random Survival Forest Model

4.3.1 AAM as a Time Variable

In this study, the Random Survival Forest (RSF) model is employed to analyze early-age fertility, focusing on the relationship between age at marriage

and the likelihood of experiencing early-age fertility, which is defined as having the first birth at the age of 20 or before. The RSF model is particularly suited for this task as it provides a robust framework for handling survival data and exploring complex interactions among multiple predictors. This model is a non-parametric extension of decision trees that excels in capturing non-linear relationships and interactions without requiring specific assumptions about the data distribution, which makes it ideal for examining the multifaceted nature of fertility outcomes.

The graphical analysis of the model reveals insights into both the performance of the model and the significance of various predictors. The plot in Figure 4.3 depicting the error rate versus the number of trees shows a clear pattern where the error rate decreases as the number of trees in the forest increases, reflecting improved model performance with more trees. This decrease continues until the error rate stabilizes, suggesting that adding trees beyond a certain point yields diminishing returns. Initially, increasing the number of trees effectively reduces the error rate, which indicates that the model benefits from greater complexity and a more comprehensive representation of the data. However, after a certain threshold, additional trees do not significantly improve the model's accuracy and may even introduce noise, emphasizing the balance needed between model complexity and performance.

The variable importance plot offers a detailed view of how different predictors contribute to the model's ability to distinguish between women who experience early-age fertility and those who do not. The most striking finding is that the respondent's education level (REL) is the most significant factor in predicting early-age fertility, with a variable importance score of 0.1946 and a relative importance of 1.0000. This indicates that educational attainment is the strongest predictor of whether a woman will experience early-age fertility, highlighting that higher education levels are associated with a lower likelihood of early-age fertility. This result underscores the critical role of education in shaping fertility outcomes, as higher educational

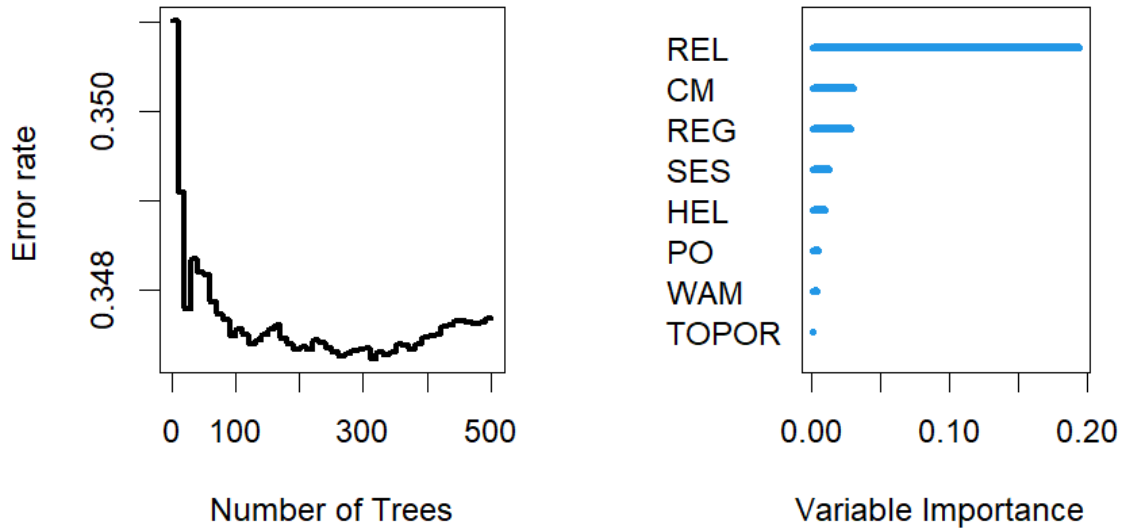


Figure 4.3: Number of tree and variable importance AAM

attainment often correlates with greater access to family planning resources and better awareness of reproductive health.

Following education level, the usage of contraceptive methods (CM) emerges as the second most important predictor, with a variable importance score of 0.0303 and a relative importance of 0.1556. This finding emphasizes that the use of contraceptives significantly impacts early-age fertility, as women who actively use contraceptive methods are less likely to experience early-age fertility. This result suggests that improving access to and education about contraceptives could be an effective strategy for reducing early-age fertility rates.

Regional factors, represented by the variable REG, also play a significant role in the model, with a variable importance score of 0.0286 and a relative importance of 0.1468. This indicates that the geographical location of the respondents affects early-age fertility outcomes, reflecting the influence of regional cultural norms, economic conditions, and availability of healthcare

services on fertility practices. Variations across different regions demonstrate that localized factors can impact fertility rates, which can be addressed through targeted regional policies and interventions.

Socio-economic status (SES), with a variable importance score of 0.0133 and a relative importance of 0.0684, shows a moderate impact on early-age fertility. This suggests that socio-economic conditions, such as wealth and access to resources, contribute to the likelihood of experiencing early-age fertility. Women from poorer socio-economic backgrounds are more likely to face early-age fertility due to fewer resources and less access to educational and healthcare opportunities.

The education level of the husband (HEL), with a variable importance score of 0.0099 and a relative importance of 0.0508, demonstrates a lower but still notable influence on early-age fertility. Although less significant than the respondent's education, the husband's education level does affect fertility outcomes, indicating that family-level educational attainment can influence reproductive decisions.

Partner occupation (PO) and work after marriage (WAM) have relatively minor effects on early-age fertility, with variable importance scores of 0.0057 and 0.0044, and relative importances of 0.0292 and 0.0225, respectively. These variables have limited impact on the likelihood of early-age fertility, suggesting that while the partner's occupation and decisions about working after marriage are factors, they are less critical compared to other variables such as education and contraceptive use.

Finally, the type of place of residence (TOPOR), with a very low variable importance score of 0.0011 and a relative importance of 0.0056, has the smallest impact on early-age fertility. This indicates that while there may be some differences between urban and rural fertility rates, this factor is less significant in the overall model compared to other predictors.

Overall, the RSF model reveals that educational attainment and contraceptive use are the most influential factors in predicting early-age fertility, with educational level being the primary determinant and contraceptive use also

playing a significant role. Regional differences and socio-economic conditions also affect fertility outcomes, but to a lesser extent. The insights gained from the model highlight the importance of focusing on educational interventions and improving access to family planning resources as key strategies for reducing early-age fertility rates.

4.3.2 RAAFB as a Time Variable

Also the Random Survival Forest (RSF) model is utilized to explore the relationship between various predictors and the age at which women experience their first birth, with the primary focus on early-age fertility as the response variable. This model examines how different factors influence the timing of first birth, providing a detailed understanding of the predictors associated with early-age fertility.

The left plot, which depicts the error rate versus the number of trees in the RSF model, reveals that as the number of trees increases, the error rate initially decreases, signifying an improvement in model performance. This decline in error rate indicates that adding more trees enhances the model's ability to make accurate predictions about the age at first birth. However, after reaching a certain number of trees, the error rate stabilizes, indicating that while additional trees initially contribute to model accuracy, they eventually lead to diminishing returns. This observation suggests that there is an optimal number of trees that balances model complexity and accuracy without unnecessary overfitting.

The right plot illustrates the importance of different variables in predicting the age at first birth. The most notable finding from this plot is that the age at marriage (AAM) stands out as the most significant predictor of the age at first birth, with a variable importance score of 0.6008 and a relative importance of 1.0000. This high score indicates a strong relationship between the age at marriage and the age at which women have their first child. Women who marry later are more likely to delay their first birth, highlighting how the timing of marriage directly affects reproductive decisions.

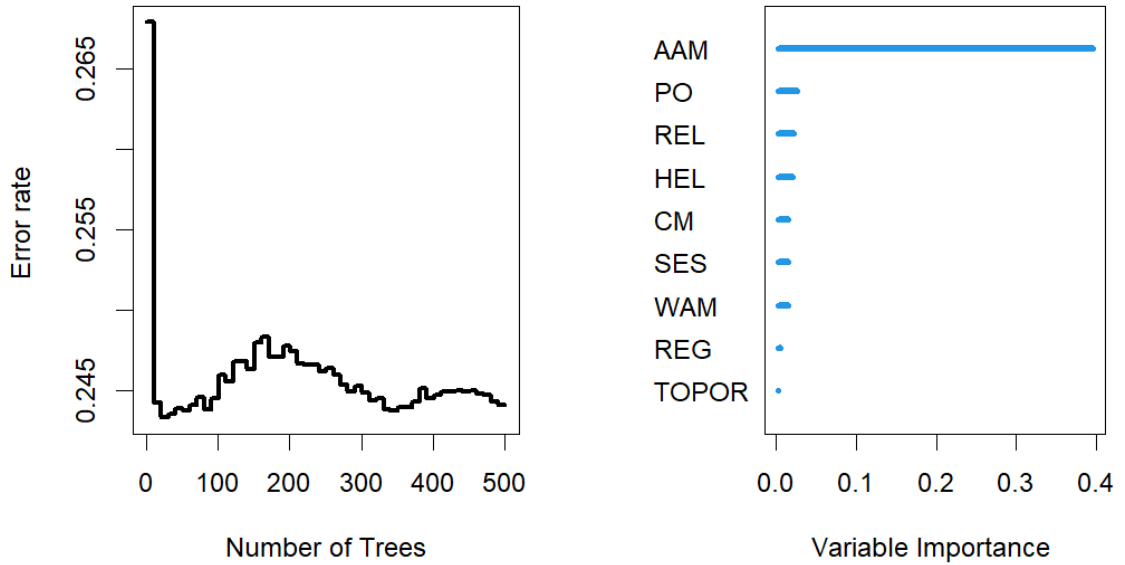


Figure 4.4: Number of tree and variable importance RAAFb

Following AAM, the next most influential predictor is the respondent’s education level (REL), which has a variable importance score of 0.0170 and a relative importance of 0.0282. Although its impact is significantly lower compared to AAM, higher educational attainment is still associated with a later age at first birth. This correlation can be attributed to the fact that women who pursue higher education tend to delay childbearing in favor of academic and career achievements.

Another important predictor is contraceptive method usage (CM), with a variable importance score of 0.0151 and a relative importance of 0.0251. This variable reflects that women who use contraceptives are able to plan their pregnancies more effectively, leading to a later age at first birth. Effective use of contraceptives enables women to space their pregnancies, which often results in delaying the age at which they have their first child.

Other variables such as partner occupation (PO) and husband education level (HEL) show lower importance scores, 0.0043 and 0.0042 respectively,

indicating that their influence on the timing of the first birth is relatively minor. These factors might still play a role in reproductive decisions by reflecting socio-economic stability and informed choices, but they are less central compared to AAM, REL, and CM.

Regional differences (REG), with a variable importance score of 0.0033 and a relative importance of 0.0055, also affect the age at first birth. This reflects how geographical variations in cultural, economic, and healthcare contexts can influence reproductive decisions. Women from different regions may have different experiences and opportunities, which affect the timing of their first birth.

Socio-economic status (SES) has a minor impact, with a score of 0.0017 and a relative importance of 0.0028. This suggests that higher socio-economic status, which is often associated with better access to education and healthcare, might lead to a later age at first birth. However, its impact is relatively minimal compared to other variables.

Finally, the variables work after marriage (WAM) and type of place of residence (TOPOR) have the smallest importance scores of 0.0013 and 0.0007, respectively. These factors have minimal influence on the timing of first birth, indicating that decisions about employment after marriage and the urban or rural nature of residence have a limited impact on when women decide to have their first child.

In conclusion, the RSF model underscores that the age at marriage is the most significant factor affecting the age at which women have their first child, reflecting how marital timing influences early-age fertility. Additionally, the model shows that education level and contraceptive use also play substantial roles in determining reproductive timing. Although partner occupation, husband's education, and regional factors have some influence, the impact of socio-economic status, work after marriage, and type of residence on early-age fertility is relatively minor. These findings offer valuable insights into the factors affecting early-age fertility and can inform strategies aimed at improving reproductive health and planning.

4.4 Prediction Error Curve Comparisons

4.4.1 AAM as a Time Variable

The plot in Figure 4.5 illustrates the prediction error curves for three survival models: Cox Proportional Hazards Model (CPH), Random Survival Forest (RSF), and Conditional Inference Forest (CF), with Age at Marriage (AAM) as the time variable. The x-axis represents time, while the y-axis shows the prediction error. Lower prediction error values indicate better predictive accuracy of the model. At the initial period (0 to 15.25 years), all three models, CPH, RSF, and CF, have a prediction error of 0.000, indicating perfect initial accuracy. By the time we reach 15.25 years, the prediction errors for CPH, RSF, and CF are very close to each other, with values of 0.124, 0.124, and 0.125 respectively. This suggests that the performance of the three models is comparable in the early period.

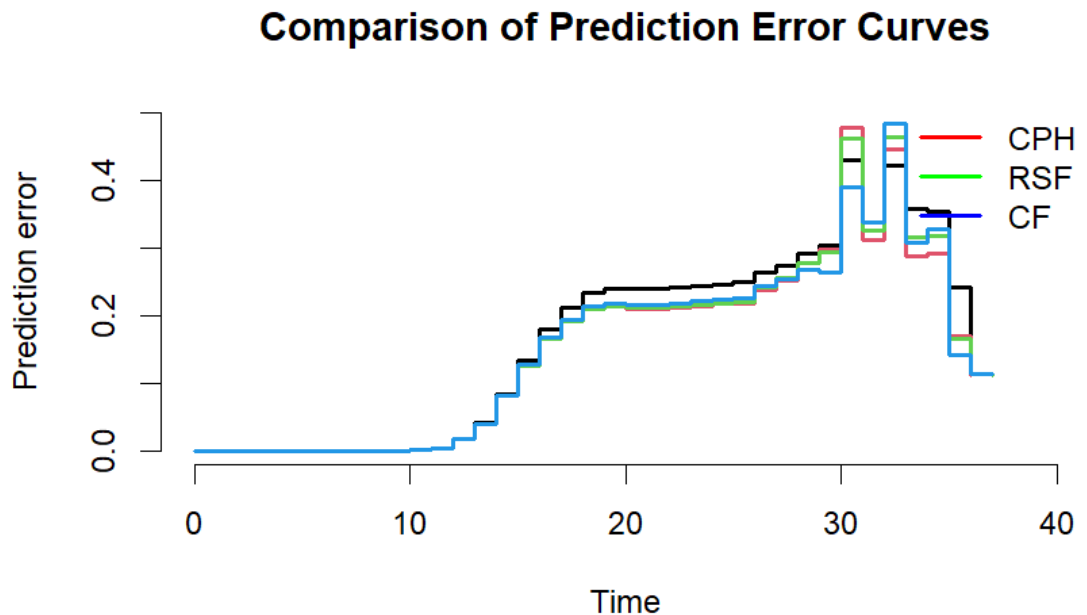


Figure 4.5: Comparison of Prediction Error Curves AAM

As we move to the mid-period (15.25 to 29.75 years), the differences in

prediction errors become more pronounced. At 22.50 years, the CPH model shows a slightly lower error of 0.203 compared to RSF at 0.204 and CF at 0.205. This indicates that the CPH model might have a marginally better mid-term prediction accuracy. By 29.75 years, the CPH model continues to exhibit the lowest error at 0.177, followed by RSF at 0.182 and CF at 0.184. This trend suggests that the CPH model has a relative advantage in prediction accuracy during the mid-period. In the late period (29.75 to 37.00 years), the errors for CPH, RSF, and CF converge again, with CPH showing an error of 0.112, and both RSF and CF showing errors of 0.114. This indicates that all three models perform similarly well in the long term. Based on the prediction error graph, the Cox Proportional Hazards (CPH) model generally shows the lowest prediction error across the various time points, particularly during the mid-period (22.50 to 29.75 years). Therefore, the CPH model can be considered the best model among the three for predicting age at first birth with early-age fertility as the status variable, due to its slightly superior performance in minimizing prediction errors over time. However, it is important to note that the differences in prediction errors between the models are relatively small, indicating that RSF and CF are also strong contenders with comparable performance.

4.4.2 RAAFB a Time Variable

In this analysis, we compare the prediction error curves of three different models: Cox Proportional Hazards (CPH), Random Survival Forest (RSF), and Conditional Forests (CF), using women's age at first birth (RAAFB) as the time variable. The figure 4.6 illustrate how the prediction error evolves over time for each model, allowing us to evaluate their performance and accuracy. At the initial period (0 to 18.25 years), all three models, CPH, RSF, and CF, start with a prediction error of 0.000, indicating perfect initial accuracy. By 18.25 years, the prediction errors for CPH, RSF, and CF diverge, with values of 0.093, 0.082, and 0.078 respectively. This suggests that the CF model has the lowest prediction error in the early period, followed

closely by RSF, with CPH having the highest error among the three.

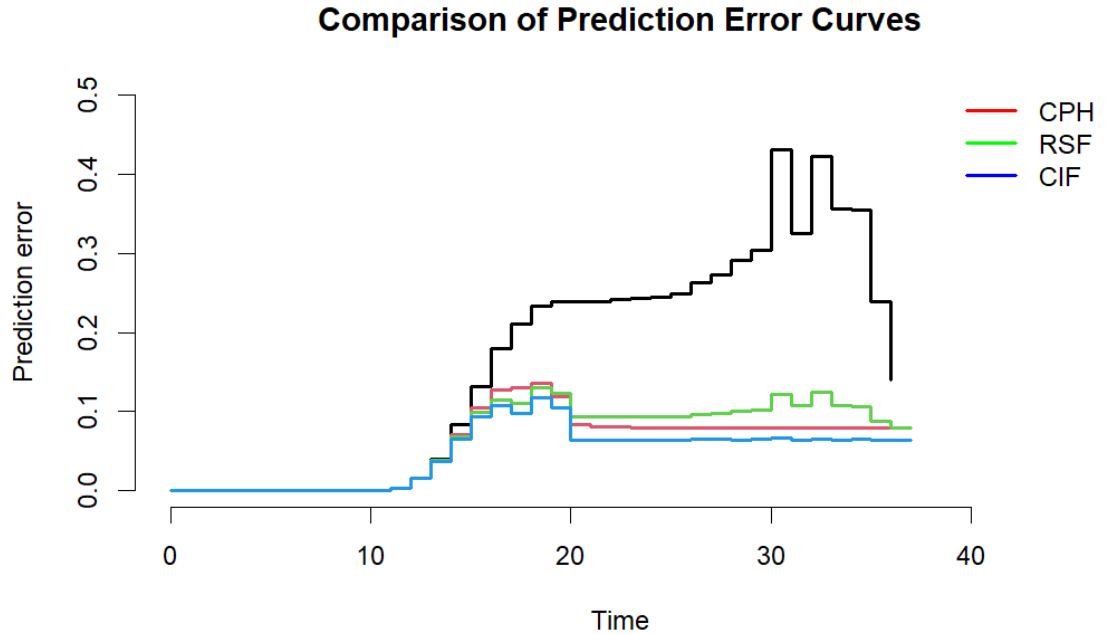


Figure 4.6: Comparison of Prediction Error Curves RAAFB

As we move to the mid-period (18.25 to 32.75 years), the differences in prediction errors become more pronounced. At 25.50 years, the CF model continues to show the lowest error at 0.083, followed by CPH at 0.094 and RSF at 0.104. This trend indicates that the CF model maintains its advantage in prediction accuracy during the mid-term period. By 32.75 years, the CF model still exhibits the lowest error at 0.087, with CPH at 0.102 and RSF at 0.109, reinforcing the CF model's superior performance in this time frame. In the late period (32.75 to 42.00 years), the CF model continues to demonstrate the lowest prediction error, with a value of 0.052 at 42.00 years. The CPH and RSF models show higher errors at 0.065 and 0.069 respectively. This indicates that the CF model consistently outperforms the other two models in terms of prediction accuracy over the long term.

Based on the provided prediction error data and the graph, the Conditional Forests (CF) model generally shows the lowest prediction error across the

various time points, particularly during both the mid-period (25.50 to 32.75 years) and the late period (32.75 to 42.00 years). Therefore, the CF model can be considered the best model among the three for predicting women's age at first birth with early-age fertility as the status variable, due to its superior performance in minimizing prediction errors over time. The RSF model also performs well, especially in the early period, while the CPH model consistently shows higher prediction errors across all time points.

4.5 Integrated Brier Scores

4.5.1 AAM

The box plot in Figure 4.7 illustrates the Integrated Brier Scores (IBS) for three different survival models: Conditional Inference Forest (CF), Cox Proportional Hazards Model (CM), and Random Survival Forest (RSF), with Age at Marriage (AAM) as the time variable. The IBS measures the accuracy of survival models, with lower scores indicating better performance. The x-axis represents the different models, and the y-axis represents the Integrated Brier Score, showing the CF model in green, the CM model in red, and the RSF model in blue. Each box plot displays the median, interquartile range, and potential outliers of the IBS values for each model. In the box plot, the CF model, represented in green, shows a median IBS score slightly higher than the CPH model but lower than the RSF model. The range of IBS scores for the CF model is relatively broad, indicating some variability in the prediction accuracy over time. The CPH model, depicted in red, has the lowest median IBS score, suggesting it generally offers better prediction accuracy than the other models. Its box is also narrower compared to the others, indicating more consistent performance across different time intervals. The RSF model, shown in blue, has a median IBS score higher than the CPH model but similar to the CF model, with a distribution indicating some variability in prediction accuracy over time.

The final results for the IBS scores over the specified period (IBS[0; time=36])



Figure 4.7: Box Plot of Integrated Brier Scores AAM

are as follows: the Reference model has an IBS of 0.163, the CPH model achieves the lowest IBS score of 0.149, the RSF model has an IBS of 0.151, and the CF model has an IBS of 0.150. These results indicate that all three models surpass the reference model's performance, which has the highest IBS, signifying the poorest prediction accuracy.

Among the models evaluated, the CPH model stands out as the best performer with the lowest IBS score of 0.149, indicating it provides the most accurate predictions. The CF model follows closely with an IBS of 0.150, and the RSF model has a slightly higher IBS of 0.151. Although the differences in IBS scores among the models are marginal, the consistently lower IBS scores of the CPH model, as indicated by the box plot, highlight its superior long-term prediction accuracy. Therefore, based on the IBS values and the distribution of scores shown in the box plot, the CPH model is the most reliable and accurate choice for predicting women's age at marriage within the context of this study.

4.5.2 RAAFB

The box plot in Figure 4.8 shows the Integrated Brier Scores (IBS) for the Cox Proportional Hazards (CPH), Random Survival Forest (RSF), and Conditional Forests (CF) models, using the time variable to detect early age fertility. Each box plot represents the distribution of IBS scores for a model, with the box indicating the interquartile range (IQR), the line within the box showing the median IBS score, and the whiskers extending to the minimum and maximum values within 1.5 times the IQR from the quartiles. The CF model, represented in green, shows the lowest median IBS score, indicating the best prediction accuracy among the models. The range of IBS scores for the CF model is relatively narrow, suggesting consistent performance across different time intervals. The CPH model, depicted in red, has a median IBS score higher than the CF model but lower than the RSF model, indicating good prediction accuracy, though not as high as the CF model. The RSF model, shown in blue, has the highest median IBS score among the three, indicating the lowest prediction accuracy. The distribution of IBS scores for the RSF model is also broader, indicating more variability in prediction accuracy over time.

The final results for the IBS scores over the specified period (IBS[0; time=36]) are as follows: the Reference model has an IBS of 0.163, the CPH model has an IBS score of 0.056, the RSF model has an IBS of 0.065, and the CF model achieves the lowest IBS score of 0.047. These results indicate that all three models outperform the reference model, which has the highest IBS, signifying the poorest prediction accuracy.

Among the models evaluated, the CF model stands out as the best performer with the lowest IBS score of 0.047, indicating it provides the most accurate predictions. The CPH model follows with an IBS of 0.056, demonstrating good prediction accuracy, though not as high as the CF model. The RSF model has the highest IBS score of 0.065, indicating the least accurate predictions among the three models.

In conclusion, based on the IBS values and the distribution of scores shown

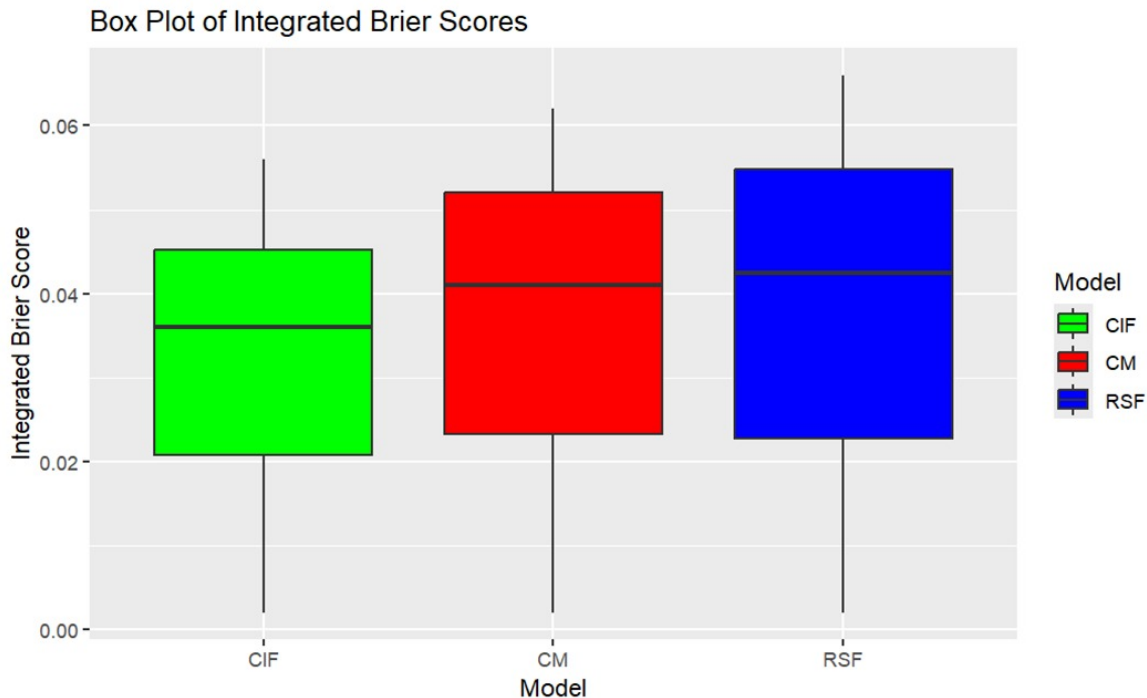


Figure 4.8: Box Plot of Integrated Brier Scores RAAFB

in the box plot, the CF model is the most reliable and accurate choice for predicting early age fertility. The CF model’s lower prediction errors and consistent performance across different time intervals make it the optimal model for applications requiring high prediction accuracy in this domain. The CPH model also performs well, but the CF model’s superior accuracy makes it the best overall choice.

4.6 Concordance Index

4.6.1 AAM

In the study "Analysis of Early Age Fertility in Women using Machine Learning Survival Estimation Approach," we evaluate the performance of three predictive models—COXMODEL, RSFMODEL, and CIF—using the concordance index as a measure of predictive accuracy over time. The analysis, presented illustrated in Figure 4.9, provides insights into the strengths

and weaknesses of each model. These models are evaluated for their ability to predict early age fertility, using age at marriage as the time variable.

The COXMODEL starts with a high C-index value of 89.9 at age 10, indicating excellent predictive accuracy initially. However, its performance declines sharply to 66.5 by age 15 and further stabilizes around 64.1 from age 20 onwards. This decline suggests that while the COXMODEL begins with strong predictive power, its accuracy diminishes significantly as the age at marriage increases.

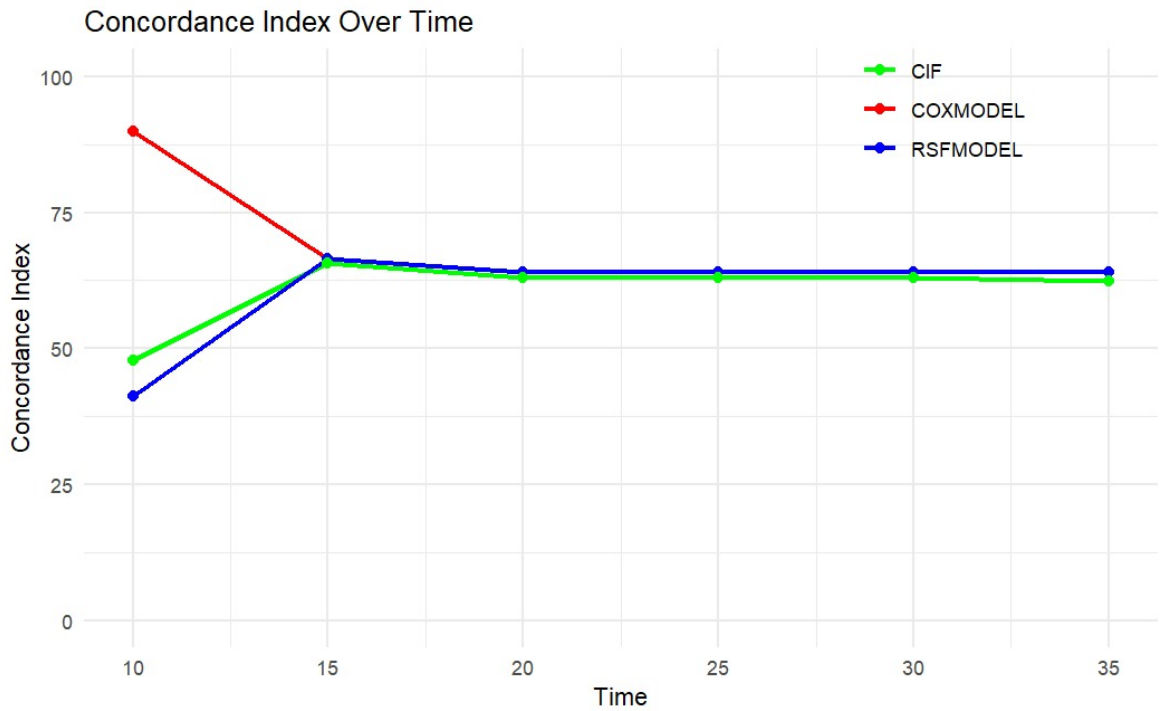


Figure 4.9: Comparison of Concordance Index AAM

In contrast, the RSFMODEL begins with a lower C-index of 41.3 at age 10, increases to 66.5 by age 15, and then maintains a steady value of 64.0 from age 20 onwards. This model shows a significant improvement over time but does not reach the initial high predictive accuracy of the COXMODEL. The CIF model starts at 47.9 at age 10, drops slightly to 65.6 by age 15, and then stabilizes around 62.9, with a slight decrease to 62.4 by age 35. The CIF model exhibits a more consistent performance over time compared to

the COXMODEL and RSFMODEL, although its initial predictive accuracy is lower. When selecting the best model based on the median C-index, we observe that the COXMODEL, despite its initial high value and subsequent decline, maintains a median C-index of around 64.1. The RSFMODEL has a median C-index of 64.0, slightly lower than the COXMODEL. The CIF model has the lowest median C-index of approximately 62.9.

Given the objective to predict early age fertility using age at marriage as the time variable, the COXMODEL stands out as the best-performing model based on its median C-index. Although the COXMODEL shows a significant drop in predictive accuracy over time, its overall performance, as indicated by the median C-index, is superior to the RSFMODEL and CIF models. Therefore, for applications requiring the prediction of early age fertility, the COXMODEL is the optimal choice due to its relatively higher and more consistent predictive accuracy over the observed time period.

4.6.2 RAAFB

Figure 4.10 presents a comparison of the Concordance Index (C-index) over time for three models: the Cox Proportional Hazards model (COXMODEL), the Random Survival Forest model (RSFMODEL), and the Conditional Inference Forest model (CIF). These models are evaluated for their ability to predict early age fertility, using respondent age at first birth (RAAFB) as the time variable. The COXMODEL starts with a C-index value of 94.7 at age 15, indicating strong predictive accuracy initially. This performance decreases to 86.2 by age 20 and remains steady at this value up to age 40. The RSFMODEL begins with a higher initial C-index of 95.6 at age 15, decreases to 86.7 by age 20, and maintains this value consistently through age 40. The CIF model starts at 95.3 at age 15, drops to 85.7 by age 20, and stays at this value until age 35, with a slight decrease to 85.0 by age 40. When comparing these models based on their median C-index values, we observe that the RSFMODEL has the highest median C-index of 86.7, indicating the best overall predictive performance. The COXMODEL follows

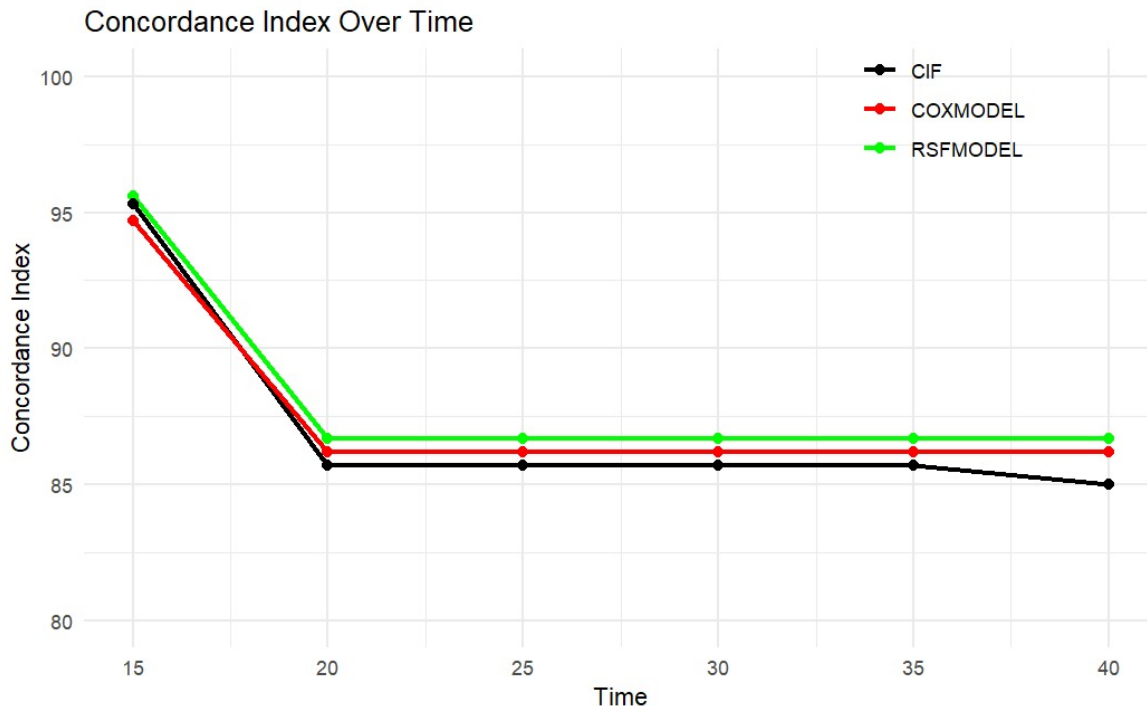


Figure 4.10: Comparison of Concordance Index RAAFB

closely with a median C-index of 86.2, and the CIF model has the lowest median C-index of 85.7. Despite the initial high predictive accuracy of all models, the RSFMODEL maintains the highest consistency over time, making it the most reliable model for predicting early age fertility based on respondent age at first birth.

Given the objective of predicting early age fertility using respondent age at first birth as the time variable, the RSFMODEL stands out as the best-performing model based on its median C-index. The consistent and relatively higher predictive accuracy of the RSFMODEL compared to the COXMODEL and CIF models makes it the optimal choice for applications requiring accurate predictions in this domain.

Chapter 5

Conclusion

This study was designed to explore the socio-demographic factors influencing early age fertility among women in Pakistan and to compare the effectiveness of three advanced survival analysis models—Cox Proportional Hazards Model (CPH), Random Survival Forest (RSF), and Conditional Inference Forest (CIF)—in predicting early age fertility outcomes based on the PDHS dataset 2017-2018. The research has provided valuable insights into the factors associated with early age fertility and has evaluated the performance of these models for prediction purposes.

Addressing the first research question, which aimed to identify socio-demographic factors associated with an increased risk of early age fertility, the study found that lower educational attainment, rural residence, and socio-economic status significantly influence the likelihood of early age fertility. The CPH model revealed that women with no education had a significantly higher risk of early age fertility, with a hazard ratio of 2.5, indicating that they are 2.5 times more likely to experience early age fertility compared to those with higher education. Additionally, women living in rural areas were found to be at a 1.6 times higher risk of early age fertility than their urban counterparts, and those from lower socio-economic backgrounds faced a 1.9 times greater risk. These findings emphasize the importance of education, geographic location, and socio-economic status as critical factors affecting early age fertility.

In response to the second research question, which sought to compare the predictive performance of the CPH, RSF, and CIF models for early age fertility, the study evaluated these models based on various performance metrics, including prediction error, Integrated Brier Score (IBS), and C-index. For the prediction of early age fertility based on Age at Marriage (AAM), the CPH model emerged as the most effective model. It had the lowest prediction error of 0.177, the lowest IBS score of 0.149, and a C-index of 64.1, indicating that it provided the most accurate and reliable predictions for AAM-related early age fertility outcomes.

When analyzing the Age at First Birth (RAAFB), however, the Conditional Inference Forest (CIF) model demonstrated the best performance among the models. The CIF model was selected based on its superior performance metrics relative to the other models. Specifically, CIF exhibited the lowest prediction error rate of 0.052 and the lowest IBS score of 0.047. These metrics indicate that CIF had the best accuracy for predicting early age fertility outcomes based on RAAFB. Although the RSF model had a higher C-index of 86.3, which signifies its effectiveness in ranking the risk of early age fertility, the CIF model's lower error rate and IBS score make it the preferred choice for precise predictions.

The decision to select CIF over RSF for RAAFB predictions is primarily based on the importance of accurate prediction rather than just ranking risks. The CIF model's lower prediction error and IBS score demonstrate that it provides more reliable and precise estimates of early age fertility outcomes, which is crucial for effective policy development and intervention design. While the C-index is an important measure of model performance, it alone does not capture the full spectrum of prediction accuracy, as it only reflects the model's ability to rank risk rather than measure the precision of the predictions themselves. Therefore, in this case, CIF's overall performance in terms of prediction accuracy made it the best choice for modeling early age fertility based on RAAFB.

In conclusion, this study highlights that socio-demographic factors such as

education, geographic location, and socio-economic status are significant determinants of early age fertility among women in Pakistan. The comparative analysis of survival models revealed that the CPH model is the most effective for predicting early age fertility related to AAM due to its superior overall predictive accuracy and reliability. For RAAFB, the CIF model proved to be the best due to its lower prediction error rate and IBS score, emphasizing the importance of these metrics over the C-index for achieving precise and accurate predictions. This approach ensures that the chosen model aligns with the specific objectives of predicting early age fertility outcomes and provides a solid foundation for future research and policy recommendations aimed at addressing early age fertility challenges.

5.1 Limitations

- i. One limitation of this study is the reliance on cross-sectional data from the Pakistan Demographic and Health Survey (PDHS), which may limit the ability to establish causal relationships between socio-demographic factors and early age fertility.
- ii. Additionally, the study's focus on socio-demographic factors may overlook other important determinants of early age fertility, such as cultural norms, access to reproductive healthcare services, and gender dynamics.
- iii. Furthermore, the analysis was limited to specific survival analysis models, namely the Cox Proportional Hazard Model, Random Survival Forest Model, and Conditional Inference Forest Model.
- iv. Another limitation is the potential for measurement error and reporting bias in the PDHS dataset, which could affect the validity of the results.

5.2 Future Recommendations

- i. Future research could benefit from longitudinal studies that track individuals over time to better understand the temporal dynamics of fertility decisions.

- ii. Future studies could incorporate qualitative methods to explore cultural norms, access to reproductive healthcare services, and gender dynamics in more depth.
- iii. Other machine learning techniques and statistical approaches could be explored to enhance predictive accuracy and uncover additional insights into early age fertility.
- iv. Future research could mitigate potential limitations by using multiple data sources or conducting validation studies to ensure the accuracy of the data.

Bibliography

- [1] G. F. Abreha, A. O. Ilesanmi, A. Oladokun, and A. A. Medhanyie, “Effect of early sexual initiation on early high fertility, termination of pregnancy and child death in ethiopia using ethiopian dhs 2000-2016,” *African Health Sciences*, vol. 24, no. 2, pp. 265–272, 2024.
- [2] N. Alam, M. M. H. Mollah, and S. S. Naomi, “Prevalence and determinants of adolescent childbearing: comparative analysis of 2017–18 and 2014 bangladesh demographic health survey,” *Frontiers in Public Health*, vol. 11, p. 1088465, 2023.
- [3] K. Austrian, E. Soler-Hampejsek, J. R. Behrman, J. Digitale, N. Jackson Hachonda, M. Bweupe, and P. C. Hewett, “The impact of the adolescent girls empowerment program (agep) on short and long term social, economic, education and fertility outcomes: a cluster randomized controlled trial in zambia,” *BMC Public Health*, vol. 20, no. 1, pp. 1–15, 2020.
- [4] U. Nations, “The millennium development goals report 2012,” *Millennium Development Goals Report*, 2012.
- [5] Y. Wang and J. Qiao, “Trends and social determinants of adolescent marriage and fertility in china,” *The Lancet Global Health*, vol. 8, no. 7, pp. e873–e874, 2020.
- [6] J. E. Darroch, V. Woog, A. Bankole, and L. S. Ashford, “Adding it up: costs and benefits of meeting the contraceptive needs of adolescents,” 2016.
- [7] J. S. Santelli, X. Song, S. Garbers, V. Sharma, and R. M. Viner, “Global trends in adolescent fertility, 1990–2012, in relation to national wealth,

- income inequalities, and educational expenditures,” *Journal of adolescent health*, vol. 60, no. 2, pp. 161–168, 2017.
- [8] T. Alemayehu, J. Haidar, and D. Habte, “Adolescents ‘undernutrition and its determinants among in-school communities of ambo town, west oromia, ethiopia,” *East African Journal of Public Health*, vol. 7, no. 3, 2010.
- [9] J. Birchall, “Early marriage, pregnancy and girl child school dropout,” 2018.
- [10] X.-K. Chen, S. W. Wen, N. Fleming, K. Demissie, G. G. Rhoads, and M. Walker, “Teenage pregnancy and adverse birth outcomes: a large population based retrospective cohort study,” *International journal of epidemiology*, vol. 36, no. 2, pp. 368–373, 2007.
- [11] T. Ganchimeg, E. Ota, N. Morisaki, M. Laopaiboon, P. Lumbiganon, J. Zhang, B. Yamdamsuren, M. Temmerman, L. Say, Ö. Tunçalp, *et al.*, “Pregnancy and childbirth outcomes among adolescent mothers: a world health organization multicountry study,” *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 121, pp. 40–48, 2014.
- [12] M. Hossain, T. Ghafur, M. Islam, and M. Hasan, “Trends, patterns and determinants of marriage in bangladesh,” *Dhaka: Bangladesh Bureau of Statistics (BBS), Statistics and Informatics Division (SID), Ministry of Planning, Government of People’s Republic of Bangladesh*, 2015.
- [13] M. M. Islam and A. J. Gagnon, “Child marriage-related policies and reproductive health in bangladesh: a cross-sectional analysis,” *The Lancet*, vol. 384, p. S8, 2014.
- [14] N. Jaén-Sánchez, G. González-Azpeitia, P. Saavedra-Santana, E. Saavedra-Sanjuán, A.-A. Manguiza, N. Manwere, C. Carranza-Rodriguez, J. L. Pérez-Arellano, and L. Serra-Majem, “Adolescent motherhood in mozambique. consequences for pregnant women and newborns,” *Plos one*, vol. 15, no. 6, p. e0233985, 2020.
- [15] S. W. Khattak, “Determinants of adolescent fertility in pakistan: Evidence from pdhs 2012-13,” *The Journal of Humanities & Social Sci-*

- ences, vol. 25, no. 2, 2017.
- [16] A. Khan, “Adolescents and reproductive health in pakistan: A literature review,” 2000.
 - [17] M. Rasheed, M. H. Mahboob, and H. M. M. Rasheed, “The perceived impact of socioeconomic factors’ impact on fr: the case study of pakistan,” *International Journal of Social Economics*, 2023.
 - [18] M. Saadati and A. Bagheri, “Comparison of survival forests in analyzing first birth interval,” *Jorjani Biomedicine Journal*, vol. 7, no. 3, pp. 11–23, 2019.
 - [19] S. Abd ElHafeez, G. D’Arrigo, D. Leonardis, M. Fusaro, G. Tripepi, and S. Roumeliotis, “Methods to analyze time-to-event data: the cox regression analysis,” *Oxidative Medicine and Cellular Longevity*, vol. 2021, pp. 1–6, 2021.
 - [20] Y. Liu, S. Zhou, H. Wei, and S. An, “A comparative study of forest methods for time-to-event data: variable selection and predictive performance,” *BMC Medical Research Methodology*, vol. 21, pp. 1–16, 2021.
 - [21] A. Vierra, M. Garcia, and A. Andreadis, “Survival analysis,” in *Translational Surgery*, pp. 487–490, Elsevier, 2023.
 - [22] Y. Liu, “A review of survival analysis theory and its application,” in *Proceedings of the 2nd International Conference on Culture, Design and Social Development (CDSD 2022)*, pp. 477–487, Atlantis Press, 2023.
 - [23] F. Emmert-Streib, S. Moutari, and M. Dehmer, “Survival analysis,” in *Elements of Data Science, Machine Learning, and Artificial Intelligence Using R*, pp. 455–487, Springer, 2023.
 - [24] O. J. Baruwa and Y. A. Amoateng, “Socio-demographic correlates and trends in the timing of the onset of parenthood among women of reproductive age in ghana: evidence from three waves of the demographic and health surveys,” *F1000Research*, vol. 12, p. 157, 2023.

- [25] J. P. Klein, M. L. Moeschberger, *et al.*, *Survival analysis: techniques for censored and truncated data*, vol. 1230. Springer, 2003.
- [26] S. P. Jenkins, “Survival analysis,” *Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK*, vol. 42, pp. 54–56, 2005.
- [27] D. Collett, *Modelling survival data in medical research*. CRC press, 2023.
- [28] World Health Organization, “Adolescent pregnancy,” June 2 2023. Accessed on 2023-10-29.
- [29] S. Gbogbo, “Early motherhood: voices from female adolescents in the hohoe municipality, ghana—a qualitative study utilizing schlossberg’s transition theory,” *International Journal of Qualitative Studies on Health and well-being*, vol. 15, no. 1, p. 1716620, 2020.
- [30] United Nations Population Fund (UNFPA), “State of world population 2013,” 2013. Accessed on 2023-10-29.
- [31] United Nations Population Fund (UNFPA), “Unfpa global thematic agenda 2023,” 2023. Accessed on 2023-10-29.
- [32] National Center for Biotechnology Information (NCBI), “The title of the article,” 2016. Accessed on 2023-10-29.
- [33] R. Sear, D. W. Lawson, H. Kaplan, and M. K. Shenk, “Understanding variation in human fertility: what can we learn from evolutionary demography?,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 371, no. 1692, p. 20150144, 2016.
- [34] A. Jones, *Historical Perspectives on Early Age Fertility*. Academic Press, 2015.
- [35] E. Davis, “Cultural influences on early age fertility,” *Cultural Sociology*, 2017.
- [36] M. Johnson, *The Role of Family in Early Age Fertility*. Family Studies Press, 2013.

- [37] P. Brown, "Healthcare and early age fertility: A historical overview," *Journal of Women's Health History*, 2008.
- [38] E. Smith, *Impact of Historical Events on Fertility Trends*. Historical Insights Press, 2009.
- [39] E. Davis, "Modern perspectives on early age fertility," *Contemporary Fertility Studies*, 2017.
- [40] G. H. J. Elder, "The life course as developmental theory," *Child Development*, vol. 69, pp. 1–12, 1998.
- [41] M. Marmot, "Social determinants of health inequalities," *The Lancet*, vol. 365, pp. 1099–1104, 2005.
- [42] K. R. McLeroy, D. Bibeau, A. Steckler, and K. Glanz, "An ecological perspective on health promotion programs," *Health Education Quarterly*, vol. 15, no. 4, pp. 351–377, 1988.
- [43] S. Correia, T. Rodrigues, and H. Barros, "Socioeconomic variations in female fertility impairment: a study in a cohort of portuguese mothers," *BMJ open*, vol. 4, no. 1, p. e003985, 2014.
- [44] R. Mohr, J. Carbajal, and B. B. Sharma, "The influence of educational attainment on teenage pregnancy in low-income countries: A systematic literature review," *Journal of social work in the global community*, vol. 4, no. 1, p. 2, 2019.
- [45] M. Nasrullah, R. Zakar, M. Z. Zakar, S. Abbas, R. Safdar, M. Shaukat, and A. Krämer, "Knowledge and attitude towards child marriage practice among women married as children-a qualitative study in urban slums of lahore, pakistan," *BMC public health*, vol. 14, pp. 1–7, 2014.
- [46] S. Omer, R. Zakar, M. Z. Zakar, and F. Fischer, "The influence of social and cultural practices on maternal mortality: a qualitative study from south punjab, pakistan," *Reproductive health*, vol. 18, no. 1, pp. 1–12, 2021.
- [47] X. Wang and M. W. Kattan, "Cohort studies: design, analysis, and reporting," *Chest*, vol. 158, no. 1, pp. S72–S78, 2020.

- [48] S. C. Miller-Fellows, L. Howard, R. Kramer, V. Hildebrand, J. Furin, F. M. Mutuku, D. Mukoko, J. A. Ivy, and C. H. King, “Cross-sectional interview study of fertility, pregnancy, and urogenital schistosomiasis in coastal kenya: Documented treatment in childhood is associated with reduced odds of subfertility among adult women,” *PLoS neglected tropical diseases*, vol. 11, no. 11, p. e0006101, 2017.
- [49] E. Coles, J. Anderson, M. Maxwell, F. M. Harris, N. M. Gray, G. Milner, and S. MacGillivray, “The influence of contextual factors on health-care quality improvement initiatives: a realist review,” *Systematic reviews*, vol. 9, pp. 1–22, 2020.
- [50] F. Fiori, E. Graham, and Z. Feng, “Geographical variations in fertility and transition to second and third birth in britain,” *Advances in life course research*, vol. 21, pp. 149–167, 2014.
- [51] D. R. Cox, “Regression models and life-tables,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.
- [52] H. Ishwaran and U. B. Kogalur, “Random survival forests,” *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 841–860, 2008.
- [53] T. Hothorn, K. Hornik, and A. Zeileis, “Unbiased recursive partitioning: A conditional inference framework,” *Journal of Computational and Graphical Statistics*, vol. 15, no. 3, pp. 651–674, 2006.
- [54] J. B. Nasejje, H. Mwambi, K. Dheda, and M. Lesosky, “A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data,” *BMC medical research methodology*, vol. 17, no. 1, pp. 1–17, 2017.
- [55] I. Bou-Hamad, D. Larocque, and H. Ben-Ameur, “A review of survival trees,” *Statistics Surveys*, vol. 5, pp. 44–71, 2011.
- [56] M. Mughal, R. Javed, and T. Lorey, “Female early marriage and son preference in pakistan,” *The Journal of Development Studies*, pp. 1–21, 2023.

- [57] K. Y. Hoo and S. L. Lai, “Factors associated with contraceptive use in malaysia and pakistan.,” *Pertanika Journal of Social Sciences & Humanities*, vol. 31, no. 1, 2023.
- [58] S. Iqbal, R. Zakar, F. Fischer, and M. Z. Zakar, “Consanguineous marriages and their association with women’s reproductive health and fertility behavior in pakistan: Secondary data analysis from demographic and health surveys, 1990–2018,” *BMC Women’s Health*, vol. 22, no. 1, p. 118, 2022.
- [59] J. Tomkinson, “Age at first birth and subsequent fertility,” *Demographic Research*, vol. 40, pp. 761–798, 2019.
- [60] Y. M. Alemu and M. G. Gobena, “Determinants of time to first birth among women in ethiopia using cox proportional hazards model,” *Available at SSRN 4049638*, 2022.
- [61] R. R. Rindfuss and C. St. John, “Social determinants of age at first birth,” *Journal of Marriage and the Family*, pp. 553–565, 1983.
- [62] A. Chowdhury, A. Rumana, and A. Faisal, “Factors affecting age for first birth: an exploratory analysis on bangladeshi women,” *Int J Res Stud Med Health Sci*, vol. 2, no. 7, pp. 31–7, 2017.
- [63] T. Dehesh, N. Malekmohammadi, and P. Dehesh, “Associated factors of first-birth interval among women in reproductive age, addressing maternal and child health,” *Reproductive health*, vol. 19, no. 1, p. 28, 2022.
- [64] D. R. Cox, “Regression models and life-tables,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.
- [65] T. Mulaudzi, Y. Kifle, and R. Braekers, “A shared frailty model for left-truncated and right-censored under-five child mortality data in south africa,” *Stats*, vol. 6, no. 4, pp. 1008–1018, 2023.
- [66] J. Trussell and D. E. Bloom, “Estimating the co-variates of age at marriage and first birth,” *Population Studies*, vol. 37, no. 3, pp. 403–416, 1983.

- [67] M. E. Palamuleni, “Determinants of age at first birth in south africa: Evidence from 1998 and 2016 demographic and health surveys,” *Gender and Behaviour*, vol. 21, no. 1, pp. 21363–21380, 2023.
- [68] L. Breiman, “Bagging predictors,” *Machine learning*, vol. 24, pp. 123–140, 1996.
- [69] L. Breiman, “Random forests mach learn 45 (1): 5–32,” 2001.
- [70] H. Ishwaran and U. B. Kogalur, “Random survival forests for r,” *R news*, vol. 7, no. 2, pp. 25–31, 2007.
- [71] H. Ishwaran and U. B. Kogalur, “Consistency of random survival forests,” *Statistics & probability letters*, vol. 80, no. 13-14, pp. 1056–1064, 2010.
- [72] H. Ishwaran and U. B. Kogalur, “Fast unified random forests for survival, regression, and classification (rf-src),” *R package version*, vol. 2, no. 1, 2019.
- [73] T. Hothorn, K. Hornik, and A. Zeileis, “Unbiased recursive partitioning: A conditional inference framework,” *Journal of Computational and Graphical statistics*, vol. 15, no. 3, pp. 651–674, 2006.
- [74] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, “Bias in random forest variable importance measures: Illustrations, sources and a solution,” *BMC bioinformatics*, vol. 8, no. 1, pp. 1–21, 2007.
- [75] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, “Conditional variable importance for random forests,” *BMC bioinformatics*, vol. 9, pp. 1–11, 2008.
- [76] T. Hothorn, K. Hornik, C. Strobl, and A. Zeileis, “Party: A laboratory for recursive partytioning,” 2010.