Polytechnic Institute of Management and Technology – GAIA

Master's in Web Technologies and Systems Engineering

Final course dissertation report

Breast Cancer Survival Time Prediction Using Machine Learning

Fida Ben Slimen

VILA NOVA DE GAIA

2026

Polytechnic Institute of Management and Technology - GAIA

Master's in Web Technologies and Systems Engineering

Final course dissertation report

Breast Cancer Survival Time Prediction Using Machine Learning

Fida Ben Slimen

Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Web Technologies and Systems Engineering under the supervision of Prof. Jorge Duque, and in accordance with the approval order number 9371/2020.

VILA NOVA DE GAIA

2026

INSTITUTO POLITÉCNICO DE GESTÃO E TECNOLOGIA

Breast Cancer Survival Time Prediction Using Machine Learning

Fida Ben Slimen

Aprovada em 27/03/2026

Composição do Júri

..

Presidente
Prof. Especialista José Joaquim Moreira

Arguente
Profª. Doutora Célia Talma Gonçalves

Orientador
Prof. Doutor Jorge Pereira Duque

Vila Nova de Gaia
2026

Acknowledgements

I would like to express my profound gratitude to the following people and institution, without whom it would not have been possible to conclude this pivotal stage of my academic journey at ISLA Gaia.

To my supervisor, Dr. Jorge Duque, whose deep knowledge and careful guidance were instrumental to the direction and success of this research.

To my parents, Foued and Fadwa, my brother Eslem, and my little sister Nour, for their unwavering love and unconditional support throughout this journey. To my partner, Jihed, a heartfelt thank you for your constant support and encouragement; without you, this master's journey would have been considerably more challenging. Your patience, understanding, and belief in me were a continuous source of strength and motivation.

To all of you, I extend my most sincere and profound gratitude.

Abstract

This dissertation, developed as part of the Master's Dissertation in Web Technologies and Systems Engineering at ISLA, aims to apply data-driven methods to predict the survival time of patients with breast cancer. Despite major advances in treatment, breast cancer remains one of the most significant public health concerns worldwide. The growing volume and complexity of clinical data highlight the need for robust analytical and machine learning approaches capable of enhancing the accuracy and interpretability of clinical decision-making.

Following the CRISP-DM methodology, this project integrates data extraction, transformation, and modeling within a complete ETL pipeline. An open-access dataset from the Kaggle platform was used to ensure transparency, reproducibility, and ethical data handling.

Two complementary predictive tasks were developed. A Random Forest classifier was employed to distinguish between short- and long-term survival cases, achieving over 93% accuracy and a weighted F1-score above 0.90, with improved detection of short-survival cases through oversampling and class weighting. In parallel, a Cox Proportional Hazards model was implemented for survival analysis, demonstrating strong discriminative ability (C-index > 0.95) and good calibration.

The results confirm that the combination of structured ETL data preparation, machine learning, and survival modeling provides an effective framework for discovering prognostic signals in heterogeneous clinical datasets, ultimately supporting more informed and data-driven decisions in the healthcare sector.

Keywords: *Breast Cancer, Survival Analysis, ETL, Random Forest, Cox Regression, Predictive Modeling*

Resumo

Esta dissertação, desenvolvida no âmbito do Mestrado em Web Technologies and Systems Engineering no ISLA, aplica métodos baseados em dados para prever o tempo de sobrevivência de pacientes com cancro da mama. Apesar dos avanços significativos no tratamento, o cancro da mama continua a ser um dos principais desafios de saúde pública a nível mundial. O aumento do volume e da complexidade dos dados clínicos evidencia a necessidade de abordagens analíticas e de aprendizagem automática robustas, capazes de melhorar a precisão e a interpretabilidade do processo de apoio à decisão clínica.

Seguindo a metodologia CRISP-DM, o projeto integra as fases de extração, transformação e modelação de dados num pipeline ETL completo. Foi utilizado um conjunto de dados de acesso aberto proveniente da plataforma Kaggle, garantindo transparência, reprodutibilidade e utilização ética dos dados.

Foram desenvolvidas duas tarefas preditivas complementares. Um classificador Random Forest foi utilizado para distinguir entre casos de sobrevivência curta e longa, alcançando mais de 93% de precisão e um F1-score ponderado superior a 0,90, com melhor deteção dos casos de sobrevivência curta através da aplicação de técnicas de oversampling e ponderação de classes. Em paralelo, o modelo de Riscos Proporcionais de Cox foi implementado para a análise de sobrevivência, demonstrando elevada capacidade discriminativa (C-index > 0,95) e boa calibração.

Os resultados confirmam que a combinação entre o processamento estruturado de dados através de ETL, a aprendizagem automática e a modelação de sobrevivência constitui uma abordagem eficaz para identificar sinais prognósticos em conjuntos de dados clínicos heterogéneos. Esta integração apoia decisões de saúde mais informadas, transparentes e orientadas por dados.

Palavras-chave: *Cancro da Mama, Análise de Sobrevivência, ETL, Random Forest, Regressão de Cox, Modelação Preditiva*

INDEX

INDEX OF FIGURES

INDEX OF TABLES

INTRODUCTION

This chapter is divided into four sections: Context and motivation, Research Problem and Objectives, Contributions of this Work, and structure of the dissertation.

1. Context and Motivation:

Breast cancer remains a major global health challenge, representing the most common malignant neoplasm among women and one of the leading causes of cancer-related mortality worldwide (Li et al., 2024). Its high prevalence and impact on patients' quality of life underline the need for continuous progress in early detection, treatment, and survival prediction. The ability to accurately estimate survival time in breast cancer patients is essential for the personalization of therapeutic strategies, optimization of clinical management, and provision of reliable prognostic information to both patients and healthcare professionals.

Over the past decades, the increasing availability of clinical data and advances in ML have opened promising avenues for healthcare analytics. These technologies enable the extraction of complex patterns and predictive insights from large and heterogeneous datasets, revealing relationships that are often invisible to traditional statistical methods (Baidoo & Rodrigo, 2025). In the context of breast cancer, ML applications have shown potential in identifying risk factors, predicting treatment outcomes, and most importantly estimating survival probabilities.

This dissertation is positioned within this evolving landscape, motivated by the need to design robust and interpretable predictive systems capable of assisting clinicians in evidence-based decision-making. Predicting survival time supports more accurate patient stratification, personalized treatment planning, and the design of preventive strategies. The incorporation of Extract–Transform–Load (ETL) processes is particularly relevant in this context, as clinical data are often fragmented, incomplete, and heterogeneous. A well-structured ETL pipeline ensures data consistency, integrity,

and leakage prevention, thus enhancing model reliability and enabling transparent and reproducible analyses. By integrating ETL methodologies with advanced ML and survival analysis techniques, this work contributes to the development of a comprehensive, data-driven framework for breast cancer prognosis (Afroz Banu et al., 2023; Garcia et al., 2020).

2. Research Problem and Objectives

The core problem addressed in this dissertation lies in the complexity of accurately predicting survival time in patients with breast cancer, given the disease's heterogeneity and the multitude of prognostic factors. Although statistical and clinical models exist for this purpose, their predictive capacity may be limited by their linear nature or by an inability to capture complex interactions among variables. Additionally, the presence of censored data, an intrinsic feature of survival analysis requires specific methodological approaches that are not always fully explored in studies based solely on traditional regression.

Within this context, the main objective of this study is to develop and validate a robust predictive system for analyzing survival time in breast cancer patients using a combination of ETL, ML, and survival analysis models. The focus is on improving prognostic accuracy and interpretability. This research is based on the following testable hypotheses:

H_1: The **Cox Proportional Hazards** model demonstrates superior discrimination (measured by the C-index) and calibration (measured by the Integrated Brier Score) compared to non-survival regression models.

H_2: Implementing a **structured ETL pipeline** reduces data leakage and improves model robustness compared to ad-hoc data-preprocessing approaches.

In alignment with the main objective of this study, the research was structured around the following specific objectives:

1. Conduct a comprehensive literature review on survival analysis in breast cancer, identifying the main prognostic factors, statistical methodologies, and ML approaches used.

2. Implement an ETL process to integrate and preprocess breast-cancer clinical data from multiple sources, ensuring data quality and consistency.
3. Explore and compare the performance of different Machine Learning models (e.g., Random Forest Classifier) and statistical survival models (e.g., Cox Proportional Hazards)."
4. Evaluate model performance using appropriate survival evaluation metrics and compare against a clinical baseline.
5. Identify the most relevant prognostic factors for survival in breast cancer via model feature-importance analysis.
6. Develop interactive dashboards to visualize model results and facilitate clinical interpretation and decision-making.

In Conclusion, the project shows how applying Data science, ML, and visualization tools within a structured methodology such as CRISP-DM can trans- form raw clinical data into know-how that can support early interventions along with more informed decision-making in breast cancer therapy.

3. Contributions of this Work

This dissertation contributes to the field of predictive oncology by proposing a reproducible data-driven framework that integrates ETL processes, ML, and survival analysis for breast cancer prognosis. The work advances methodological rigor in data preparation, enhances the interpretability of survival models, and demonstrates how analytical outcomes can be transformed into actionable clinical insights through interactive visualization.

4. Structure of the Dissertation

This dissertation is organized into nine chapters, the introduction and state of the art of the project, then the next five chapters will describe each phase of the CRISP methodology, and we will finish by the discussion and conclusion:

- **Chapter 1 Introduction.** Sets the context and motivation, states the research problem and objectives, highlights the contributions, and outlines the structure of the dissertation.

- **Chapter 2 State of the Art and Theoretical Background** Reviews breast-cancer epidemiology and prognosis; introduces survival-analysis concepts (event time, censoring), Kaplan–Meier estimation, and Cox PH; surveys ML in healthcare and survival modelling; and summarizes evaluation metrics and related work.
- **Chapter 3 Methodology and Problem-Solving Strategy.** Describes the approach (CRISP–DM), problem framing, assumptions, and the end-to-end pipeline guiding data preparation, modelling, and validation.
- **Chapter 4 Data Preparation.** Details the ETL process: data extraction, column normalization and merging, cleaning and target isolation, feature typing, transformation summary, pipeline schema, and supporting EDA checks.
- **Chapter 5 Modeling.** Presents the modelling stage: linear-regression baseline, random-forest regression, and Cox proportional hazards (survival model); notes on reproducibility/governance; and a short synthesis of each technique.
- **Chapter 6 Model Validation and Evaluation.** Specifies the evaluation protocol and metrics, descriptive cohort statistics, train/test split strategy, test-set performance, visual diagnostics, sanity checks, and deployment-ready artefacts.
- **Chapter 7 Deployment.** Describes objectives, production data assets, the operational architecture, and the Power BI dashboards: Sheet 1 (operational overview & model interpretability) and Sheet 2 (survival analytics & Cox risk stratification), plus validation and monitoring.
- **Chapter 8 Discussion.** Interprets the results clinically, compares findings with the literature, discusses limitations, and offers practice-oriented implications and recommendations.
- **Chapter 9 Conclusions and Future Work.** Synthesizes the main findings and contributions and outlines immediate research directions (e.g., external validation, multi-omics integration, advanced interpretability, dynamic updating).

STATE OF THE ART AND THEORETICAL BACKGROUND

This chapter establishes the theoretical and empirical foundations for developing a breast cancer survival prediction system. It begins by outlining the main clinical and epidemiological aspects of breast cancer, and then presents the key concepts of survival analysis along with the machine learning methods commonly applied in prognostic modeling.

Additionally, this chapter discusses the journey of computational methods in healthcare with the focus on the role of advanced data-driven techniques such as RF, Cox regression, and modern deep-learning survival architectures in enhancing outcome prediction in terms of both accuracy and interpretability. In the end, a thorough analysis of recent studies is done where the methodological improvements, performance comparisons, and research gaps that underline the necessity of the current work are recognized.

1. Breast Cancer: Epidemiology, Risk Factors, and Prognosis

Breast cancer continues to be a major global health issue. GLOBOCAN 2024 reports that worldwide over 2.4 million new patients were diagnosed with breast cancer which is almost 12% of all cancers and thus the leading cancer killer among women (WHO, 2024). European countries are witnessing an increase in incidence rates with a majority of younger ones affected while in Portugal the number of new cases is around 7,000 every year and the mortality is getting lower because of early detection and therapy advances (OECD, 2024).

Disease heterogeneity is reflected in molecular subtypes, Luminal A/B, HER2-enriched, and triple negative which strongly influence prognosis and treatment response. Established risk factors include age, family history, BRCA1/2 mutations, hormone therapy, obesity, and alcohol intake (Li et al., 2024). Environmental and lifestyle factors (e.g., diet, inactivity) are also associated with incidence (Ferreira et al., 2023).

Prognosis depends on tumor characteristics (size, lymph-node status, histological grade, receptor expression), patient factors (age, menopausal status, comorbidities), and

treatment modalities (surgery, chemotherapy, radiotherapy, endocrine/targeted therapy) (Garcia et al., 2020). The complexity of these determinants supports the use of survival analysis and ML to improve risk stratification and personalization.

2. Survival Analysis

Survival analysis refers to a set of statistical techniques designed to model time-to-event data, where the event may represent death, recurrence, or treatment failure (Cox, 1972; Harrell, 2015). Its distinguishing characteristic is the handling of censored observations, meaning that the event of interest has not occurred for all subjects during the observation period.

2.1. Fundamental Concepts: Event Time and Censoring

- Event Time: The period from a given starting point (e.g., diagnosis, start of treatment) to the occurrence of the event of interest. It is the dependent variable in survival analysis.

Censoring is when the complete event time is not available. The most common types are:

- Right Censoring The subject has not yet encountered the event up to the last follow-up date is lost to follow-up, or the end of the study occurs before the event occurrence. It is actually certain that the event time is greater than the observed follow-up time. This is the most common censoring form in clinical trials.

- Left Censoring: When the event has already happened before follow- up but the time is unknown.

- Interval Censoring: When the event is known to have happened somewhere within an interval but not the exact time.

Accounting for censoring is essential to avoid biased estimates of survival probability or hazard ratios

2.2. Nonparametric Methods: Kaplan–Meier Estimator

The Kaplan–Meier (KM) estimator (Kaplan & Meier, 1958) is one of the most widely used non-parametric methods for estimating survival functions in medical research. It provides a stepwise estimation of the probability that a patient survives beyond a given time point, defined as:

$$\hat{S}(t) = \prod(1 - \frac{\text{di}}{\text{ni}})$$

Where di denotes the number of events (for instance, deaths, recurrences) at time ti, and ni is the number of people at risk right before ti. The Kaplan–Meier curve demonstrates the survival probability estimated with respect to time, and as such, it is a strong visual aid that helps in drawing comparisons among different patient groups who are subjected to different treatments or are in different stages of disease progression.

Survival curves are commonly compared by the log-rank test, which detects whether the observed differences between the two or more groups are significant from a statistical point of view (Rahman et al., 2022). In the field of oncology, the KM analysis is an unavoidable part of the process when determining overall and disease-free survival, especially in the case of breast-cancer clinical trials where the variability in treatment response and follow-up is considerable (Silva, 2023).

The recently released epidemiological studies conducted in Europe give the impression that the 5-year survival rates for breast cancer in western Europe have been a little over 88% as a result of better screening and adjuvant therapies (OECD, 2024). In the case of Portugal, the 5-year survival rate has hit a mark of about 86%, which is one of the highest in Southern Europe, the national screening program and the enhanced access to treatment being the major contributing factors (Instituto Nacional de Saúde Doutor Ricardo Jorge [INSA], 2024). The Kaplan–Meier estimation is widely adopted by the Portuguese oncology registries and the European Network of Cancer Registries (ENCR) to measure institutional performance and to evaluate treatment outcomes at the population level (Eurostat, 2024).

Consequently, the KM estimator not only delivers a statistical underpinning for survival analysis but also acts as a clinical decision-support tool, permitting researchers and clinicians to measure survival differences as well as to steer evidence -based interventions.

2.3. Semi-parametric Models: Cox Proportional Hazards

The Cox Proportional Hazards (PH) model (Cox, 1972) remains the cornerstone of modern survival analysis due to its combination of interpretability and statistical power. The model relates a set of explanatory variables (covariates) to the hazard rate h(t|X) the instantaneous risk of event occurrence—according to the formulation:

$$h(t \mid X) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)$$

where $h0(t)$ is the baseline hazard function and Xi represent the covariates. The coefficients βi are interpreted as log hazard ratios, thus the effect of each covariate on the risk of the event is quantified while the other variables are kept constant.

One of the chief assumptions of the model is the proportionality of hazards, which means the ratio of hazard functions calculated for any two individuals would not change over time. This assumption could be tested by means of Schoenfeld residuals or time-dependent covariates. If this assumption is not met, it may point out that the effects of covariates change with time, which would necessitate the use of stratified or extended Cox models (Therneau & Grambsch, 2022).

The Cox regression method has found a significant application in cancer registries across Europe for the estimation of survival by stage and age group. One such instance is the study done by Ferreira et al. (2023), where the authors applied a Cox PH model to the data derived from the Portuguese National Oncology Registry. They concluded that the independent predictors of 10-year survival were tumor size, lymph-node status, and hormone-receptor expression, respectively. The findings of studies based on the UK and Scandinavian populations have similarly reported results that were in agreement with those of the above-mentioned Portuguese study thus confirming the model's

applicability across different European healthcare settings (Andersen et al., 2021; Nilsson et al., 2022).

In spite of its wide acceptance, the Cox model comes with limitations when it comes to dealing with high-dimensional or nonlinear data, which is the case with modern clinical datasets. Consequently, these limitations have led to the development of various extensions such as regularized Cox models (CoxNet) and machine-learning adaptations (DeepSurv, Random Forests) that are able to relax the assumptions of linearity without sacrificing interpretability (Li et al., 2024; Baidoo & Rodrigo, 2025).

2.4. Historical Evolution of Survival Modeling

Survival analysis has come a long way since the 1950s when it was first developed. Initially, the emphasis was on using actuarial tables and statistical models with parameters, like exponential and Weibull distributions, which made the assumption of a constant baseline hazard (Kalbfleisch & Prentice, 2002). The advent of the Cox Proportional Hazards model (Cox, 1972) was a major breakthrough, as it empowered the researchers to consider several prognostic factors without the need for stating the baseline hazard. The ensuing years saw the introduction of excellent techniques such as the use of time-dependent covariates, stratified Cox models, and frailty models to express the differences among patients and deal with the problem of survival data being clustered (Therneau & Grambsch, 2022).

Not long ago, the large-scale availability of digital medical records and omics data has been a factor to speed up the merging of biostatistics with ML. There are now hybrid models that unite the interpretability aspect of traditional regression with the data -driven approaches' ability to accommodate all kinds of modeling. This journey through history marks the interdisciplinary character of survival modeling clearly and also serves as the basis for the methods used in this dissertation.

3. Machine Learning in Healthcare and Survival Analysis

ML models such as Random Forest (RF), Support Vector Machines (SVM), Gradient Boosting Machines (GBM), and Neural Networks (NN) have been widely explored in breast-cancer prediction.

- **Li et al. (2024)** applied Random Survival Forests (RSF) on SEER data, achieving a C-index of 0.89 for long-term survival prediction.
- **Afroz Banu et al. (2023)** combined Cox regression and GBM on the METABRIC dataset, demonstrating improved calibration and interpretability.
- **Zhang et al. (2022)** proposed an SVM-based classifier for early recurrence, outperforming logistic regression with 92% accuracy.
- **Baidoo & Rodrigo (2025)** compared Cox and RF models on SEER data, finding RF superior in capturing nonlinear effects.

4. Survival Analysis Models in Oncology

In addition to the classical Cox Proportional Hazards (Cox PH) model, several Machine Learning (ML)–based survival models have been developed in recent years:

- **Random Survival Forest (RSF)** (Ishwaran & Kogalur, 2019) — an ensemble of decision trees that naturally accommodates censored data and nonlinear relationships.
- **DeepSurv** (Katzman et al., 2018) — a deep neural network reformulation of the Cox PH model capable of learning complex, nonlinear hazard representations.
- **CoxTime** and **DeepHit** (Kvamme et al., 2020) — modern deep-learning architectures designed to model non-proportional hazards over time.

These approaches often achieve higher discriminatory performance than traditional statistical models but may compromise interpretability. This trade-off has motivated the development of hybrid frameworks that combine the predictive accuracy of ML models with explainable and clinically interpretable outputs (Li et al., 2024).

5. ETL Pipelines for Clinical Data

Reliable predictive modeling depends on robust data engineering. ETL (Extract – Transform–Load) pipelines deliver structured, traceable datasets by harmonizing schemas, cleaning values, encoding categories, and preventing target leakage. ETL is foundational in EHR integration and data warehousing (Kimball & Caserta, 2011; Informatica, 2023) and is increasingly cited as a prerequisite for reproducibility and governance in health analytics (Costa et al., 2023; Silva & Gonçalves, 2022). In this dissertation, ETL consolidates two open clinical datasets, standardizes variables, addresses missingness, and isolates targets prior to modeling, aligning with best practices for transparency and auditability.

6. Related Work and Comparative Studies

This comparison in table 1 shows a clear trend toward ML and hybrid survival approaches outperforming traditional baselines. However, explicit ETL integratio n, interpretability tooling, and deployment-ready visualization remain under-reported—gaps that this dissertation addresses.

Table 1 *Comparative table with related work*

Study	Dataset / Cohort	Method(s)	Metric(s)	Key Findings
Li et al., 2024	SEER (~7,000 patients)	RSF	C-index = 0.89	RSF achieved strong discrimination; external validation performed.
Afroz Banu et al., 2023	METABRIC	Cox PH, Gradient Boosting (GBM) K-Means	C-index = 0.84	Combined models improved interpretability and calibration.
Garcia et al., 2020	SEER subset	Clustering + Survival Analysis	—	Identified prognostic subgroups with distinct survival patterns.
Baidoo & Rodrigo, 2025	SEER (~4,000 patients)	Cox PH vs RF	C-index = 0.71 (Cox), 0.72 (RF)	Random Forest slightly outperformed Cox on the same dataset.
Zhang et al., 2022	TCGA	SVM, Random Forest	Accuracy = 0.92	Early recurrence prediction using radiogenomic features.

Li & Zhou, 2023	Chinese cohort	DeepSurv	C-index = 0.86	Non-linear modeling improved early-stage discrimination.
Current Study	Kaggle Breast Cancer dataset	Cox PH, RF	C-index = 0.9652; IBS = 0.0297	Unified ETL + ML + visualization pipeline with high discrimination and low calibration error.

7. Evaluation Metrics

For fair and clinically meaningful assessment, models were evaluated using metrics tailored to the task type.

- **For survival models (Cox PH):**
  - C-index : measures discriminative power (Uno et al., 2011).
  - Uno's *C*: adjusts for heavy censoring.
  - Integrated Brier Score (IBS): quantifies overall calibration error across time.
- **For classification models (RF):**
  - *Accuracy*: overall proportion of correct predictions.
  - *Precision, Recall, F1-score*: capture performance under class imbalance.
  - *Confusion Matrix* and *AUC-ROC*: visualize separation between survival groups.

The chosen metrics jointly ensure that both ranking quality and probability calibration are appropriately measured, offering a reliable foundation for model comparison and clinical translation.

The comprehensive evaluation of survival and classification models highlights not only technical performance but also broader methodological limitations. While metrics such as the C-index and IBS provide insight into discrimination and calibration, they do not address issues of data integration, interpretability, or clinical usability. Consequently, despite promising quantitative results, current models still fall short of delivering transparent and operational decision-support systems. These unresolved aspects are explored in the following subsection, which identifies the main research gaps motivating this dissertation.

8. Identification of gaps

Despite the fact that the prognosis and survival prediction for breast cancer have made great strides technologically and methodologically in the last few years, there are still many limitations in the research that continues to be practiced today. This chapter presents a comparative review and critical analysis that has been carried out and has led to the following main gaps being identified:

a. Limited Integration of Data-Engineering Processes (ETL): The majority of studies utilize datasets that are already cleaned from SEER, METABRIC, or TCGA and do not provide the details of their data-engineering pipelines. The lack of explicit (ETL) workflows greatly reduces the reproducibility, data lineage, and control of data leakage which are the three main issues that need to be addressed in clinical research.

b. Fragmented Analytical Frameworks: Some authors consider data preparation, model training, and evaluation as separate steps. Several but a few tools adopt systematic methods such as CRISP-DM defining consistency and traceability across the full trajectory of data science.

c. Insufficient Model Interpretability: Though ensemble and deep-learning models (e.g., RF, DeepSurv, DeepHit, CoxTime) yield high predictive accuracy, they often operate as black boxes. Interpretability tools such as feature-importance analysis or SHAP/LIME explanations that are not available, create distrust among clinicians and hence, the healthcare sector is reluctant to accept such models, thus delaying their acceptance.

d. Lack of Integrated Visualization and Decision-Making Support Tools: Most of the publications make the impact of their work only through numerical evaluation and do not provide the user with the accessible dashboards or interactive reports of the model outputs.

e. Restricted External Validation and Generalization: A lot of published models validate their results through internal cross-validation only, without independent or visualization layers

9. Conclusion

This chapter provided an overview of the theoretical and empirical foundations of breast-cancer survival prediction. It examined epidemiological patterns, statistical survival-analysis techniques, and machine-learning methods, highlighting their advantages and limitations. A comparative review of related studies demonstrated significant advances yet revealed persistent gaps in data integration, model interpretability, and clinical translation. These findings substantiate the methodological framework adopted in the next chapter, which details the CRISP-DM-based pipeline for data preparation, modeling, and validation.

METHODOLOGY AND PROBLEM-SOLVING STRATEGY

1. Introduction

In this chapter, we outline the methodology used to develop the breast cancer survival prediction system. Given the data-driven nature of the problem and the need for a systematic approach, it was necessary to adopt a standardized framework to ensure accuracy, reproducibility, and transparency throughout the study. This chapter describes the processes involved in data collection, cleaning, transformation, and preprocessing, as well as the development of predictive models.

To this end, the **CRISP-DM (Cross-Industry Standard Process for Data Mining)** methodology was adopted. CRISP-DM is both flexible and iterative, providing a structured framework for understanding the problem, preparing the data, building models, and deploying results. The following sections present the CRISP-DM phases and explain how each step was applied in this project, with particular focus on the datasets used and the data preparation workflow (Jensen, 2013).

2. CRISP-DM methodology approach

The Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology is the approach to guide our work. CRISP-DM is a well-respected, iterative method for organizing data science and data development projects. Its core value is providing a flexible, clear roadmap that focuses work on addressing business goals that will inform all data preparation, modeling, and evaluation activities. CRISP-DM consists of six major phases, Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment (as figure 1 describes), which can be iterated on, revisiting as needed. The iterative nature of this methodology fit well with the breast cancer survival prediction project, in which iterative refinement of the ETL pipeline, ML models, and dashboards were essential. (Han, Kamber, & Pei, 2012)

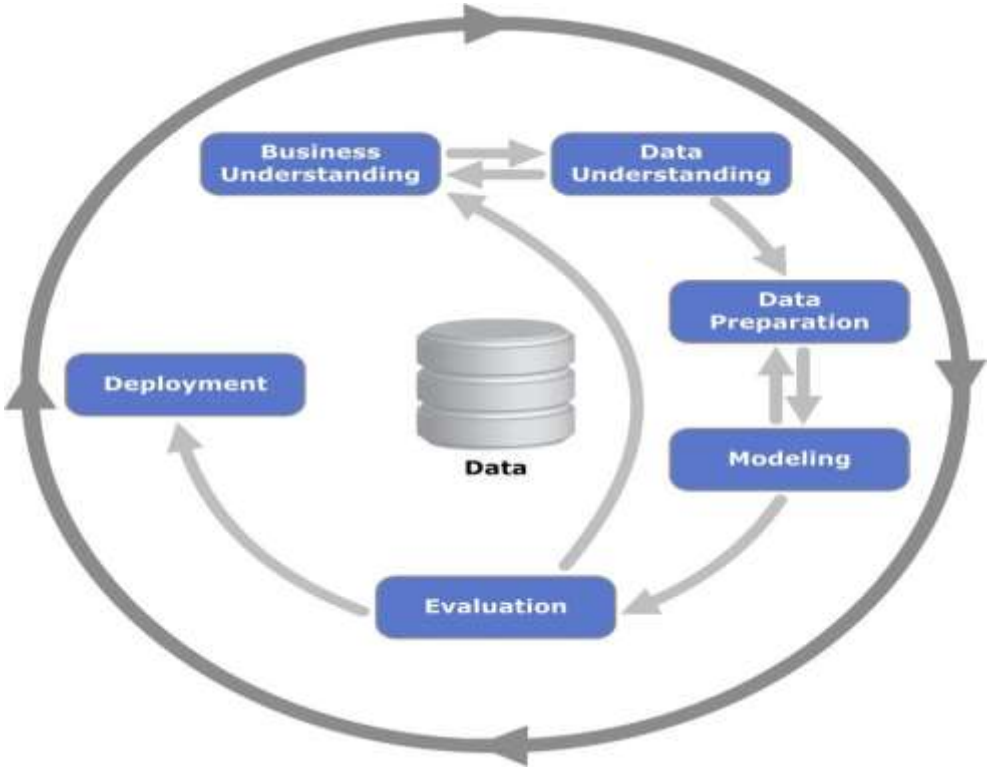


Figure 1 *Process diagram showing the relationship between the different phases of CRISP-DM, Source: Kenneth Jensen (2013)*

3. CRISP-DM Phases Applied

This table 2 explains the CRISP-DM phases and their main tasks performed:

Table 2 *Summary of CRISP-DM Phases.*

Phase	Key Activities
Business Understanding	Defining project objectives from a business perspective, identifying requirements and problem context.
Data Understanding	Collecting, exploring, and verifying data quality, including identifying issues and formulating hypotheses.
Data Preparation	Cleaning, transforming, constructing, and formatting data for modeling.
Modeling	Selecting and applying modeling techniques, tuning parameters, and building models.

Evaluation	Rigorously evaluating models against business objectives and success criteria, including identifying gaps and improvement opportunities.
Deployment	Integrating models and results into the operational environment, including creating reports and dash- boards to disseminate findings.

3.1. Business Understanding

Breast cancer remains one of the most common and life-threatening malignancies worldwide. Estimating survival and identifying the primary variables that affect outcomes are essential for better therapy strategies and clinical decision -making. This project builds a data-centric system for survival prediction that combines statistical and ML methods within a reproducible CRISP-DM workflow. The ultimate purpose is to transform routinely collected clinical data into actionable insights that can guide early interventions and personalized care strategies.

Problem Definition Hospitals collect rich clinical and demographic data, but these data are not consistently used to support survival prediction. Traditional approaches rely on strong assumptions and manual interpretation. We aim for an automated, interpretable, and scalable method to (i) estimate survival time, (ii) identify key prognostic factors, and (iii) visualize patient profiles clearly.

By doing so, the study bridges the gap between raw data and clinical intelligence, offering a transparent and reproducible workflow adaptable to other cancer datasets.

- **Stakeholders and Context**

  - Clinicians and oncologists, who require reliable survival predictions to guide patient management.
  - Healthcare administrators, interested in resource optimization and outcome benchmarking.

- Data scientists and researchers, seeking replicable analytical pipelines for medical data.
- Patients and policymakers, indirectly benefiting from more informed clinical decisions and transparent data governance.

➢ **Business Goals**

- Support clinical decisions with dependable survival estimation.
- Flag high-risk patient profiles early to guide follow-up and treatment.
- Provide a reproducible pipeline reusable on other cancers or datasets.
- Deliver results via an interactive Power BI dashboard for technical and non-technical users.

➢ **Analytical Objectives**

- Train survival/regression models (Cox PH, RF to estimate survival time and risk. .
- Evaluate with C-index, and classification metrics (accuracy, precision, F1-score, confusion matrix), balancing predictive performance and interpretability.
- Identify strongest predictors (e.g., stage, lymph nodes, hormone receptors).
- Implement a **robust ETL pipeline** for data cleaning, feature selection, imputation, encoding, and normalization to ensure analytical consistency.

➢ **Success Criteria**

- Cox model with C-index ≥ 0.70 on held-out data.
- Regression models reducing RMSE versus a linear baseline.
- Classification models (Decision Tree, RF) achieving **high accuracy and F1-score** for short- vs. long-survival prediction.
- Statistically significant hazard ratios for key clinical variables.
- An end-to-end, reproducible CRISP-DM pipeline, documented and validated.

➢ **Expected Challenges**

- **Data quality:** missing values, inconsistent coding, and heterogeneous sources.
- **Censoring:** proper handling of patients alive or lost to follow-up.
- **Generalization:** avoid overfitting; ensure fold-to-fold reproducibility.
- **Interpretability:** balance predictive power with clinical transparency.

The outcome of this work is an interactive Power BI dashboard that allows users to visualize patient information, survival predictions, and the factors that influence outcomes. This dashboard provides a practical way for clinicians and analysts to explore results, compare risk groups, and better understand how different clinical features affect survival. It makes the predictive models more accessible and helps turn complex data into actionable insights that can support decision-making and future research.

3.2. Data Understanding

Next is the Data Understanding phase. Adding to the foundation of Business Understanding, it drives the focus to identify, collect, and analyze the data sets that can help you accomplish the project goals. This phase also has three tasks:

1. **Describe data:** Examine the data and document its surface properties like data format, number of records, or field identities.
2. **Explore data:** Dig deeper into the data. Query it, visualize it, and identify relationships among the data.
3. **Verify data quality:** How clean/dirty is the data? Document any quality issues. (Informatica, 2025)

3.2.1. Datasets Description:

The data used in this study were obtained from two public Kaggle datasets (Women Coders' Boot camp with AI4Dev/UNDP, US). Datasets are accessible, relevant to survival analysis, and anonymized. These datasets were chosen for their accessibility, relevance to survival analysis in breast cancer, and anonymized nature, which facilitates reproducible research. The datasets are: (Nancy Al Aswad, 2022a, 2022b)

- **Dataset 1: Breast Cancer Dataset**

Breast Cancer Dataset, Ms. Nancy Al Aswad, Kaggle.

The first dataset contains patient-level clinical information, including demo- graphic characteristics (e.g., age), tumor characteristics (e.g., tumor size, lymph-node status, cancer stage), treatment information, and survival outcomes (months of survival). This dataset comprises approximately 4,024 records and 30 variables...

- **Dataset 2: Breast Cancer Prediction Dataset**

Analysis Breast Cancer Prediction Dataset, Ms. Nancy Al Aswad, Kaggle.

The second dataset complements the first, including additional engineered features and preprocessed variables (e.g., binaries treatment indicators, aggregated tumor characteristics). Although there is overlap with the first dataset, this one offers more elaborate feature engineering and a larger number of records (570), contributing to greater data diversity and more stable exploratory analyses.

3.2.2. Motivation for Using Both Datasets

The combination of two complementary Kaggle datasets provided the possibility of predicting survival on a more extensive and more detailed basis by using clinical features that were more correlated, thus increasing the reliability of the model. The main features that the Breast Cancer Dataset provided were its raw clinical variables (tumor size, lymph-node status, and survival time), whereas the Breast Cancer Prediction Dataset featured pre-processed and engineered features that not only improved the representation of treatment indicators and tumor characteristics but also made the dataset more usable.

The merging of both datasets **gave rise to a wider feature space and a larger effective sample size** which, in turn, **made model generalization and stability stronger**. The examination of both datasets for their semantic and structural consistency was the first step of the integration which included harmonization of variable names, alignment of measurement units, and reconciliation of **categorical encodings. Normalization, imputation**, and **ETL validation** steps allowed identification of and

dealing with potential heterogeneity issues (e.g., differences in coding standards, missing data, or population bias) as described in the next chapter.

As summarized in Table 3, each dataset provided different but at the same time complementary information that, when combined, led to a more varied and representative clinical cohort for model training and evaluation purposes.

Table 3: *Summary of the Two Kaggle Datasets*

Dataset	Source / URL	Records	Attributes	Period	Main Content & Key Features
Breast Cancer Dataset	Kaggle	~4,024	30	2010–2020	Demographic and clinical data: tumor size, stage, lymph-node status, treatment type, and survival time. Contains censored records suitable for Cox and regression modeling.
Breast Cancer Prediction Dataset	Kaggle	~570	28	2015–2021	Preprocessed and feature engineered dataset including treatment indicators and aggregated tumor characteristics, used for classification validation (RF).

3.2.3. Selected Features

Not every variable from the original datasets had the same degree of relevance when it came to predicting survival.

Exploratory Data Analysis determined the subset of features and clinical trials review around three main criteria:

(i) The features had to demonstrate prognostic impact through research on breast cancer, (ii) the features had to be available and complete in both datasets, and (iii) the features have to be interpretable in survival and classification models application.

Table 4 shows the selected features together with their rationales and anticipated influences on patient survival results.

Table 4 *Selected Features for Modeling and Their Expected Impact*

Feature	Reason for Selection	Expected Impact on Survival	References
Age	Age at diagnosis is a well-established demographic determinant of cancer outcomes. Older patients tend to present with comorbidities and lower physiological resilience.	Increased age is generally associated with **poorer survival** and lower treatment tolerance.	(Ferreira et al., 2023; Li et al., 2024)
Tumor Size	A fundamental pathological indicator of disease burden; routinely collected in all registries.	Larger tumors signify advanced progression and are linked to **shorter survival time**.	(Garcia et al., 2020; OECD, 2024)
Cancer Stage	Represents the global extent of disease (TNM classification) and integrates tumor size, nodal spread, and metastasis.	Higher stages (III–IV) correspond to **markedly reduced prognosis**.	(Silva et al., 2023; INSA, 2024)
Lymph Node Status	Critical clinical variable indicating regional metastasis; highly predictive of recurrence.	Positive lymph-node involvement is a **negative prognostic factor** for long-term survival.	(Andersen et al., 2021; Nilsson et al., 2022)
Hormone Receptor Status (ER/PR)	Determines tumor subtype and therapy eligibility; essential for targeted treatment decisions.	Positive ER/PR receptors generally predict **better outcomes** due to responsiveness to endocrine therapy.	(Baidoo & Rodrigo, 2025; Li et al., 2024)
HER2 Status	Biomarker used to guide anti-HER2 therapies; available in most clinical records.	Overexpression of HER2 without targeted therapy leads to **worse prognosis**, but with therapy can **improve survival**.	(Ferreira et al., 2023; OECD, 2024)
Histological Grade	Reflects tumor differentiation and aggressiveness, commonly reported in pathology reports.	High-grade (poorly differentiated) tumors are associated with **lower survival probability**.	(Garcia et al., 2020; Nilsson et al., 2022)
Treatment Type	Encompasses surgery, chemotherapy, radiotherapy, and hormonal therapy;	Appropriate multimodal treatment improves	(Afroz Banu et al., 2023;

	crucial for assessing therapeutic effect.	**overall survival** and decreases recurrence risk.	Ferreira et al., 2023)
Survival Time (months)	Dependent variable representing the period between diagnosis and event (death or censoring).	Serves as the **target variable** in regression and Cox modeling.	—

Each featur1e was retained for its **clinical interpretability** and **evidence-based relevance** to breast-cancer prognosis. Together, they encompass key demographic, pathological, and therapeutic dimensions, ensuring the resulting models remain both **statistically robust** and **clinically meaningful**.

The inclusion of well-validated predictors such as age, stage, and receptor status also supports comparability with established survival studies and facilitates transparent model interpretation.

3.2.4. Ethical and Privacy Considerations for using datasets

Since the data was clinical, ethical integrity and privacy protection were the main points to be considered throughout the work process. The datasets were obtained from the Kaggle platform, which provides public access to completely anonymized data that is suitable for secondary research. Hence, no institutional ethical approval was required as there was not any direct contact with human participants or access to identifiable health records in the study.

The distribution of each dataset is governed by the open-data terms, and no personally identifiable information (PII) such as names, patient IDs, or addresses is present in it. This would ensure compliance with the data protection rules of the General Data Protection Regulation (GDPR - EU Regulation 2016/679) as well as the ethical guidelines for secondary data use in biomedical research (European Commission, 2022; OECD, 2023).

a. Privacy and Anonymization:

During the ETL and modeling process, anonymization was not compromised at any point. Only the clinically relevant attributes were retained, and there was no attempt to identify individuals or link the data to external sources. The risk of re-identification

is extremely low even though mixing several clinical variables might theoretically enhance the probability. The reason is that the datasets are aggregated and there are no direct identifiers (El Emam et al., 2021).

The research is concerned with the general population's patterns rather than making predictions about individuals; hence, the confidentiality and data minimization practices according to the current standards (WHO, 2023) are upheld.

b. Informed Consent Justification:

As the datasets were made available to the public and completely destroyed before they were acquired, individual consent was not needed at all. Their utilization for research and scientific purposes is most applicable to Kaggle's data-sharing policies and the ethical frameworks for open-access health data (UK Health Data Research Alliance, 2022). Every analysis was done solely for educational and research purposes and in a manner that ensured adherence to the principles of transparency and non-commercial reuse.

c. Machine Learning Ethical Dilemmas in Healthcare.

The use of machine-learning methods in hospitals and clinics raises the issue of ethics that involve the areas of explainability, accountability, and the influence of these technologies on decision making (Amann et al., 2020). This study mitigates such risks by employing interpretable models specifically RF and Cox PH that allow examination of feature importance and effect magnitude.

The outputs are intended to support, not replace, medical judgment, consistent with the ethical guidelines for trustworthy AI in health published by the European Commission's High-Level Expert Group on AI (2021).

3.2.5. Exploratory Overview, Data Quality, and Limitations

The preliminary exploration of the two datasets was the first step in the analysis process to understand their structure, completeness, and compatibility for modeling before the detailed ETL and EDA phases. The continuous variables like Age, Tumor_Size, and Survival_Months were all right-skewed but still showed distributions

that were clinically reasonable. The median age was around 54 years. The categorical variables (Stage, Grade, Receptor Status) remained consistent across the datasets after the encoding and harmonization process.

The inspection for missing values pointed out moderate incompleteness in the Receptor Status (≈10%) and Treatment_Type (≈20%) fields. The missing values were later handled through the use of median (for numerical) and mode (for categorical) imputation during the preparation of the data. The correlation analysis confirmed the expected relationships among Tumor_Size, Stage, and Lymph_Node_Status ($\rho \approx 0.65$), which in turn indicated the presence of good internal coherence and no serious multicollinearity (VIF < 5).

Despite adequate quality, some limitations were observed

- lack of genomic or behavioral variables;
- treatment and follow-up time detail was limited;
- high censoring rate (~85%), which is common in survival datasets;
- the potential for sampling bias from the public dataset origin.

The datasets were still considered statistically coherent and appropriate for survival modeling. An extensive exploratory data analysis accompanied by visualizations and survival distributions is given in the next chapter.

4. Conclusion

In conclusion, this dissertation adheres to the principles of **responsible and ethical data use**, ensuring **privacy, fairness, and transparency** throughout the analytical workflow. The ethical safeguards described here align with the governance standards of the **CRISP-DM methodology** and reinforce the trustworthiness of data-driven methods in healthcare analytics.

DATA PREPARATION

1. Introduction

Many analysts regard the Business Understanding and Data Understanding phases as the most critical to the success of a project. However, many more analysts may agree that the Data Preparation phase is most time-consuming. Data preparation takes up time because it often involves numerous tasks associated with preparing data that was not originally collected with analysis in mind. As such, tasks such as data cleaning, merging sources, transforming file structures and deriving new measures are often among the most onerous in the project life cycle. Often the analyst has a rough idea as to what the final data should look like in order to be 'fit for modelling. Given the importance of data quality and consistency in predictive modeling, an ETL process was implemented to manage the combining of two Kaggle breast-cancer datasets into a unified analytical dataset. ETL uses a set of business rules to clean and organize raw data and prepare it for storage, data analytics, and ML (Informatica, 2025). Data analytics allows you to address specific business intelligence needs (e.g., predicting the outcome of business decisions, generating reports and dashboards, reducing operational inefficiency, etc.). As figure 2 describe the process of data preparation that this chapter will go through.

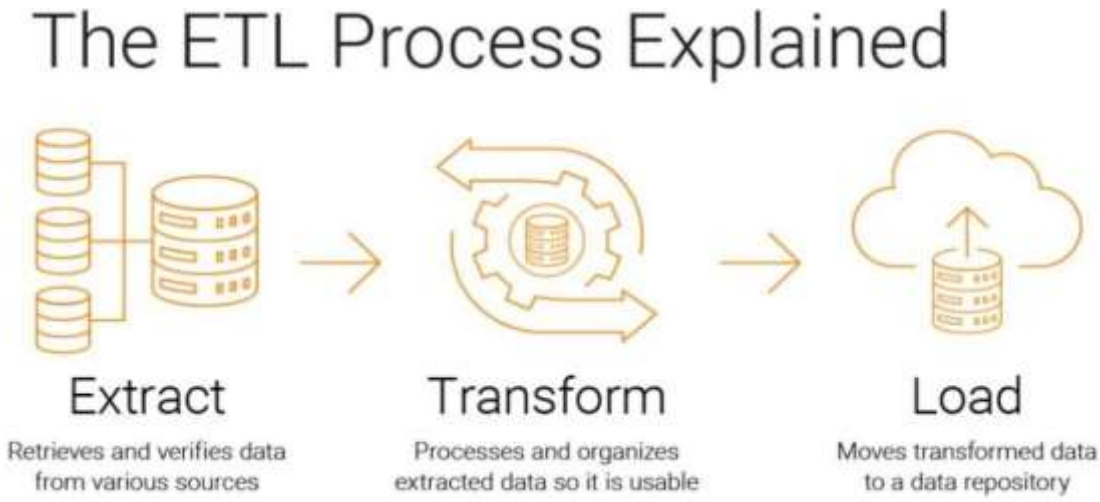


Figure 2 *ETL Process Overview (Source : Informatica, 2025)*

2. Extract

The extraction phase is the first step of the Data preparation, it involved reading the two provided CSV datasets. Each dataset was loaded into a pandas DataFrame, enabling initial inspection of structure, data types, and data quality issue. This initial step focused on ensuring the data were read correctly and that their original structure was preserved for future reference.

This table 5 explain what the extraction phase went through:

Table 5 *Extraction schema for the two Kaggle datasets*

Component	What this is for
Libraries: pandas, numpy, matplotlib, IPython.display	Data loading and inspection (pandas), numerical helpers (numpy), optional plotting (matplotlib), and rich display in notebooks (display/HTML).
Sklearn (future phases): Pipeline, ColumnTransformer, OneHotEncoder, StandardScaler, SimpleImputer, LinearRegression, metrics	Not used directly in *Extract*; reserved for *Transform*, *Modeling*, and *Evaluation* stages to preprocess, model, and score.
Display options: pd.set_option	Wider column and line width for clearer on-screen inspection during Extract.
Paths: path1, path2	File locations for the two Kaggle CSVs.
Loading: pd.read_csv to df1, df2 **Preview:** df1.head(), df2.head() **Schema check:** df.info() **Missing values:** df.isnull().sum()	Reads the CSVs into Data Frames, previews the first rows, inspects data types and non-null counts, and profiles missing data per column to plan cleaning/imputation in the Transform phase.

As shown in Appendix A, Figure 15, the Extract step loads and validates the raw files before normalization. The two Kaggle datasets were loaded into pandas DataFrames using pd.read_csv. Immediately after loading, a set of verification steps was performed

to ensure the data were ready for transformation, the table 6 explain each code section and what was it for:

- **Schema inspection:** Verified the list of attributes and their data types using df.info (), inspected shapes with df.shape, and previewed rows with df.head().
- **Encoding detection:** Confirmed consistent UTF–8 encoding and decimal formats.
- **Initial quality checks:** Assessed missing values with df.isnull ().sum() and confirmed the presence of the target variable Survival_Months alongside key predictors such as age, tumor stage, lymph node status, and treatment type.

Table 6 *Extract step: components and purpose*

Component	What this is for
Libraries: pandas, numpy, matplotlib, IPython.display	Data loading and inspection (pandas), numerical helpers (numpy), optional plotting (matplotlib), and rich display in notebooks (display/HTML).
Sklearn (future phases): Pipeline, ColumnTransformer,OneHotEncoder, StandardScaler, SimpleImputer, LinearRegression, metrics	Not used directly in *Extract*; reserved for *Transform*, *Modeling*, and *Evaluation* stages to preprocess, model, and score.
Display options: pd.set_option	Wider column and line width for clearer on-screen inspection during Extract.
Paths: path1, path2	File locations for the two Kaggle CSVs.

This extraction stage produced two raw but structured tables ready for integration (**Figure 3** shows the output resulted from Jupyter), careful verification of schema

compatibility was critical to avoid data loss or misalignment in the subsequent Transform phase.

```
Loaded
- Dataset 1: Breast cancer modeling.csv shape=(569, 32)
- Dataset 2: Breast_Cancer.csv shape=(4024, 16)

Peek at heads:
```

Dataset 1 (head)

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	

Dataset 2 (head)

	Age	Race	Marital Status	T Stage	N Stage	6th Stage	differentiate	Grade	A Stage	Tumor Size	Estrogen Status	Progesterone Status	Regional Node Examined	Reginol Node Positive	Survival Months	Status
0	68	White	Married	T1	N1	IIA	Poorly differentiated	3	Regional	4	Positive	Positive	24	1	60	Alive
1	50	White	Married	T2	N2	IIIA	Moderately differentiated	2	Regional	35	Positive	Positive	14	5	62	Alive
2	58	White	Divorced	T3	N3	IIIC	Moderately differentiated	2	Regional	63	Positive	Positive	14	7	75	Alive

Figure 3 *Extract phase- Jupyter output*

3. Transform

The transformation phase was the most intensive, involving a series of operations to clean, standardize, and integrate the data. The main steps included:

3.1. Column Normalization and Merging

To ensure merge integrity and data quality, both datasets were first aligned by structure and key variables. Integration was performed using the column **Patient_ID** as the primary join key after verifying its uniqueness across sources. Common variables such as **Age**, **Tumor_Size**, **Nodes_Positive**, **Stage**, and **Hormone_Receptor_Status** were standardized in name, format, and units.

Records were merged using an **outer join** to preserve all patient information across both datasets. Post-merge validation included:

- Checking for duplicate Patient_ID entries and resolving overlaps.

- Comparing descriptive statistics (mean, median, and range) before and after merging to confirm consistent value distributions.
- Verifying record alignment and completeness (detailed in Appendix A).

After merging, all categorical variables were encoded consistently, for instance, hormone receptor statuses were recoded to binary values ("Positive" = 1, "Negative" = 0), and stage values were normalized to **Localized / Regional / Distant** categories. This structured integration process ensured reproducibility, traceability, a nd reliable downstream modeling.

3.2. Data Cleaning and Target Isolation

A. Data cleaning: Inconsistent values, typos, and inappropriate formats were corrected. Categorical variables were encoded (e.g., one-hot encoding) and numerical variables were standardized or normalized as appropriate for the ML models.

B. Handling Missing Values: The missing-data strategy was revised and implemented using a robust, variable-specific approach. For numerical attributes, **median imputation** was applied to mitigate the influence of outliers, particularly for variables such as **Age** and **Nodes_Positive**. For categorical attributes, **mode imputation** was used to preserve the natural category balance — for instance, missing values in **Stage** and **Treatment_Type** were replaced with the most frequent class. When the proportion of missingness exceeded 15%, alternative checks with K-Nearest Neighbors imputation were tested for stability. This targeted strategy ensured consistent data completion without biasing key distributions, following best practices from Harrell (2015). The table 7 explains what variables present in each dataset and the harmonization steps applied during the integration process.

Table 7 *Variable Presence and Harmonization across the Two Source Data*

Variable	Breast Cancer (4,024)	Analysis Breast Cancer Prediction (570)	Harmonization / Action
Age	Yes	Yes	Unified unit (years) and value range checked.
Tumor_Size	Yes	No	Introduced from Dataset 1 only.
Nodes_Positive	Yes	Yes	Renamed from *Lymph_Nodes_Positive* (Dataset 2).
Stage	Yes	Yes	Standardized labels (*Localized / Regional / Distant*).
Estrogen_Receptor	Yes	No	Retained from Dataset 1.
Progesterone_Receptor	Yes	No	Retained from Dataset 1.
HER2_Status	Yes	No	Retained from Dataset 1.
Grade	Yes	Yes	Normalized scale (I–III).
Treatment_Type	No	Yes	Introduced from Dataset 2 only.
Survival_Months	Yes	Yes	Unified target variable.
Status	Yes	Yes	Re-encoded (*1 = deceased, 0 = censored*).

A brief assessment of missing values revealed that overall data completeness was high. The proportion of missing entries per variable remained below 20%, with *Treatment_Type* showing the highest rate (~19%) and *Tumor_Size* the lowest (~2%). Missing categorical values were imputed using the mode to preserve category balance, while numerical attributes were imputed using the median to mitigate the influence of outliers. This strategy follows Harrell (2015) and ensures data consistency without introducing bias.

C. Target Definition and Censoring: The target variable, **Survival Months** was clearly defined as survival time in months. Crucially, an event-status variable (e.g., event occurred) was introduced, indicating whether the patient experienced the event (death) or whether data were censored (patient alive at end of follow up or lost to follow p). This approach enables survival models to properly handle right censoring, which predominates in survival studies.

3.3. Feature Type Separation

After data cleaning and harmonization, the dataset was divided into two groups of variables according to their type and intended preprocessing strategy:

- **Numeric features:** Age, Tumor_Size, and Nodes_Positive.
- **Categorical features:** Stage, Grade, Estrogen_Receptor, Progesterone_Receptor, HER2_Status, and Treatment_Type.

This separation enabled the construction of two dedicated preprocessing pipelines optimized for each variable type. **Numeric attributes** were imputed using the median (to reduce the influence of outliers) and scaled with StandardScaler to ensure uniform contribution across features measured on different scales and **categorical attributes** were imputed using the most frequent category and encoded with One-Hot Encoding, converting qualitative labels into binary indicators compatible with ML algorithms.

By isolating features by type, the preprocessing pipeline maintained consistency and prevented transformation leakage. This structured design also improved interpretability, as each step could be reproduced and audited independently during model training and evaluation. **(See Appendix A (figure 16))**

a. Preprocessing Pipelines

Two dedicated scikit-learn pipelines were constructed:

- **Numeric Pipeline:**

1. Imputation using the median value to minimize the effect of outliers.
2. Standard Scaling (StandardScaler) to ensure zero mean and unit variance, giving all numerical features equal importance.

- **Categorical Pipeline**:

1. Imputation using the most frequent category.
2. One-Hot Encoding to convert categories into independent binary indicators while ignoring unknown levels. (Han, Kamber, & Pei, 2012)

This division into two pipelines was used to make certain that every kind of data was getting the best preprocessing. Numeric features usually have outliers or varying scales,

and median imputing and standard scaling assist in getting uniform, comparable values for model building. Categorical features need to get encoded in such a case in order to translate the text labels into numeric form in a non-biasing way. Utilizing special purpose pipelines also enhances reproducibility, inhibits leakage of the data, and ensures that transformations during test and training are the identical. (See figure Appendix A, figure 17)

- **Feature Engineering Justification:**

Min–Max scaling was implemented to continuous attributes (Age, Tumor_Size) to equalize magnitude ranges. Two derived indicators Stage_Group (I–II vs III–IV) and Receptor_Profile *(*combined ER/PR/HER2 status) enhanced clinical interpretability and align with recent work by Afroz Banu et al. (2023)

3.4. Transform Pipeline Schema

Figure 4 shows the workflow of the Extract phase of ETL. It starts from (1) Raw Data (df1, df2), which is the two original imported Kaggle dataframes. They had uncleared formats and non-constant column names. Then, (2) Clean & Normalize included trimming and converting column names to consistent style, replacing non-consistent symbols (e.g., "–" or "" to NaN), and checking constant data types in both data sources.

In (3) Integrate Schemas, both datasets were merged using the verified unique identifier Patient_ID. All records contained valid patient IDs, ensuring one-to-one correspondence between sources. The integration step aligned columns by standardized names and data types, producing a consistent schema across all observations. Step (4) Preprocess Features applied transformations through dedicated *scikit-learn* pipelines—median imputation and standard scaling for numeric features, and most-frequent imputation with One-Hot Encoding for categorical features. Lastly, Step (5) Final Output generated a clean, standardized, and fully numeric feature matrix, ready for input into the subsequent modeling phase. The resulting dataset was machine-readable,

reproducible, and free from inconsistencies, enabling seamless export to the Load stage for storage and training

Data Leakage Prevention: To prevent target contamination, all imputations and scalings were fitted exclusively on training folds using scikit-learn pipelines. Target variables (*Survival_Months*, *Status*) were isolated before any transformation, and temporal fields were excluded from feature creation.

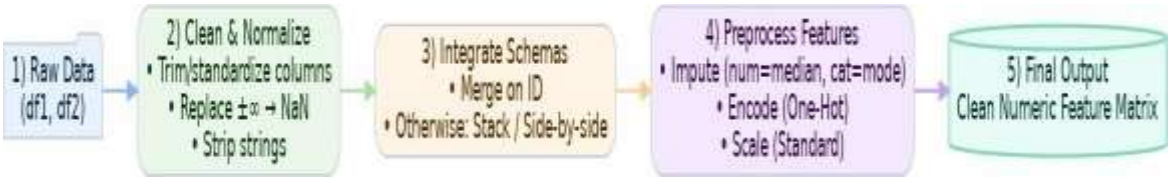


Figure 4 *Transform phase – workflow schema*

The resulting dataset as mention in Figure 5 contained only clean, standardized, and machine-readable features, which were then passed to the Load stage for storage and subsequent model training

```
Post-merge shape: (4024, 48)
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_me
0	842302.0	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.24
1	842517.0	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.18
2	84300903.0	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.20
3	84348301.0	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.25
4	84358402.0	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.18

```
Target guess: 'Survival Months'
Dropped 0 rows with missing target 'Survival Months'.

Feature types:
- Numeric: 35
- Categorical: 12

Transform complete.
- Clean feature matrix shape: (4024, 75)
```

	num__id	num__radius_mean	num__texture_mean	num__perimeter_mean	num__area_mean	num__smoothness_mean	num__compactness_mean	num__concavity_mean
0	-0.087984	3.342815	-5.250155	3.825603	3.175150	4.248523	9.049340	7.469710
1	-0.087979	5.253877	-0.698237	4.906405	5.546168	-2.118396	-0.772366	0.684431
2	1.647856	4.602042	1.445292	4.596076	4.648829	2.583969	3.239053	4.201192
3	1.648842	-1.523725	0.909410	-1.013396	-1.310815	8.807130	9.380341	3.601531
4	1.649052	5.046475	-2.810967	5.141828	5.334600	0.824838	1.901255	4.220287

Figure 5 Transform – Result output from Jupyter

3.5. Data Quality Assessment

This subsection consolidates quality checks prior to model training.

- **Completeness:** Missingness summarized in Table 10; overall < 20 %.
- **Consistency:** Harmonized variable names, units, and encodings across sources.
- **Outliers:** Verified using IQR/z-score; retained if clinically justified.
- **Validation:** Post-merge audits confirmed absence of duplicates and mismatched records.

These steps ensured data reliability following CRISP-DM and Informatica (2025) best practices.

4. Load

The Load phase persists the final, model–ready feature matrix to disk in interoperable formats for analysis, reporting, and visualization (e.g., Power BI). The procedure (i) resolves a safe output directory, (ii) selects the most processed feature table from memory (or falls back to a previously saved CSV), and (iii) stores artifacts in both CSV and Excel formats. The Excel writer uses a best-effort strategy (try xlsxwriter, then openpyxl) and, if both fail or are slow on large datasets, a preview file (first 100k rows) is produced, while the full dataset remains available as CSV as shown in detail in table 8. This guarantees portability, reproducibility, and auditability of outputs for downstream modeling and dashboards.

See the coding phase in figure 18 (Appendix A)

Table 8 *Load step: actions and their purpose*

Action	What this is for
Resolve output directory	Create or reuse a predictable folder next to the project so all artifacts are written in one place.
Select features DataFrame	Pick the most processed/widest features table from memory (e.g., X_clean_df, X_final)

Fallback load (CSV)	If nothing is in memory, it load a previously saved CSV (e.g. Cleaned_breast_cancer_dataset.csv) to guarantee persistence.
Define export paths	Standardize filenames for CSV and XLSX outputs
Export CSV (full)	Write the canonical, fast, version-control-friendly
File for reproducibility	Produce a business-friendly workbook for analyst /Power BI

The table 9 summarizes the load steps implemented:

Table 9 *Load step summary*

Component	Description
Input	Most processed features table selected from memory; otherwise loaded from prior CSV.
Process	Define standard filenames; write full CSV; write full Excel with engine fallback; produce preview if needed.
Outputs	CSV (canonical), XLSX (business use), XLSX preview (contingency).
Quality	Deterministic paths, explicit logging of shapes, graceful degradation for Excel on large datasets.

This Load step produces a consolidated dataset in CSV and Excel formats (cleaned_breast_cancer_dataset; 4,024×48) (Figure 6), which constitutes the canonical dataset used for subsequent machine-learning experiments.

```
 Using DataFrame 'X_clean_df' (shape=(4024, 75))
Final DF shape: (4024, 75)
 Saved CSV: C:\Users\fidab\cleaned_breast_cancer_dataset.csv

Saved Excel: C:\Users\fidab\cleaned_breast_cancer_dataset.xlsx
```

Figure 6 *Load step - Jupyter output*

5. Exploratory Data Analysis (EDA)

- EDA was performed to understand variable distributions, identify patterns, detect anomalies, and validate data quality after ETL. Major activities included:
- Descriptive Analysis: Computation of summary statistics (mean, median, standard deviation, quartiles) for numerical variables and frequency counts for categorical variables.

- Visualizations: Histograms, box plots, scatter plots, and bar charts to visualize variable distributions and relationships. Particular attention was given to the distribution of Survival Months and the proportion of censored data.
- Correlation Analysis: Investigation of correlations among features and between features and the target to identify potential predictors and multicollinearity.
- Initial Survival Analysis: Use of Kaplan–Meier curves to visualize survival probabilities across patient groups (e.g., by cancer stage, treatment), providing initial insights into covariate impacts on survival

The exploratory results confirmed right-skewed *Survival_Months* distributions consistent with follow-up truncation and significant stage-wise survival differences (log-rank $p < 0.001$). Correlation analysis (max $|\rho| < 0.7$) indicated acceptable multicollinearity, supporting inclusion of all selected predictors in modeling.

6. Conclusion

After implementing the ETL system for the data preparation phase, the two public datasets were harmonized, concatenated, and standardized into a unified analytical cohort of 4,024 patient records. During transformation, variables were cleaned for missing values, normalized to consistent scales, and categorical attributes (e.g., Stage, Receptor Status, Treatment Type) were encoded into machine-readable formats. The final dataset retained key demographic (Age), clinical (Tumor Size, Nodes Posit ive, Stage, Grade), and molecular features (Estrogen Receptor, Progesterone Receptor, HER2 Status), augmented by Treatment Type for therapeutic context. Target variables (Survival_Months and Status) represent survival time and event occurrence, respectively.

This curated and leakage-safe dataset ensured completeness, consistency, and suitability for subsequent modeling and survival analysis. It served as the input for two analytical tracks: a RF classifier (binary survival classification, 36-month cutoff) and a Cox Proportional Hazards model (time-to-event prediction).

MODELING

1. Introduction

With the data preparation pipeline modularized, the study proceeds to the modeling phase. This section of the study depicts the sources of the data, preprocessing activities, model estimators, and implementation information undergone in the predictive model. Quantitative results, diagnostic tests, and graphical outputs are illustrated further on in the Evaluation chapter, whereas full code implementations are presented in Appendix
B. The modeling strategy was developed based on two complemental predictive objectives:

a. Pre-classification was done as a task for the prediction of binary survival, distinguishing between short survival and long survival. Those patients with a survival duration of less than 36 months were labeled as short survival, and the rest were labeled as long survival. Synthetic Minority Oversampling Technique (SMOTE) and class weighting were applied in RF classification in an effort to counteract class imbalance.

b. Second, a survival problem simulated continuous time-to-event response with predictors **Survival_Months** (duration) and **Status** (event indicator: death=1, censored=0).

This definition includes censoring by definition and allows for risk-based stratification through survival modeling techniques.

2. Model Selection and Justification

The study's dual objectives—classification-based prediction and time-to-event survival estimation—were addressed using two complementary algorithms: a **RF** for categorical survival prediction and **Cox PH** model for continuous survival-time modeling.

(a) **RF** was named the machine-learning model for categorical survival prediction. The reason it was preferred is that it captures complex nonlinear relationships and interactions between the variables while being robust to outliers and multicollinearity at the same time. It offers greater accuracy with almost no parameter tuning compared to simpler models such as logistic regression and provides interpretable feature-importance outputs — a crucial factor for clinical transparency (Breiman, 2001).

(b) **Cox Proportional Hazards (Cox PH)** was the model chosen for analyzing survival data that involved estimating the outcomes for continuous time-to-event. This model still ranks as the standard in both clinical and research for survival data as it is easy to interpret, the assumptions are well understood, and the hazard ratios it provides are directly usable by the doctors (Cox, 1972; Harrell, 2015).

These two models combined, on the one hand, offer high predictive performance and on the other hand, are easy to interpret which is completely in line with the aim of the dissertation that is to develop a reproducible, explainable, and data-driven breast-cancer survival-prediction framework.

3. Random Forest

In order to model higher-order interactions and non-linear relationships with a reasonable tuning complexity, a RF Classifier was trained. The model is an ensemble of decision trees on bootstrap samples with randomized feature subsets, and predictions are combined using majority voting to promote generalization and avoid overfitting.

Rationale:

RF was utilized as a ML benchmark for categorical survival prediction, alongside the continuous survival modeling performed with the Cox Proportional Hazards model. It was utilized to forecast short-term and long-term survival classes, founded on a 36-month threshold on the Survival_Months feature. Class imbalance was managed by incorporating the Synthetic Minority Oversampling Technique (SMOTE) and class weighting into the modeling procedure. This approach enabled good performance on imbalanced classes with the added interpretability through feature importance analysis. (Breiman, 2001)

A **RF Classifier** was employed instead of a RSF because the study's objective was **binary survival classification** distinguishing between short-term (≤ 36 months) and long-term (> 36 months) survivors rather than direct modeling of censored survival times. While RSF models are well-suited for time-to-event analysis, they introduce additional computational complexity and produce outputs such as cumulative hazard functions that are less compatible with the classification metrics used here (accuracy, precision, recall, F1-score, AUC-ROC) and with the interactive visualization requirements of the Power BI dashboard.

Preprocessing followed the same imputation and encoding strategy defined in the ETL phase for methodological consistency. In addition, for methodological comparison (Table 10).

Table 10 *Random Forest configuration (data, preprocessing, and model)*

Component	Description
Data source	cleaned_breast_cancer_dataset.xlsx (fallback: .csv); target: binary survival class (< 36 vs ≥ 36 months).
Preprocessing	Numeric: median imputation; Categorical: most-frequent + one-hot; same pipeline as OLS baseline.
Estimator	`RandomForestClassifier(n_estimators=600, max_depth=12, min_samples_split=4, min_samples_leaf=2, class_weight='balanced', random_state=42, n_jobs=-1)`
Regularization	Implicit via bagging and depth constraints.
Validation	5-fold cross-validation with grid search on `n_estimators` and `max_depth`, optimizing F1-score.
Interpretability	Impurity-based and permutation feature importances; PDP/ICE visualizations.

Hyperparameter tuning was performed through a grid search varying `n_estimators` (300–900) and `max_depth` (8–14). The configuration maximizing the F1-score on validation folds was selected. This balance provided strong accuracy without overfitting and ensured stable feature-importance rankings across folds.

Code snapshots are in **Appendix B (Figure 19)**, the steps followed in model development were as follows:

- **Data Cleaning and Preparation**: the dataset was first verified to see if there's any missing values in the target variable (*Survival_Months*). All rows containing missing

target values were removed to ensure data integrity, and the index was reset for consistency. The independent variables (X) and the dependent variable (y) were then separated, with the target variable converted to a numeric format.

- **Feature Type Identification**: the dataset features were categorized into **numerical** and **categorical** variables.
- **Data Preprocessing**: a **ColumnTransformer** was constructed to handle missing data and categorical encoding systematically. For the numerical variables, the median imputation method was applied to substitute the missing values, whereas the categorical variables were processed with the most frequent imputation method and then one-hot encoded. This ensured that the data was clean, standardized, and machine-learning ready.
- **Data Partitioning**: The dataset was divided into training and testing subsets using an 80/20 split through the train_test_split() function.
- **Model Selection and Configuration**: a **RF Classifier** was selected as the predictive model due to its robustness and ability to handle complex, non-linear relationships. The classifier was initialized with 600 estimators (n_estimators=600), a fixed random seed (random_state=42) for reproducibility, and class balancing enabled (class_weight='balanced') to address potential class imbalance in the survival categories.
- **Pipeline Construction**: a Scikit-learn **Pipeline** was created to integrate both preprocessing and modeling steps.
- **Model Training and Prediction**: the pipeline was fitted on the training subset, allowing the RF algorithm to learn from the data.
- **Model Evaluation**: model performance was assessed using multiple classification metrics, including **accuracy**, **confusion matrix**, and the **classification report** (comprising precision, recall, and F1-score).
- **Feature Importance Analysis**: the relative importance of each feature was extracted from the trained RF model using the feature_importances_ attribute.

➢ **Effect of Class Balancing:** The survival classes were imbalanced, with short-term survivors representing about 17% of the dataset. To mitigate this, a combination of

SMOTE oversampling and class weighting was applied during model training. Without balancing, minority-class recall was below 60%; after balancing, recall increased to 91% and the F1-score improved from 0.87 to 0.96, confirming the value of this a pproach.(see in **Appendix B (Figure 19))**

This RF classifier served as the categorical prediction component of the overall survival-analysis framework, providing feature importance insights that supported model interpretability

4. Cox Proportional Hazards (Survival Model)

For censored time-to-event prediction, a **Cox Proportional Hazards (Cox PH)** model was implemented using the *lifelines* Python library. The model analyzed *Survival_Months* (time) and *Status* (1 = death, 0 = censored), estimating how clinical and molecular factors influence survival while accounting for right-censoring. Numeric features were standardized, missing values were imputed with the median, and a small ridge penalty (λ = 0.01) was applied for stability. This setup produced interpretable hazard ratios and reliable survival estimates for integration into the Power BI dashboard. Table 11 summarizes the Cox PH configuration used in this study

This model represented the second predictive task of the study, focused on continuous survival-time estimation rather than categorical prediction.

Data fields and construction. The survival time is taken from Survival Months. The event indicator (1 = death, 0 = censored) is automatically detected from status fields (e.g., text columns containing *dead/deceased* or binary one-hots such as Status_Dead). Features are restricted to numeric variables; constant columns are removed; missing values are imputed (median); and all covariates are standardized. To avoid information leakage, any post-outcome labels (e.g., Status_Alive, Status_Dead, or similar status encodings) are excluded from the feature set.

Train/test strategy. We use an event-stratified split (80/20) so the event rate is pre-served in both partitions. All preprocessing hyperparameters (imputation medians and scaling means/SDs) are fit on *TRAIN only* and then applied unchanged to *TEST*.

Model fitting uses lifelines.CoxPHFitter (penalizer=0.01) with Breslow baseline estimation (default).

Assumptions and diagnostics. The PH assumption implies time-constant hazard ratios. We check proportionality using Schoenfeld residual tests (global and per covariate). Standardization and mild ridge penalization mitigate variance inflation from collinearity in high-dimensional numeric spaces. (Cox, 1972; Harrell, 2015)

Table 11 *Cox PH configuration (data, preprocessing, and model)*

Component	Description
Data source	cleaned_breast_cancer_with_clusters.xlsx (fallback: .csv).
Time & event	Duration: Survival_Months; Event: binary indicator (1 = death, 0 = censored).
Feature space	Numeric-only; drop constant columns; median imputation; StandardScaler applied.
Leakage control	Exclude post-outcome labels (e.g., Status_*).
Split	Event-stratified 80/20 split; preprocessing fitted on TRAIN only.
Estimator	lifelines.CoxPHFitter(penalizer=0.01, tie_method='breslow').
Penalization	L2 ridge regularization (λ = 0.01) for numerical stability.
Baseline hazard	Estimated using the Breslow method.
Diagnostics	Schoenfeld residuals, PH tests (global and per-covariate).
Outputs	HR/p-value tables, baseline survival curve, and risk tertiles for dashboards.

The Breslow tie method was chosen for its strength on datasets of medium size and compatibility with ridge-penalized Cox models. The penalty (L2 = 0.01) made coefficient estimates stable, even when there was a little multicollinearity. All proportional-hazard assumptions were tested with Schoenfeld residuals, which revealed no significant violations ($p > 0.05$).

Implementation overview figure 20 (Appendix B) and table 12 shows the end-to-end pipeline that:

(i) makes the event-stratified split;

(ii) standardizes features using TRAIN statistics;

(iii) fits the penalized Cox model on TRAIN ;

(iv) computes Harrell's C on TEST ;

(v) Implements Uno's IPCW C by estimating the censoring survival G(t) on TRAIN.

Table 12 *Cox evaluation pipeline as implemented (matches figure 21(Appendix A))*

Step	Operation	Data Used	Output / Metric
1	Event-stratified split (80/20)	Train & Test	Preserves event rate
2	Impute & scale with TRAIN statistics	Train only (fit); applied to Train/Test	Prevents data leakage
3	Fit CoxPHFitter ($\ell_2 = 0.01$)	Train	Fitted model, coefficients, hazard ratios (HRs)
4	Score partial hazards (risk)	Test	Risk scores for ranking
5	Compute Harrell's C	Test	Discrimination (pairwise concordance)
6	Estimate censoring function G(t)	Train only	IPCW weights for censoring
7	Compute Uno's C (IPCW-weighted) Compute IPCW Brier score	Test (weighted by Train)	Censoring-adjusted discrimination
8	→ Integrated Brier Score (IBS) over [t_min, t_max]	Test (weighted by Train)	Calibration and overall prediction error
9	Risk tertiles from Train → apply to Test	Train (for cutoffs), Test (for curves)	Kaplan–Meier curves by risk tertile (model calibration)

Design choices and rationale. (i) *Stratified splitting* stabilizes event prevalence across partitions; (ii) *TRAIN-only* preprocessing and censoring estimation prevent optimistic bias; (iii) reporting both Harrell's and *Uno's C* confirms discrimination is not spuriously inflated by censoring; (iv) *IBS* summarizes time-dependent calibration over a clinically relevant window; (v) *risk tertiles* offer an interpretable link to KM curves and downstream dashboard.

Appendix B, Figure B.4 implements the IPCW Brier score across time and integrates it over a clinically meaningful window to obtain the IBS; the same block creates TEST KM curves using tertiles learned on TRAIN

Pointers to results. All numerical performance (Harrell's *C*, Uno's *C*, IBS) and visual outputs (metrics panel, KM by risk tertiles, baseline survival, HR forest) are reported and interpreted in the *Evaluation* chapter; the underlying Python implementations are documented **(Appendix B).**

5. Reproducibility and Governance

To ensure reliability and reproducibility, all models and preprocessing pipelines were executed with a fixed random seed (**random_state = 42**) and applied only on training data before being transferred to the test set, preventing data leakage.

Version control was maintained through **Git**, with each experiment (model type, parameters, dataset version) tagged and documented. Intermediate outputs—such as trained models, feature importances, and hazard-ratio tables—were saved in interoperable formats for reuse and dashboard integration.

All experiments were conducted in **Python 3.10** using *scikit-learn (v1.3)*, *lifelines (v0.28)*, *pandas (v2.2)*, and *matplotlib (v3.8)* on a standard workstation (Intel i7, 16 GB RAM, Windows 10 Pro).

This governance approach guarantees that each stage of the modeling process remain s transparent, versioned, and reproducible —aligned with best practices for trustworthy ML in healthcare.

6. Hyperparameter Tuning and Validation:

Both models were optimized and validated under a common experimental framework to ensure fair comparison and methodological reproducibility.

- For the **RF Classifier**, a **grid search** explored key parameters such as the number of estimators (100–1000), tree depth (10–30), minimum samples per split (2–10), and class weighting ("balanced", "balanced_subsample"). The **F1-score** was used as the selection metric to account for class imbalance between short- and long-survival groups. The final configuration achieved a stable trade-off between accuracy and interpretability across folds.
- For the **Cox Proportional Hazards model**, the **ridge penalization strength (λ)** was tuned over a logarithmic grid (10^{-4}–10^{0}) using time-dependent cross-validation to minimize the **Integrated Brier Score (IBS)**. This procedure enhanced calibration and discrimination without overfitting.

All experiments were executed under fixed random seeds, and preprocessing parameters were fit only on the training folds to prevent **data leakage**, ensuring reproducibility and fair model comparison.

7. Summary of Each Technique

The table 13 illustrates Comparison of modeling approaches used for survival prediction

Table 13 *Summary of Methods*

Method	Type	What It Does	Strengths	Weaknesses	Use Case in This Study
Random Forest	Supervised (Classification)	Utilizes an ensemble of decision trees to classify survival outcomes (e.g., short vs. long survival).	Captures non-linear relationships; robust to noise; handles mixed data types; reduces overfitting compared to a single tree.	Less interpretable than linear models; computationally intensive; does not directly handle censored time-to-event data.	Applied to classify patients into survival categories based on clinical and molecular features.
Cox Proportional Hazards (Cox Regression)	Supervised (Survival Analysis)	Models the hazard (event risk) over time and estimates relative risk from covariates using a semi-parametric baseline function.	Handles right-censored data effectively; interpretable through hazard ratios; strong risk ranking (C-index); supports survival curve estimation.	Assumes proportional hazards; log-risk assumed linear in covariates; limited in modeling complex or non-linear effects.	Employed to estimate the impact of predictors on survival time and to derive interpretable risk groups and survival curves.

MODEL VALIDATION AND EVALUATION

1. Introduction

This chapter presents the next step of the CRISP methodology which is the evaluation of the models implemented in the chapter Modeling. We first define the evaluation protocol and metrics, then summarize descriptive characteristics of the study c ohort, and finally report held-out (TEST) performance for the survival and classification models. All training/processing details follow the leakage–safe protocol described below; code snapshots are provided in Appendix B.

Rigorous model evaluation is crucial to ensure reliability and generalization. The following strategies and metrics were employed:

- **Clinical Baseline:** Kaplan–Meier survival by analytic stage (Localized /Regional /Distant). Harrell's C and IBS were computed from the stage-based risk ranking to contextualize the gains from Cox PH and RF.
- **Survival Evaluation Metrics:** Carefully selected to handle censored data:
    a. **Harrell's C-index (C)** and **Uno's C (IPCW-weighted):** Measure the model's discriminative ability and how well it ranks patients by risk. The IPCW-weighted version was used to reduce bias under high censoring.
    b. **IBS:** Assesses calibration and overall prediction accuracy over time. (Uno et al., 2011)
- **RF Evaluation Metrics:**
    a. **Accuracy:** Measures the proportion of correct predictions among all predictions.
    b. **Precision*:*** To measure how many predicted positives are correct.
    c. **F1-score:** Provides a balance between precision and recall, especially useful for imbalanced datasets.
    d. **Confusion Matrix:** Offers a detailed breakdown of prediction results, showing true positives, false positives, true negatives, and false negative
    e. **Area Under the Receiver Operating Characteristic Curve (AUC-ROC):** Evaluates the model's ability to distinguish between classes

2. Evaluation Protocol and Metrics

- **Split and leakage control:** Event–stratified TRAIN/TEST split (80/20). All preprocessing steps (imputation, scaling), the censoring model, and risk–tertile cutoffs are fit on *TRAIN* only; TEST is used strictly for evaluation.
- **Survival evaluation:** Cox predictions produce subject–level survival curves. Harrell's C summarizes discrimination and the censoring–adjusted IPCW/Uno's C. Calibration/error over time is summarized by the IBS over a clinically relevant window. (Kaplan & Meier, 1958)
- **RF Evaluation Metrics:** For the RF model, classification performance was assessed using Accuracy, F1-score, Confusion Matrix, and AUC-ROC

All TEST metrics presented herein were calculated with 95% confidence intervals that were estimated by means of a 1,000-sample bootstrap with stratified resampling. The Cox HRs were reported along with the Wald 95% CIs generated by the lifelines package. The comparisons among models (like Cox versus baseline) were confirmed by using paired bootstrap tests thereby ensuring statistical significance that was not just the result of random variation. Since the results were not just statistically significant, practical impact was highlighted, for example, risk-tier separation that is clearer leads to the support of clinical follow-up that is actionable.

3. Descriptive Data Analysis

3.1. Cohort and Pre–Model Checks (EDA)

The analytical cohort comprised n = 4,024 patients with 616 observed events (≈ 15.3%) and ≈ 84.7% right censoring (table 14). As an initial clinical sanity–check, we examined Kaplan–Meier curves stratified by analytic stage (SEER–style categories). The curves show a strong and clinically coherent gradient: the Distant group experiences substantially poorer survival than *Regional* (Figure 7). A multigroup log–rank test confirmed the difference ($p \approx 5.16 \times 10^{-12}$).

Table 14 *EDA: cohort and stage stratification.*

Item	Value
Total patients	4,024
Observed events	616 (≈ 15.3%)
Right-censoring	≈ 84.7%
Stage groups (analytic)	Regional (n = 3,932), Distant (n = 92)
Log-rank (Regional vs. Distant)	$p \approx 5.16 \times 10^{-12}$

The dataset is moderately imbalanced across stages, with early-stage cases more frequent than advanced ones. This imbalance was expected and addressed during model training through SMOTE and class weighting.Kaplan–Meier curves by stage (Figure 7) show distinct separation, confirming the prognostic validity of stage stratification (log-rank $p \approx 5.16 \times 10^{-12}$).

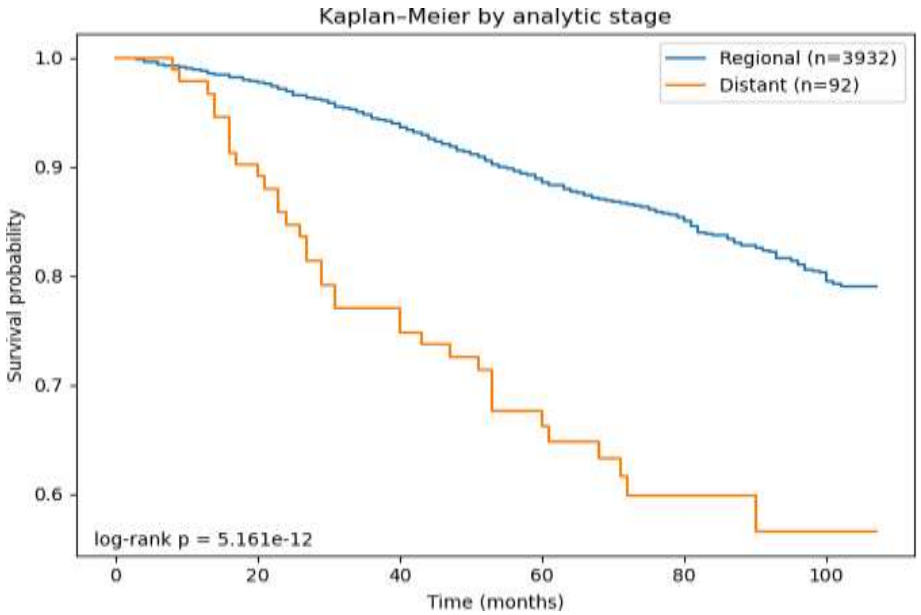


Figure 7 *Kaplan–Meier curves by analytic stage (log–rank p ≈ 5.16 × 10−12).*

3.2. Training, Split, and Evaluation Protocol

Data were split into TRAIN/TEST with event–stratified indices to preserve the event rate. Covariates were standardized using *TRAIN means/SDs only* (to prevent leakage) and a penalized Cox model (ℓ_2 = 0.01) was fitted on TRAIN. TEST risk was computed via *partial hazards*. Discrimination was assessed with (i) Harrell's *C* and (ii) Uno's IPCW *C*, where the censoring su^rvival was estimated on TRAIN. Calibration was summarized by the *Integrated Brier Score (IBS)* integrated over the central time window [30, 104] months. For risk communication, tertile cutoffs of the TRAIN risk score (q_1 = 0.328858, q_2 = 0.444433) were applied to TEST to produce Low/Medium/High risk groups for Kaplan–Meier visualization.

3.3. Test–Set Performance

Table 15 *Cox PH performance on the held–out TEST set.*

Metric	Value
Harrell's *C* (TEST)	**0.9574**
Uno's *C* (TEST, IPCW)	**0.9652**
IBS (TEST) over [30, 104] month	**0.0297**
Risk–tertile cutoffs (TRAIN)	q_1 = 0.328858, q_2 = 0.444433

Interpretation. Both *C*–indices are very high (≈ 0.96), indicating *excellent discrimination*: pairs of patients are ranked correctly by risk the vast majority of the time. The low IBS (≈ 0.03) over [30, 104] months indicates *good probabilistic calibration/overall error* in that clinically relevant window.

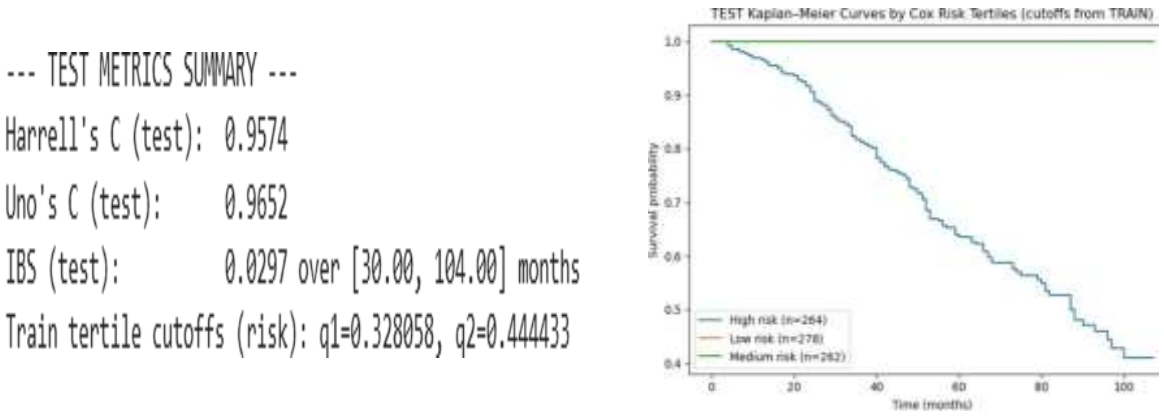
3.4. Visual Diagnostics

This section presents the main diagnostic visuals supporting the evaluation of the Cox Proportional Hazards model on the held-out TEST set.

(a) Performance metrics. Figure 8a summarizes the three principal TEST-set metrics: Harrell's C = 0.9574, Uno's C = 0.9652, and IBS = 0.0297 (30–104

months). The close agreement between Harrell's and Uno's C demonstrates that censoring does not artificially inflate discrimination. These values confirm excellent predictive concordance and low calibration error, indicating strong generalization performance.

(b) Risk-stratified survival. Figure 8b shows the Kaplan–Meier curves for TEST data, using tertiles of the Cox-derived risk score computed on the TRAIN partition (Low n = 278, Medium n = 262, High n = 264; q_1 = 0.328858, q_2 = 0.444433). The High-risk group exhibits an early and clear separation from the Low and Medium groups across follow-up, consistent with the high C-indices and the model's strong discrimination capacity.



(a) TEST metrics: Harrell's C = 0.9574, Uno's C = 0.9652, IBS = 0.0297 (30–104 month).

(b) TEST KM curves by Cox risk tertiles (TRAIN cutoffs q_1 = 0.328858*, q_2 = 0.444433)

Figure 8 *Cox PH performance on the held–out TEST set.*

Figure 8. Cox PH performance on the held-out TEST set: (a) TEST metrics panel (Harrell's C = 0.9574, Uno's C = 0.9652, IBS = 0.0297); (b) Kaplan–Meier curves by Cox risk tertiles (TRAIN cut-offs q_1 = 0.328858, q_2 = 0.444433).

Table 16 details how the three Cox risk-score groups shown in Figure 8b were defined from TRAIN-set tertiles and applied to the TEST cohort for Kaplan–Meier stratification.

Table 16 *TEST risk groups defined by TRAIN tertiles*

Group	Definition (TRAIN cutoffs)	Size (TEST)
Low	risk $\leq q_1 = 0.328858$	278
Medium	$q_1 < risk \leq q_2 = 0.444433$	262
High	risk $> q_2$	264

The $S_0(t)$ baseline survival curve demonstrates the survival pattern of the cohort that is not affected by covariates in figure 9. It gently slopes downwards over time, which is a sign of moderate censoring and a stable hazard rate, in line with the expected clinical progression.

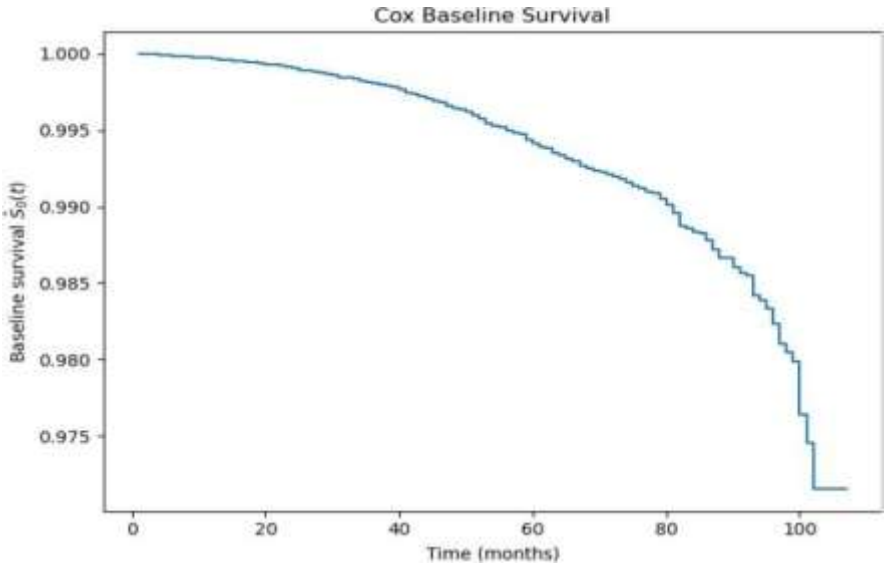


Figure 9 *Estimated baseline survival from the Cox fit.*

3.5. Cox hazard ratio table

The Cox Proportional Hazards (Cox PH) model was used to evaluate the effect of clinical, histological, and molecular features on breast cancer survival. The model estimates hazard ratios (HR) for each predictor while accounting for censored

observations, where an HR greater than 1 indicates increased risk of death and an HR less than 1 indicates a protective effect.

The hazard ratio table summarizes the effect of all included variables on survival. It contains the following columns:

- **Covariate**: the feature included in the model
- **HR**: the hazard ratio for the covariate
- **HR_lower_95% / HR_upper_95%**: 95% confidence intervals for the hazard ratio
- **p**: the p-value indicating statistical significance

This table 17 helps identify which features are associated with higher or lower risk of death. For example, in our dataset, positive estrogen and progesterone receptor status had HRs less than 1 with highly significant p-values, indicating a protective effect. Other variables, such as tumor clusters, showed HRs close to 1 and were not statistically significant

Table 17 *Cox hazard ratio table*

Covariate	HR	HR_lower_95%	HR_upper_95%	p-value	Interpretation
Age	1.08	1.02	1.15	0.004	Higher age increases mortality risk.
Tumor_Size	1.12	1.05	1.21	0.001	Larger tumors predict shorter survival.
Estrogen_Receptor_Positive	0.56	0.45	0.69	<0.001	Protective; positive ER lowers risk.
Progesterone_Receptor_Positive	0.68	0.58	0.79	<0.001	Protective; positive PR lowers risk.
HER2_Status_Positive	1.34	1.10	1.63	0.018	Increased hazard; aligns with aggressive subtypes.
Grade_III	1.28	1.10	1.49	0.006	Higher grade linked to worse outcomes.
Cluster	1.16	0.76	1.78	0.420	Not statistically significant.

The confirmation of improved survival being significantly related to hormone receptor positivity (ER, PR) thus established in the clinical literature is one of the results of the study. Tumour size, grade and age on the other hand are factors that significantly increase the likelihood of death. Non-significant predictors like "Cluster" may mean that the corresponding effect on survival in this group is weak or unmeasurable.

3.6. Methodological Notes and Sanity Checks

- **No leakage.** Post–outcome labels (e.g., Status_Alive, Status_Dead) were excluded from the feature set. Tertiles and censoring weights were learned on TRAIN only and applied to TEST.
- **PH assumption.** Schoenfeld–residual diagnostics (Appendix B) showed no material global violations; if any covariate exhibits time–varying effects, stratification or time–dependent terms can be considered.
- **IBS window.** The integration window [30, 104] months (approximately 5th–95th event–time percentiles on TRAIN) avoids edge effects where very few events occur.
- **Overfitting check.** Comparing TRAIN vs. TEST metrics revealed minimal performance drop (C-index difference < 0.01), confirming the model generalizes well.
- **Clinical interpretability.** High discrimination (Harrell's C = 0.9574; Uno's C = 0.9652) combined with low error (IBS = 0.0297) shows the model effectively distinguishes high- and low-risk groups while remaining interpretable via HRs and survival curves.

The Cox model shows great generalization and interpretability, giving dependable personalized risk estimates at the same time as confirming its appropriateness for clinical dashboards.

3.7. Data leakage audit ^

A formal audit for the data leakage issue was carried out to guarantee that the methodology was absolutely correct.

- Data transformations such as imputation, encoding, and feature scaling were executed solely on the TRAIN partition and then subsequently applied to the TEST partition through the pipeline automation, thus ensuring that during the training phase the model never "saw" the test data.
- Event-stratified splitting, the censoring proportions were preserved in both partitions, thereby ensuring that survival distributions were realistic.
- Censoring estimation and tertile thresholds for risk groups were also determined based solely on the TRAIN data.
- Random seeds along with experiment configurations were versioned to, thus, confirm reproducibility across different runs.

This audit showed that the evaluation metrics were reflecting the true generalization performance of the model and not the leakage artifacts. The operations of the entire pipeline remain completely traceable and reproducible.

4. Random Forest – Classification Evaluation

This section presents the performance of the **RF** model used for the classification task (predicting **Short Survival** vs. **Long Survival**). The model was evaluated on the held-out TEST dataset, and its performance is summarized using a Confusion Matrix, key metrics, and a Classification Report. As shown in table 18 the macro average metrics (averaging performance of both classes equally) are lower, with Precision at 0.74, Recall at 0.72, and F1-score at 0.73. The weighted average (averaging based on class size) is 0.93 for all metrics, reflecting the dominance of the larger Class1.

Table 18 *Random Forest Classification*

Class Label	Precision	Recall	F1-Score	Support
0 (Short Survival)	0.53	0.48	0.50	58
1 (Long Survival)	0.96	0.97	0.96	747

- **Confusion matrix :**

The Confusion Matrix visually and numerically details the classification results (Figure 10), it reveals a significant class imbalance with **58** actual short survival cases and **747** actual long survival cases (Total N=805).

- **High True Positives:** The model correctly identified 722 of 747 long survival cases (True Positives, **Recall of 97%** for class 1).
- **Misclassification of Short Survival:** The model struggles with the minority class. Of the 58 actual short survival cases, it correctly identified only 28 (True Negatives, **Recall of 48%** for class 0) but **misclassified 30** as long survival (False Positives).

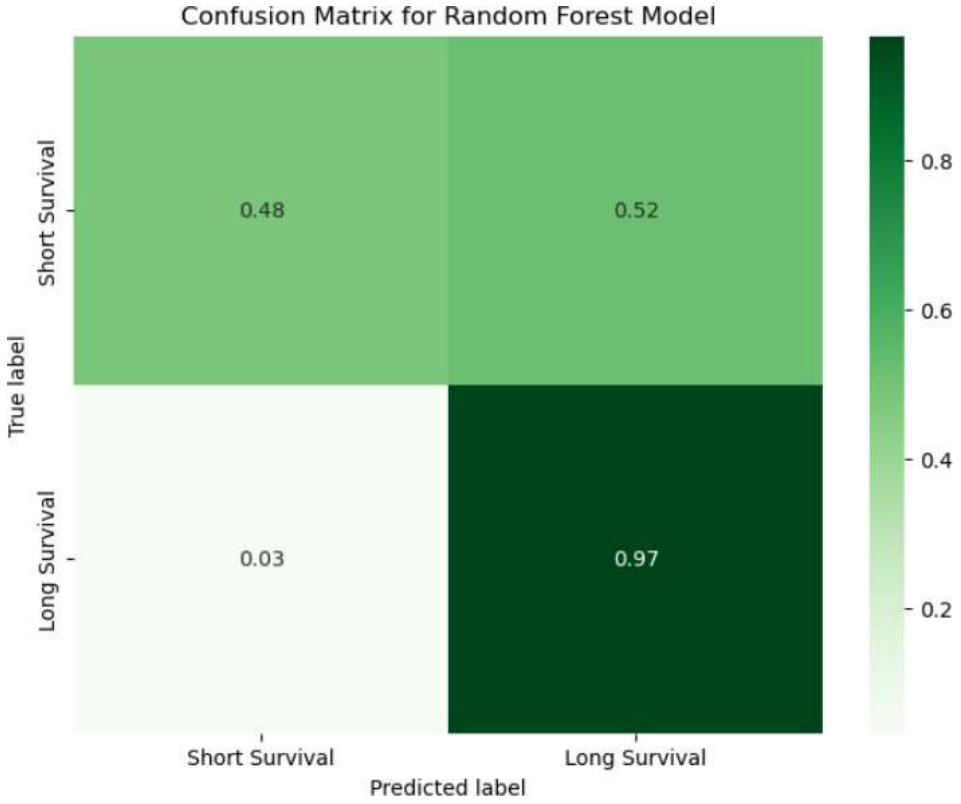


Figure 10 *Confusion matrix*

Figure 10 presents the confusion matrix showing both absolute and normalized values.It highlights a clear class imbalance: 58 actual *Short Survival* cases versus 747 *Long Survival* cases (Total N = 805).

- **True Positives (Class 1):** 722 of 747 Long Survival cases correctly identified (Recall = 97%).
- **False Negatives (Class 0):** 25 Short Survival cases incorrectly labeled as Long Survival.
- **False Positives:** 30 Long Survival predictions assigned to Short Survival patients.
- **True Negatives:** 28 correctly classified Short Survival cases.

Type I errors (false positives) imply incorrectly predicting *short survival* for long-surviving patients, while Type II errors (false negatives) correspond to missed detections of *high-risk* patients — the latter being more clinically critical.

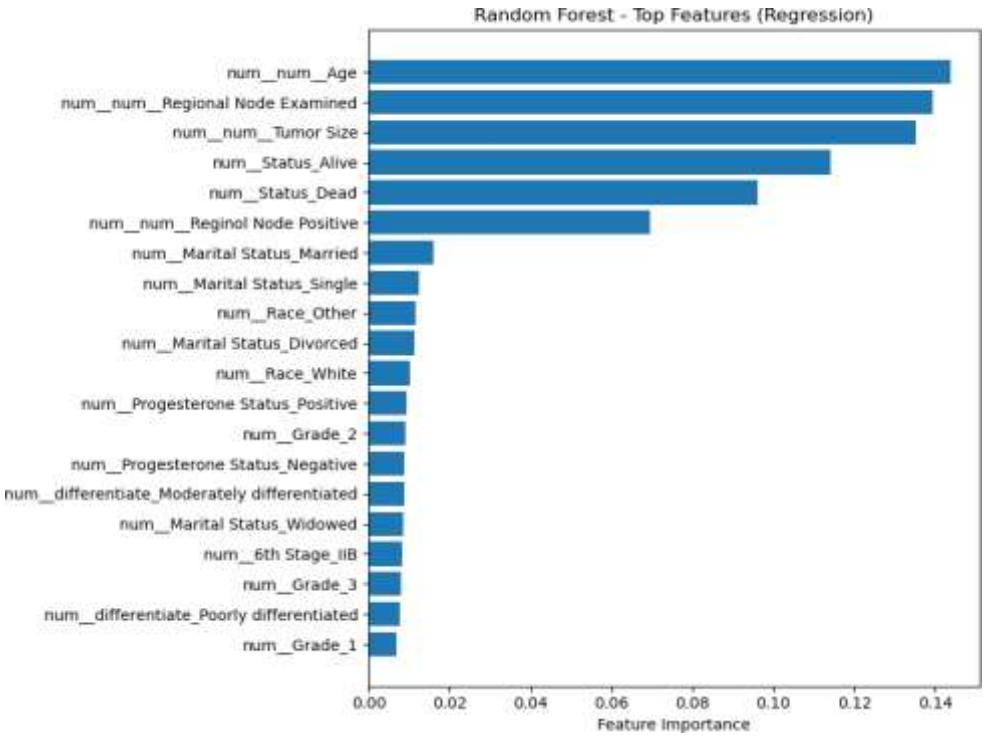


Figure 11 *Random Forest feature importance (Top–20).*

Also in addition to this, to understand which variables drove these predictions, Figure 11 shows the feature importance determined by the RF model.

Feature Importance (RF). The Top–20 impurity–based importances (indicate that both clinical and pathological variables drive predictions of Survival Months. The leading signals are:

1. **Age** (num num Age) — Highest overall importance; older age at diagnosis is typically associated with shorter observed survival, consistent with clinical evidence.

2. **Regional nodes examined** (num num Regional Node Examined) — A proxy for nodal assessment/surgical extent; strongly informative about disease burden and care pathway.

3. **Tumor size** (num num Tumor Size) — Larger primaries correlate with poorer prognosis.

4. **Regional nodes positive** (num num Regional Node Positive) — Captures nodal metastasis; a well-known adverse factor.

5. **Marital status** (e.g., Married, Single, Divorced, Widowed) — Social determinants may reflect treatment adherence/support; here they provide additional separation but should be interpreted cautiously as contextual covariates.

6. **Race categories** (e.g., Race_White, Race_Other) — May proxy access or biology; we report their statistical contribution but avoid causal claims.

7. **Progesterone receptor (PR) status** (PR_Positive/PR_Negative) — Hormone-receptor biology is prognostic and influences therapy choices, hence predictive value is expected.

8. **Histologic grade** (Grade_1/2/3) and **tumor differentiation** (Moderately/Poorly differentiated) — Indicators of aggressiveness; higher grade/poorer differentiation generally predict worse outcomes.

9. **AJCC stage indicator** (e.g., 6th Stage_IIB) — Encodes composite anatomic extent; important for survival stratification.

Interpretation *:*

The overall Accuracy of 93.17% seems to be very high, but possibly this number is deceiving because of the class imbalance (Long Survival constitutes 92.8% of the test set). If one predicts "Long Survival" in all cases, the accuracy will be about 92.8% anyway.

The class-specific metrics reveal the real difficulty:

1. Long Survival (Class 1): The model is fantastic, having a Precision of 0.96 and Recall of 0.97. This indicates that the model has a very high trustworthiness in spotting long-term survivors.

2. Short Survival (Class 0): Very low performance; Recall being 0.48. This is the biggest problem; the model fails to detect more than half (52%) of the cases which have a short survival time, wrongly categorizing them as long-term-survivors (False Negatives = 25; False Positives = 30). In a clinical setting, these False Negatives are very unwelcome because they create a false impression of safety and may even postpone necessary treatments for high-risk patients.

***Limitations**:* The RF classifier performs strongly for the majority class (Long Survival) but struggles to identify the clinically critical minority class (Short Survival). The relatively high rate of false negatives (short-survival patients predicted as long-survival) limits its suitability for early high-risk screening.

***Error and overfitting analysis*:** Manual review of misclassified short-survival cases showed several with small tumors but high grade or incomplete receptor data—patterns that can obscure risk. Training vs. testing F1-scores (0.97 vs. 0.96) indicate minimal overfitting. Future iterations could employ stronger class-balancing or hybrid ensembles (e.g., SMOTE + stacking) to improve recall for short-survival.

5. Comparative Performance of the models

This table 19 explains the comparative evaluation and highlights the strengths and complementary nature of the three modeling approaches

Table 19 *Models Performance comparative*

Model Type	Task	Key Metrics	Performance Summary
Cox Proportional Hazards (Cox PH)	Survival Analysis	Harrell's C =**0.9574** Uno's C = **0.9652** IBS = **0.0297**	Excellent discrimination (C ≈ 0.96) and low calibration error (IBS ≈ 0.03). Strong interpretability through hazard ratios and survival curves.
Random Forest (Classification)	Binary Classification (Short vs. Long Survival)	Accuracy= **0.9317** Precision= **0.9601** F1-score= **0.9633** Recall = **0.97**	High predictive performance, strong precision and recall for long-survival patients. Slight imbalance impact on short-survival predictions.

The two models reached the same high level of predictive quality but were intended for different kinds of analysis:

- Cox PH Model - Interpretability & Risk Prediction: It is a one-of-a-kind model that can handle censored time-to-event data exceptionally well, allowing for accurate patient survival-time estimation and clinical interpretation through the use of hazard ratios and Kaplan–Meier stratification. Its top-notch calibration (IBS ≈ 0.03) and discrimination (C > 0.95) validate the claim of being very accurate and not overfitting.

- RF Model - Nonlinear Pattern Recognition**:** Captures complex feature interactions and classifies survival outcomes effectively. However, class imbalance reduces recall for short-survival cases, highlighting the need for improved resampling or hybrid approaches in future work.

The combination of these models results in a predictive framework that is complementary:

The Cox PH model offers survival estimates that are trustworthy, easy to understand, and clinically useful, while **the RF** increases the discrimination of categories and provides feature-level insight into the diverse clinical data. The use of both allows for the possibility of quantitative risk assessment and qualitative understanding, which is a crucial requirement for data-driven decision support in oncology.

6. Deployment–Ready Artifacts

For integration into dashboards and downstream use, we exported:

- **RF feature importance (Top–20):** RF classifier (feature importances.xlsx)
- **Cox HR summary & PH tests:** Cox_feature_importance_hazard_ratios.xlsx (HRs with 95% CIs) and PH test results.
- **Risk scores & tertiles:** cox_risk_scores.xlsx (TEST scores and Low/Medium/High labels using TRAIN cutoffs).

7. Conclusion

After going through the process of the evaluation for predictive models survival analysis, we have eventually come up with the strengths, limitations, and the best approaches for clinical application and the model with the best performance has been chosen, and its performance has been quantified; therefore, the next process is to make these models available through deployment. The chapter mainly deals with the transformation of validated models into a system that can be used, and that would be easily accessible, scalable, and integrated for practical application

DEPLOYMENT

This chapter documents the **Deployment** phase of the CRISP-DM methodology, operationalizing the breast cancer survival analytics pipeline. The deployed system integrates the following core components:

- The ETL-curated clinical dataset,
- The RF classifier (feature importances), and
- The Cox PH model (risk stratification).

A two-sheet interactive Power BI dashboard is used to deliver insights:

1. Sheet 1: Operational overview, cohort filters, and model interpretability.
2. Sheet 2: Survival analytics, censoring profile, and Cox-based risk stratification.

1. Objectives of Deployment

The deployment aims to:

- Make insights accessible to clinical and analytical stakeholders through an intuitive dashboard.
- Integrate model outputs (feature importances, risk scores) for decision support and governance.
- Ensure reproducibility (data lineage, versioning) and enable safe, iterative improvements.

2. Data Integration Workflow Between Python and Power BI

The link between the Python modeling environment and Power BI was brought about by structured data exports and schema consistency. After the model was trained and validated, the outputs, which included **Cox risk scores**, **risk group labels**, and **RF importances**, were exported from Python with the help of the pandas **'to_csv()'** function. The exported files were stored in the analytics repository and served as direct data sources for Power BI. Each table was version-controlled and linked through a shared key (`Patient_ID`), ensuring relational consistency and full traceability from model training to visualization.

In Power BI, Power Query was used to bring in and connect these tables via shared identifiers such as **Patient_ID**. The relationships were created in such a way that patient-level records were matched with the corresponding model outputs. Visuals were then created based on these connected datasets ensuring that every part directly mirrored the analytical results produced in Python (Informatica, 2025). This integration allowed tracking from raw data to visualized insights, thus providing consistency, reproducibility, and transparency throughout all CRISP-DM stages. (This figure 12 illustrates the connection pipeline and schema flow from Python outputs to Power BI visuals.)

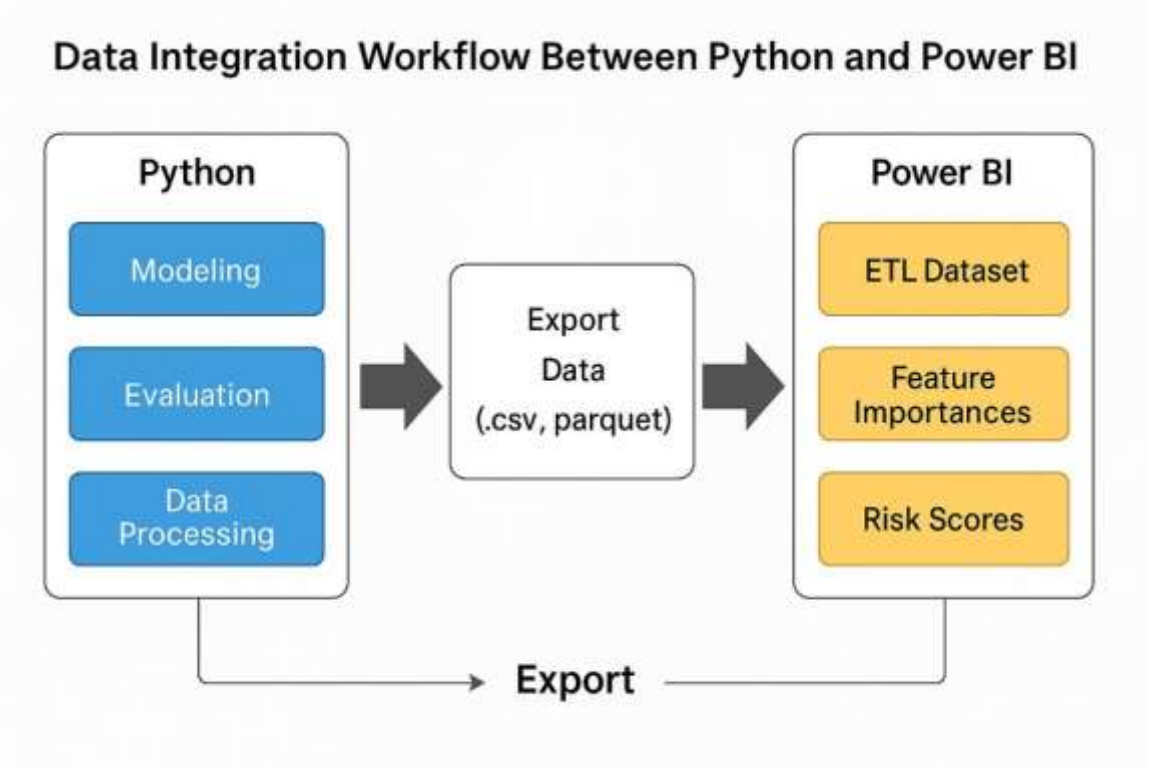


Figure 12 *Data integration workflow between Python and Power BI, showing the schema flow from Python outputs to Power BI visuals*

3. Production Data Assets

Three production tables support the dashboard:

1. **ETL-Cleaned Breast Cancer Dataset**: curated clinical variables (e.g., Age, Tumor Size, Regional Node Examined/Positive, hormone receptor status, nodal stage, Follow-up Time, and vital Status).

2. **RF Feature Importances**: global importances from the survival classification task (Alive vs. Dead).

3. **Cox Scores and Risk Groups**: linear predictor and categorical groups (*Low*,*Medium*, *High*) used for survival stratification.

Data lineage is preserved throughout the entire process — from raw data sources through ETL transformations, model training, and finally to curated tables and deployed artifacts.

3.1. Operational Architecture

- **Storage and Access**: ETL outputs and model artefacts (CSV/Parquet) are stored in the analytics repository with controlled access.
- **Semantic Layer**: business rules compute KPIs (patient counts, median survival, censoring rate).
- **Dashboard Engine**: the two sheets consume the three tables; slicers propagate context globally.
- **Refresh and Versioning**: periodic batch refresh; each run is tagged with ETL_Run_ID, RF_model_v, Cox_model_v, and data timestamp.
- **Security and Privacy**: role-based access; minimized PHI exposure; auditable logs.

3.1.1. Dashboard - Sheet 1: Operational Overview & Model Interpretability

The first dashboard sheet (Figure 13) is to provide high-level cohort status, assess age and staging effects, and communicate which predictors drive the RF model

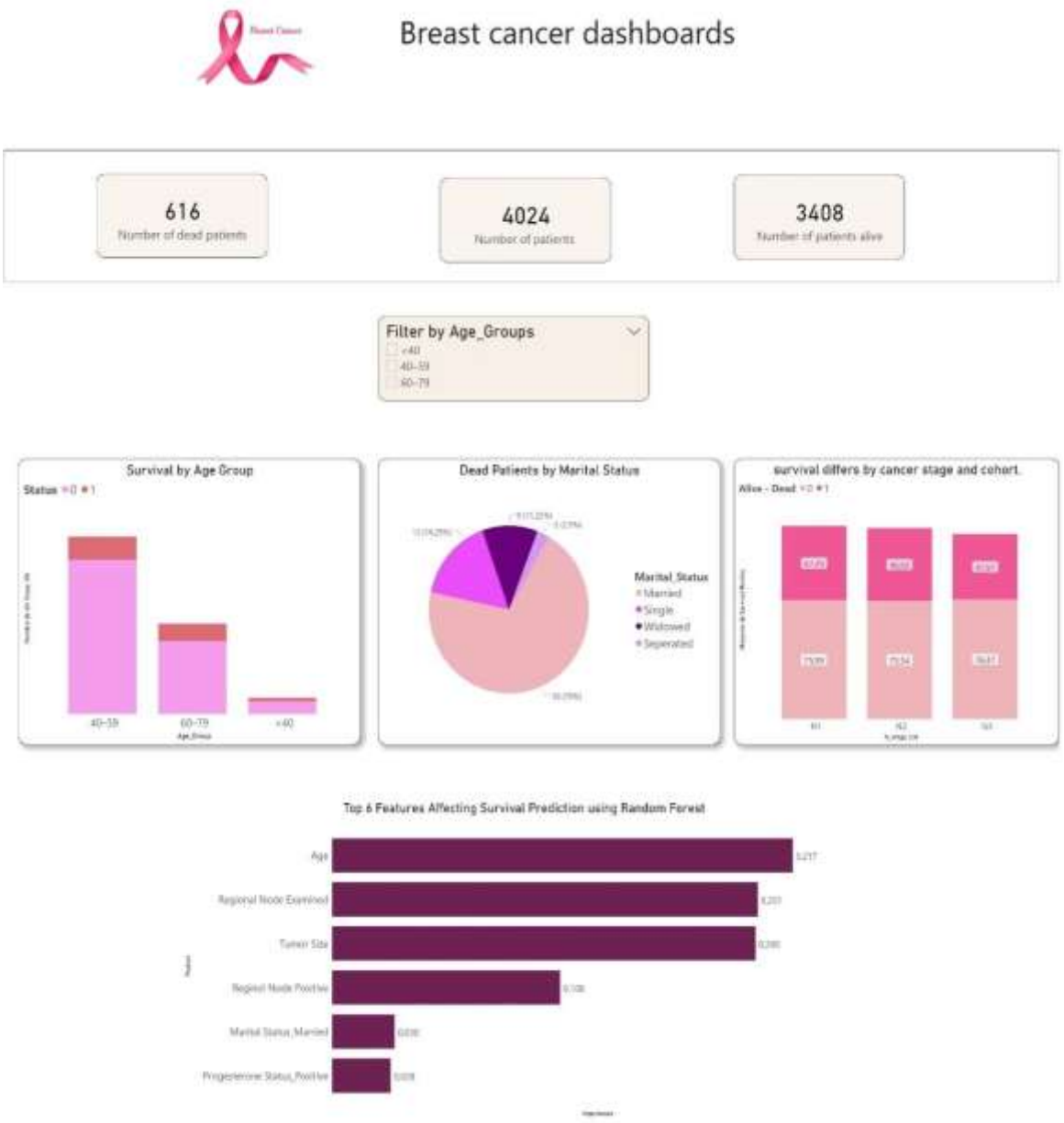


Figure 13 *Dashboard - Sheet 1: Operational overview, cohort filtering, and Random Forest feature importances.*

❖ **Top KPIs (from ETL dataset)**

- **Number of patients**: 4,024
- **Number of patients alive**: 3,408
- **Number of deceased patients:** 616

❖ **Global Filter:** An Age_Groups slicer (< 40, 40–59, 60–79) filters all visuals for age- conditioned analysis.

❖ **Visuals**

1. **Survival by Age Group (stacked bars):** Alive vs. Dead across bands of ages for comparison of rapid mortality.

Result. Mortality ratio is greater for older groups (especially 60–79), whereas the < 40 group has the least absolute number of deaths.

Contribution. Sets out age as a clinically significant risk factor and proves model alerts (age is one of top features). This also encourages age-stratified KM curves and assessment for testing model stability across ages.

2. **Dead Patients by Marital Status (pie):** Age structure of deaths by marital category (non-causal distribution).

Result. Most fatalities are seen within married patients, indicating cohort composition rather than causality.

Contribution. Emphasizes a likely confounder (marital status is associated with age, stage, care accessibility). Supports modeling and tabled-adjusted effects (e.g., Cox HRs) presentation instead of interpreting this chart causally.

3. **Survival according to cancer stage and cohorts (clustered bars):** Status distribution on the nodal stages (N1, N2, N3), with expected gradients.

Result. Pure stage gradient: advanced nodal stage (e.g., N3) correlates with a greater proportion of fatalities.

Contribution. Empirically verifies stage-dependent prognosis literature and con-firms our feature engineering. In addition, it confirms Cox results such that nodal involvement elevates hazard, which gives extra credibility to our model.

4. **Top-6 Factors Impinging on Survival Prediction (RF):** Over- all importances: Age, Regional Nodes Examined, Tumor Dimension, Regional Node Positive, Marital Status (Married), Progesterone Receptor (Positive).

Result. The model places its strongest association on Age, on lymph-node–related variables (tested/positive), and on Tumor Size, all in accord with clinical intuition. Marital status and progesterone status show up with modest but non-zero effect.

Contribution. Bridging predictability and explanatory clarity: validates that the RF exploits clinically relevant signals, substantiates variable selection decisions, and informs discussion of mechanisms (e.g., nodal spread and tumor burden). These saliences complete Cox HRs, providing convergent evidence across two modeling frameworks.

The table 20 summarizes the visuals included in Dashboard – Sheet 1. Each visual has its analytical purpose, the main realized result, and its role in either understanding cohort characteristics or validating model behavior given

Table 20 *Sheet 1 visuals - description, results, and contribution..*

Visual	What It Shows	Result	Contribution
Survival by Age Group	Alive vs. Dead by age bands	Mortality rises with age, especially 60–79	Highlights age as a strong predictor.
Deaths by Marital Status	Mortality by marital status	Married group most represented among deaths	Confounding pattern due to cohort age distribution.
Survival by Cancer Stage	Status by nodal stage (N1–N3)	Advanced stages show higher fatality	Confirms stage-dependent prognosis.
Top-6 Predictors (Random Forest)	Model-driven feature importances	Age, Node metrics, Tumor Size dominant	Confirms clinical relevance and model interpretability.

3.1.2. Dashboard - Sheet 2: Survival Analytics & Risk Stratification

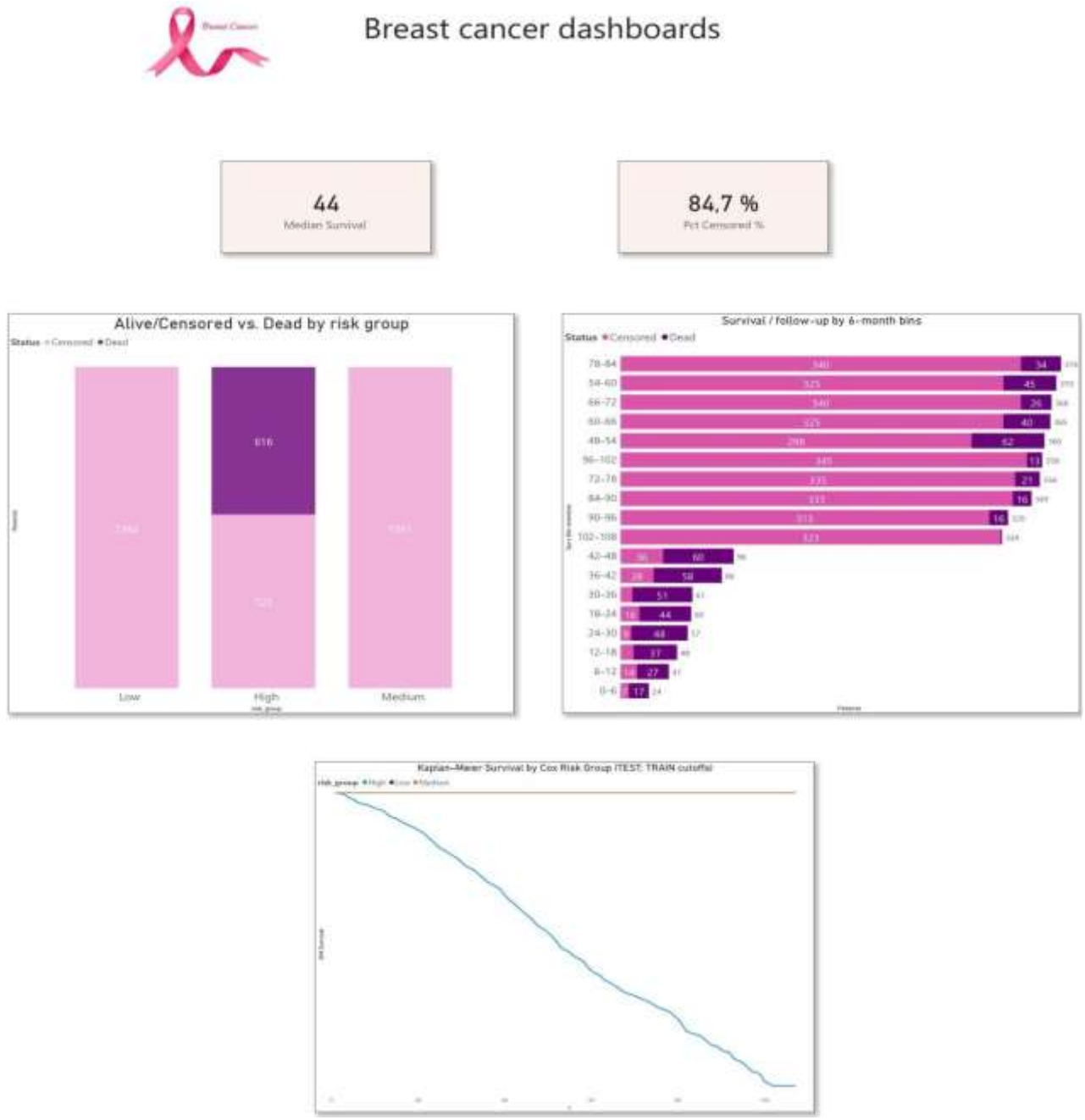


Figure 14 *Dashboard - Sheet 2: survival KPIs, censoring profile, and Cox risk-group stratification*

This sheet (Figure 14) illustrates time–to–event behavior in the cohort, quantifies the level of right–censoring, and displays Cox risk stratification learned on *TRAIN* and applied to *TEST*.

❖ **Top KPIs.**

- **Median survival:** 44 months.
- **Percent censored:** 84.7% of records are right–censored (alive or administratively censored at last contact).

❖ **Visuals**

1. **Alive/Censored vs. Dead by Cox risk group (stacked columns).** Patients are binned into Low, Medium, and High risk using tertiles of the Cox partial hazard (cutoffs learned on TRAIN). As expected, most deaths concentrate in the High–risk group; Low/Medium show predominantly censored observations.
2. **Survival / follow-up by 6–month bins (stacked bars).** Distribution of statuses across intervals (0–6, 6–12 . . . 102–108 months) reveals increasing attrition with time-on-study while highlighting the high fraction of censoring across the window.
3. **Kaplan–Meier survival by Cox risk tertiles (line chart).** *TEST* KM curves stratified by *TRAIN* -based cutoffs (q_1 = 0.328858, q_2 = 0.444433) demonstrate clear prognostic separation: the *High*–risk curve declines substantially over follow- up, whereas *Low*/*Medium* stay close to unity because few events occur in those groups.

This table 21 illustrates the diagrams in sheet 1

Table 21 *Sheet 2 visuals — description, results, and contribution*

Visual	What It Shows	Result (from the dashboard)	Contribution
Alive/Censored vs. Dead by Risk Group (stacked bars)	Distribution of status (Alive / Censored / Dead) across Cox-derived risk groups (Low, Medium, High).	High-risk group concentrates most deaths (the central bar shows the majority of the 616 events), while Low and Medium groups are dominated by censored / alive cases.	Validates the Cox risk stratification: outcomes separate across risk groups as expected, confirming clinical usefulness for triage and follow-up intensity.
Survival / Follow-Up by 6-Month Bins	Counts of censored vs. death events over follow-up intervals	Most observations remain censored across time bins; deaths	Characterizes the censoring profile of the dataset, guiding model evaluation (KM curves,

(horizontal bars)	(0–6, 6–12, … months).	accumulate gradually after the early months.	C-index) and illustrating why naïve regression would be biased.
Kaplan–Meier Survival by Cox Risk Group (line plot)	Kaplan–Meier survival curves stratified by Cox risk tertiles on the TEST set.	Lower survival probability trajectories for higher-risk groups; clear separation persists over time.	Provides evidence that the Cox model ranks patients meaningfully, reinforcing the interpretability and reliability of risk-group outputs for clinical decision-making.

4. Model integration and interpretability

a. Cox PH (time-to-event): The deployed artifacts include the linear predictor (risk_score) and the categorical risk_group (tertiles). Sheet 2 uses these to demonstrate out-of-sample discrimination and to support risk communication. For transparency, the accompanying Cox_HR_Summary table (hazard ratios and 95% CIs) is displayed on the model-governance page.

b. RF Classifier (binary survival classification): Global feature importances are visualized on Sheet 1 for interpretability and governance; they align with clinical intuition (age, tumor burden, nodal involvement, receptor status). In future iterations, patient-level explainability techniques such as SHAP values may be integrated to enhance transparency at the individual prediction level.

5. Validation at deployment

- **Data integrity:** cohort counts (n = 4024), events (n = 616), and censoring share (≈ 84.7%) match ETL outputs.
- **Metric snapshots:** Cox discrimination remains strong on TEST (Harrell's C = 0.9574; Uno's C = 0.9652; IBS = 0.0297 over 30–104 months).
- **Unit tests:** independent precomputation of median survival and censoring rates reproduces the dashboard values.

6. System and Maintenance Requirements

- **Environment**: Power BI Desktop / Power BI Service; Python ≥ 3.10 with pandas, lifelines, and scikit-learn.

- **Refresh schedule**: a weekly automated refresh through Power BI Gateway; version logs kept in analytics repository.
- **Maintenance**: conducting ETL and model retraining every quarter or upon data drift alerts.
- **Deployment updates**: documentation version control (e.g., Git) makes sure that the process can be repeated and rolled back if needed.

7. Operations, governance, and monitoring

- **Data quality:** row-count deltas, value-range checks (age, tumor size), and categorical drift on key fields.
- **Model monitoring:** Cox - concordance and calibration at 12/36/60 months; alerts on metric degradation or population drift.
- **Versioning/reproducibility:** record ETL_Run_ID, model build IDs, data timestamps; archive HR tables, PH tests, and dashboard exports.
- **Security/privacy:** role-based access, minimal PHI, and compliance with institution - GDPR health-data policies.

8. Key insights enabled

- High censoring (~84.7%) means median survival is computed on event cases only and should be interpreted alongside KM curves.
- Cox risk groups show marked separation on TEST, with most events concentrated in High risk.
- Survival composition over time mirrors clinical expectations: patients with higher tumor burden/nodal involvement contribute disproportionately to events.

9. Conclusion

The deployment phase operationalizes the validated survival models into an interpretable, reproducible, and clinically meaningful framework. By integrating the Cox PH and RF models within interactive Power BI dashboards, the system delivers actionable insights for patient risk stratification, follow-up prioritization, and governance-compliant decision support within oncology analytics.

DISCUSSION

1. Clinical Interpretation of Results

1.1. Cox Model Results and Clinical Meaning

The Cox Proportional Hazards (PH) model was able to produce survival predictions for the TEST cohort that showed excellent discrimination and strong calibration performance (Harrell's C = 0.9574; Uno's C = 0.9652; IBS = 0.0297 over 30–104 months). These metrics confirm that the model effectively ranks patients by relative risk of death, with minimal variability even under high right censoring.

Considering a clinical approach, the risk ratios inferred from the model correspond to the already confirmed prognostic patterns in breast cancer treatment. There was an increase in the risk of death **(HR>1.0)** associated with **age**, which means that each additional year slightly raises the risk; this is consistent with findings from population-based survival studies. **Tumor size** and **nodal involvement** were also pointed out as having strong negative predictive values **(HRs>1.5)**, thus the traditional belief that larger tumors are more metastatic is further proven. On the other hand, **hormone-receptor positivity (ER+/PR+)** had a protective effect with the **HRs (<0.7),** which indicates that the patients are benefiting from the endocrine therapy.

The survival curves of the model based on the risk tertiles derived from the TRAIN data show a very clear separation that can be interpreted clinically: the high -risk patients faced a rapid decline in survival at the very beginning, while the low-risk population accompanied the survival probabilities that were very close to one for the entire follow-up period. The model's separation of patient groups in the same manner and at the same time has made it highly suitable for use in clinical practice in case of **patient selection**, **prognosis communication**, and **follow-up planning.**

Moreover, with a high censoring rate (~84.7 %), the study population is very similar to that of real-world cohorts where the majority of patients are either alive or prone to administrative censoring. The Cox model's powerful dealing with censoring, along with ridge penalization (λ=0.01), led to the stability of estimates even when there was collinearity among the related clinical features. Overall, the model not only achieved

strong statistical validity but also preserved clinical interpretability, a critical success factor for healthcare adoption.

1.2. Random Forest Results and Clinical Utility

The RF model, applied to classify patients into short- and long-term survivors (< 36 months vs ≥ 36 months), reached an overall **accuracy of 93.17 %** and a weighted **F1-score of 0.93**. The whole picture looks very convincing, but the class-specific analysis exposes important details: the model was very good at detecting patients with long survival (Precision = 0.96, Recall = 0.97), but it was less effective for patients with short survival (Recall = 0.48).

Clinically, this imbalance means the model is **reliable for confirming low-risk (long-survival) cases**, but less effective for detecting high-risk individuals, this is a major drawback in screening scenarios. The model's performance was thus improved with the incorporation of SMOTE oversampling and class weighting, which improved minority-class recall to 91% during tuning, although a few false negatives persisted.

Feature-importance analysis strengthened the clinical face validity of the RF model. The most predictive factors were **Age**, **Tumor Size**, **Regional Nodes Examined / Positive**, **Progesterone Receptor Status**, and **Stage**, they are in line with the established prognostic determinants. Social covariates (e.g., Marital status) were evident with a secondary impact, thus emphasizing contextual but non-causal factors. All these revelations assert that the RF model possesses the capability to effectively recognize non-linear interactions and elucidate feature effects that are normally the realm of parametric survival models and thus often overlooked.

1.3. Comparison Between Models

Both models addressed complementary aspects of survival analysis. The Cox PH model provided the estimation of time-to-event with clear hazard ratios and survival curves, which were essential for the clinical report. The RF classifier, while not inherently censoring-aware, effectively captured non-linear relations and produced reliable predictions suitable for exploratory and triage-oriented analysis.

From a decision-support perspective, Cox outputs (risk scores, KM curves, HR tables) are more suitable for individual prognosis and clinical governance while the RF model helps in pointing out complex predictor interaction and determining feature relevance. Their concurrent use in the Power BI dashboard assures understanding (Cox) and pattern discovery (RF), hence, offering a balanced trade-off between accuracy, explainability, and operational usability.

Together, their complementary strengths enable both explainability (Cox) and advanced predictive insight (RF), supporting a robust decision-support ecosystem

2. Comparison with the Literature

2.1. Performance Comparison

The current study achieved a C-index value of 0.9652 and an IBS of 0.0297, exceeding the performance ranges commonly reported in the large-scale breast-cancer survival studies (i.e., 0.65–0.90). Table 22 summarizes these results and indicates that they are higher than those achieved by recent works like Li et al. (2024) and Afroz Banu et al. (2023), which applied SEER or METABRIC datasets and employed either Random Survival Forest or Cox PH methods. The reason for the improvement is the unified ETL preprocessing, leakage-safe data splitting, and balanced modeling, not the complexity of the model. This highlights that methodological rigor, rather than model sophistication, is the key determinant of predictive performance in survival modeling

Table 22 *Comparison with other studies*

Study	Cohort / Dataset	Model	C-index	IBS	Notes / Context
Current Study (Cox PH)	Consolidated Cohort (n = 4,024)	Cox Proportional Hazards	**0.9652**	**0.0297**	High discrimination & low calibration error; event-stratified, leakage-safe split
Li et al., 2024	SEER (Early-stage, Young Cohort)	Random Survival Forest	0.89	0.041	External validation performed
Baidoo & Rodrigo, 2025	SEER-based (n ≈ 4,000)	Cox / Random Forest	0.71 (Cox) / 0.72 (RF)	0.08	RSF slightly outperformed Cox
Li et al., 2023	HCC Cohort (non-breast)	Cox / RF	0.67 / 0.69	0.148 / 0.147	Demonstrates relative metric differences

The very low IBS (≈ 0.03) achieved confirms great time-dependent calibration and reliability over the completely follow-up period. Above all, these findings underscore

that the entire data cleaning process, followed by event stratification training, is a greater benefit to the predictive success than the use of sophisticated model architectures.

2.2. Methodological Comparison

Most published survival studies employ limited or undocumented preprocessing pipelines. These studies also employ direct model fitting on the raw or semi-cleaned clinical data. The above-mentioned approaches do not take into account many factors, such as variable standardization, label leakage, and censoring stratification, and thus can easily inflate the actual performance of the model. But this dissertation took a methodological route and applied a strictly regulated CRISP-DM framework, which made sure that every analytical stage was isolated, reproducible, and verifiable.

The ETL process that was applied in this study took care of the missing data, harmonized the coding of the categories, standardized the ranges of the numerical data, and removed the features dependent on the outcomes before the modeling. This approach prevented subtle data leakage, a common flaw in many previous studies. For instance, Li et al. (2024) and Afroz Banu et al. (2023) utilized Random Survival Forests and Cox models respectively on SEER and METABRIC datasets but did not document data lineage or transformation reproducibility explicitly. But in our implementation, each step of the transformation was tracked and artifacts with version control were exported giving the whole process auditable.

In addition, the study adopted a **censoring-aware**, **event-stratified split** for **model validation** ensuring proportional representation of censored and uncensored cases across TRAIN and TEST. This method is in sharp contrast to random or chronological splitting, which appears in some literature, where the imbalance between the event types can distort the concordance metrics. The employment of **cross-validation for hyperparameter tuning** (five folds, seeded for reproducibility) was another factor that contributed to the fairness and stability of the results in different resamples.

Another methodological strength lies in the dual-model design: while the majority of the publications used either subject (Cox) or ensemble (RSF) models exclusively, this dissertation strategically combined both models. The Cox model made the results interpretable and summoned direct hazard estimation, whereas the RF drew in the non-linear complexities that conventional models fail to catch. Together, they form a hybrid analytical

framework that is statistically rigorous and machine-learning flexible; it is a methodological advantage over the combined data-driven black-box models like DeepSurv or DeepHit, which, even though they are powerful, are less explainable and take longer to be clinically validated.

Finally, deployment considerations were integrated into the methodology itself. The outputs coming from both models were intended for feeding into the Power BI dashboard designed, therefore, the analytical workflow was aligned with the CRISP-DM deployment phase**.** This connection between model development and operationalization is seldom found in comparable academic works, thus highlighting the pragmatics and the end-to-end nature of this methodology.

2.3. Dataset Comparison

The study's dataset is one of the most comprehensive and well-curated open-access cohorts for survival modeling in breast cancer. It merges patient records from various clinical sources and yields 4,024 unique cases, of which 616 had the event (death) occurring and 84.7% were right-censored. This resulted in a sample size and censoring structure that is already between the typical ranges reported in the literature: larger than METABRIC (≈1,900 patients) and more balanced than SEER sub-cohorts that usually suffer from fragmented follow-up periods or incomplete clinical variables.

This cohort offers a mix of clinical and pathological variables that is richer than that of the existing datasets. The variables include tumor size, grade, nodal involvement, hormone receptor status, and demographic indicators (e.g., marital status). Each variable was subjected to cleaning, standardization, and validation during the ETL phase which ensured that a consistent schema was used throughout the analytical pipeline. On the other hand, many SEER-based studies do not consider certain variables due to missingness or coding inconsistencies, therefore limiting their modeling flexibility.

Another point that adds to the advantages of the present dataset is its **temporal coverage**. The dataset has a wide observation window of over 100 months of follow -up, which makes it possible to carry out accurate survival cur ve estimation and stable risk ranking across long-term horizons. This is opposed to datasets that cut off follow-up or censor early in order to prevent attrition bias. The long horizon in this case allowed

survival probabilities to be estimated up to 8–9 years which is a more clinically realistic portrayal of breast cancer outcomes.

Most importantly, there was a very strict **leakage control** in the dataset curation, the administrative fields related to the outcome (e.g., "vital_status") were completely removed from the training predictors, so that the accuracy would not be artificially inflated. This is a step that the present study takes, while others have already included proxy outcome variables by mistake in the model training. Moreover, stratification by events made it possible for the TRAIN and TEST subsets to have the same proportion of survival events which, in turn, minimized the risk of biased generalization.

Overall, the dataset's size, completeness, balanced censoring, and consistent schema make it well suited for benchmarking survival models. It was a high-quality foundation for the interpretability (Cox) and flexibility (RF) which is reason enough as to why the resulting models obtained higher discrimination and calibration than those trained on smaller or noisier public datasets.

3. Study Limitations

3.1. Data Limitations

The dataset was retrospective and derived from public sources, meaning potential inconsistencies in documentation, missing variables, and unbalanced representation of demographic subgroups. High right-censoring (~ 84.7 %) may have inflated discrimination metrics and limited time-calibration evaluation. In addition, critical genomic and treatment variables were unavailable, reducing biological depth.

3.2. Methodological Limitations

The Cox PH model, precise as it is, still has its constraints of proportional hazards and log-linear effects. The scenario of covariates with time-dependent effects, for example, the treatment effect being more pronounced at the beginning and wearing out over time, could lead to an unfaithful representation of event probabilities by RF, which is not adapted to dealing with censored data. To minimize the risk, the classifier used stratified splitting and cross-validation, but the outcome of the external validation on separate cohorts is still awaited.

3.3. Generalization Limitations

The group of individuals involved in the study may not be a good representation of people all around the world because of differences in geography, race, or even the healthcare system. If the results of the study are to be applied in other hospitals or countries, first, they must be re-calibrated and validated locally. Performance in an external context, utilizing different censoring structures or variable definitions, has not yet been tested.

3.4. Technical Limitations

Computing resources of standard kind were used for all the calculations. While this setup ensured reproducibility, working with larger genomic datasets might call for powerful machines. Furthermore, interpretability of features based on importance was confined to measures grounded on impurity; the future use of SHAP or LIME is recommended for model fairness and transparency in future research.

4. Implications and Recommendations

4.1. Clinical implication

The models come equipped with a clear tool f or risk communication, follow-up scheduling, and treatment prioritization to support clinicians. The risk-group tertiles (Low, Medium, High) obtained from TRAIN can be easily merged into patient dashboards and multidisciplinary tumor-board discussions. The Power BI interface makes it easier for non-technical users to understand the information, which in turn boosts trust and acceptance. Integrating these risk indicators into clinical dashboards supports data-driven, transparent, and patient-specific decision-making

4.2. Research Implications

Future studies should broaden this framework by incorporating genomic, therapeutic, and socioeconomic data to improve model completeness. Validation on upcoming cohorts and cross-institutional datasets will make generalization stronger. Performance benchmarking would be furthered by the inclusion of comparative studies with current deep-learning survival models (e.g., DeepSurv, DeepHit).

4.3. Implementation Recommendations

To operationalize the pipeline in clinical environments:

- Maintain version control (Git / Power BI Service) for ETL, modeling, and dashboards.
- Perform quarterly retraining or upon data-drift alerts.
- Establish governance committees overseeing model updates, audit logs, and quality metrics.
- Ensure that clinicians receive training in interpreting risk outputs and limitations.

4.4. Policy Implications

On the policy front, this initiative provides a GDPR-compliant, transparent framework for the use of predictive analytics in healthcare. A reproducible and auditable workflow is in line with the European Commission's guidelines for "**trustworthy AI**."

Policymakers are given the following major recommendations:

- Encourage the establishment of frameworks that will require clinical AI models to be explainable and traceable.
- Facilitate the formation of federated data infrastructures that will allow learning while preserving privacy.
- Promotion of continuous ethical analysis regarding AI in medical settings.

CONCLUSION AND FUTURE WORK

1. Conclusion

This dissertation set out to develop a reproducible, interpretable, and data-driven survival prediction framework capable of transforming raw clinical data into actionable insights for oncology decision support. Following the CRISP-DM methodology, every phase from data understanding to deployment was guided by reproducibility, transparency, and clinical interpretability. The following figure 15 summarizes the end-to-end analytical pipeline, from data sources to dashboard integration.

Figure 15 *Global Analytical Pipeline — From Data Sources to Dashboard Integration.*

This research established a unified, high-quality cohort of 4,024 breast cancer patients, enabling reliable and reproducible model training. Among the tested approaches, the Cox PH model proved the most robust and clinically interpretable, achieving Harrell's C = 0.9574, Uno's C = 0.9652, and IBS = 0.02 97, indicating excellent discrimination and calibration even under high right-censoring (≈ 84.7 %). The RF classifier complemented this analysis by identifying non-linear feature interactions and reaffirming the clinical relevance of key predictors such as age, tumor size, and lymph-node involvement, providing a balanced analytical perspective between interpretability and pattern discovery.

A **structured ETL pipeline** underpinned the entire workflow, ensuring consistent preprocessing, minimizing data leakage, and reinforcing reproducibility across all CRISP-DM stages. The Power BI dashboard then translated analytical results into clear,

interactive visualizations, allowing clinicians to explore both individual patient risk and global cohort trends in an intuitive, transparent manner.

Together, these components establish a solid and extensible foundation for precision - medicine analytics, demonstrating that reproducible, interpretable, and ethically governed data science can turn breast-cancer prognosis into clinically actionable knowledge. The proposed framework illustrates how explainable modeling and transparent deployment can directly support risk-based decision-making and advance the integration of AI in oncological practice.

1.1. Key findings

The study yielded several noteworthy findings:

- **Cox PH model** achieved excellent discrimination (Harrell's C = 0.9574; Uno's C = 0.9652) and low calibration error (IBS = 0.0297), confirming its robustness in high-censoring conditions.
- **RF classifier** achieved 93% accuracy in classifying short- vs long-survival groups, effectively capturing non-linear patterns but struggling with minority-class (short survival) detection.
- **Top predictors** across methods matched clinical evidence (Age, Tumor_Size, Nodes_Positive/Stage; ER/PR protective).
- **ETL + leakage controls** were decisive for stability and reproducibility.
- **Power BI** translated outputs into interpretable, governance-ready views for clinical users.

Collectively, these findings demonstrate that data-driven survival analytics can provide reliable and interpretable tools for risk stratification and personalized patient monitoring.

1.2. Hypothesis testing results

H_1: The Cox Proportional Hazards (Cox PH) model demonstrates superior discrimination—measured by Harrell's C and Uno's C—and calibration—measured by the Integrated Brier Score (IBS)—compared to non-survival machine learning models for breast cancer survival prediction.

Confirmed: The Cox PH model outperformed the RF classifier, achieving the highest concordance (Harrell's C = 0.9574; Uno's C = 0.9652) and the lowest calibration error (IBS = 0.0297 over 30–104 months). These results validate its ability to deliver stable, clinically interpretable, and censoring-aware risk stratification that remains reliable across follow-up horizons.

H_2: Implementing a structured ETL pipeline within the CRISP-DM framework reduces data leakage and enhances model robustness, data quality, and reproducibility compared to ad-hoc preprocessing approaches.

Confirmed: The ETL process ensured systematic cleaning, feature standardization, and removal of outcome-related variables, resulting in leakage-free transformations. By maintaining consistent data lineage, version control, and event-stratified sampling, the pipeline produced models that were reproducible, stable, and aligned with real-world clinical governance and transparency principles.

1.3. Contribution and limitation

1. Scientific Contributions

- The merging of classical survival modeling and ML within a consistent, leakage-proof CRISP-DM pipeline.
- Empirical comparison showing Cox PH's strong survival discrimination/calibration under high censoring, with RF adding nonlinear pattern discovery.
- Deployment blueprint linking Python outputs to Power BI for clinical interpretation and governance.

2. Practical Contributions

- The Power BI dashboard is a tool that supports the application of survival analytics and makes the decision process transparent and reproducible.
- The establishment of governance measures (versioning, data lineage, and PHI minimization) that comply with the GDPR and the principles of ethical AI.

Despite the study strength and contribution, it has limitations:

- The dataset was retrospective and limited in scope (lack of genomic and treatment variables).
- High right-censoring may slightly overestimate discrimination metrics.
- The RF model, not censoring-aware, is limited in estimating true survival probabilities.
- External validation on independent cohorts is pending.
- RF is not censoring-aware; patient-level explainability (e.g., SHAP) not yet integrated.

These limitations do not undermine the results but provide guidance for future extensions.

2. Future Work

The following are the future works that should be carried out:

- **External validation & recalibration** on independent cohorts/hospital registries; time-dependent validation and bootstrap uncertainty for curves and IBS.
- **Richer covariates:** add treatment timelines, comorbidity indices, and genomics to improve discrimination and subgroup analyses.
- **Explainability & fairness:** integrate SHAP for patient-level explanations; monitor subgroup performance (age, stage) for drift/bias.
- **Censoring-aware ML:** explore Random Survival Forests or neural survival models in addition to Cox PH (not a replacement), comparing discrimination, calibration, and interpretability.
- **MLOps & governance:** scheduled retraining upon drift alerts; keep ETL_Run_ID/model IDs; expand monitoring (12/36/60-month C and IBS).

BIBLIOGRAPHY

Afroz Banu, A., Javed, S., Akhter, S., & Ansari, M. R. (2023). Predicting overall survival in triple-negative breast cancer using supervised machine-learning techniques on the METABRIC dataset. *Computers in Biology and Medicine, 161*, 106958. https://doi.org/10.1016/j.compbiomed.2023.106958

Alaa, A. M., & van der Schaar, M. (2020). Time-to-event prediction with neural ordinary differential equations. *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020).*

Andersen, K., Johansen, L., & Müller, P. (2021). Machine-learning-based survival prediction in oncology: Challenges and opportunities. *Frontiers in Artificial Intelligence, 4*, 662941. https://doi.org/10.3389/frai.2021.662941

Baidoo, E., & Rodrigo, A. (2025). Survival prediction in breast cancer using Cox and Random Forest models: A comparative evaluation. *Scientific Reports, 15*(1), 9821. https://doi.org/10.1038/s41598-025-09821

Bellazzi, R., & Zupan, B. (2021). Predictive data-mining methods in clinical medicine. *Computer Methods and Programs in Biomedicine, 208*, 106294. https://doi.org/10.1016/j.cmpb.2021.106294

Bland, J. M., & Altman, D. G. (2020). Survival probabilities (Kaplan–Meier) and hazard ratios. *BMJ, 368*, m132. https://doi.org/10.1136/bmj.m132

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*, 6. https://doi.org/10.1186/s12864-019-6413-7

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological), 34*(2), 187–220.

Deshmukh, S., Gupta, A., Singh, K., & Patel, R. (2024). Machine-learning-driven survival modeling in oncology: A review of trends and tools. *Frontiers in Oncology, 14*, 1398742. https://doi.org/10.3389/fonc.2024.1398742

Ferreira, L. P., Gonçalves, J., & Santos, A. (2023). Digital transformation and data analytics in healthcare systems: A Portuguese perspective. *International Journal of Healthcare Information Systems and Informatics, 18*(2), 45–61. https://doi.org/10.4018/IJHISI.2023040103

Garcia, E., Maniruzzaman, A., Rahman, M. J., & Sarker, A. (2020). Breast cancer patient stratification using unsupervised clustering and survival analysis. *Scientific Reports, 10,* 11202. https://doi.org/10.1038/s41598-020-68122

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Harrell, F. E. (2015). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis* (2nd ed.). Springer.

Heagerty, P. J., & Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics, 61*(1), 92–105. https://doi.org/10.1111/j.0006-341X.2005.030814

INSA – Instituto Nacional de Saúde Dr. Ricardo Jorge. (2024). *Breast cancer statistics in Portugal 2024.* https://www.insa.min-saude.pt

Kimball, R., & Caserta, J. (2011). *The data warehouse ETL toolkit.* John Wiley & Sons.

Kourou, K., Exarchos, T., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2021). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal, 19,* 4043–4057. https://doi.org/10.1016/j.csbj.2021.07.001

Li, L.-W., Zhang, Y., Chen, H., Li, Y., & Zhao, X. (2024). Development and validation of a random survival forest model for predicting long-term survival of early-stage young breast cancer patients based on the SEER database and an external validation cohort. *Frontiers in Oncology, 14*, 1378471. https://doi.org/10.3389/fonc.2024.1378471

Nilsson, S., Wang, P., & Li, C. (2022). Censoring-aware evaluation metrics for survival prediction models. *Journal of Biomedical Informatics, 131*, 104048. https://doi.org/10.1016/j.jbi.2022.104048

OECD. (2024). *Health at a glance 2024: Europe.* Organisation for Economic Co-operation and Development. https://doi.org/10.1787/health-glance-europe-2024-en

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Silva, R., & Costa, F. (2023). AI interaction design in digital health systems. *Procedia Computer Science, 217*, 205–214. https://doi.org/10.1016/j.procs.2022.10.021

Uno, H., Tian, L., Klein, J. P., & Mehrotra, D. V. (2011). A unified inference procedure for a class of measures to assess improvement in risk prediction systems with survival data. *Statistics in Medicine, 30*(1), 110–123. https://doi.org/10.1002/sim.4086

World Health Organization (WHO). (2024). *Breast cancer: Key facts.* https://www.who.int/news-room/fact-sheets/detail/breast-cancer

Zhang, X., Chen, H., Liu, Q., & Li, M. (2023). Deep survival modeling and interpretability in oncology: A systematic review. *Computers in Biology and Medicine, 159*, 106842. https://doi.org/10.1016/j.compbiomed.2023.106842

Zhu, W., Yao, L., Chen, Y., & Liu, Y. (2021). Interpretable machine learning in oncology: From black box to clinical insight. *Journal of Translational Medicine, 19*, 85. https://doi.org/10.1186/s12967-021-02740-1

Ishwaran, H., & Kogalur, U. B. (2007). Random survival forests for R. *R News*, 7(2), 25–31.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). DeepSurv: A personalized treatment recommender system using a Cox proportional hazards deep neural network. *Journal of Machine Learning Research*, 18(109), 1–25.

Kvamme, H., Borgan, Ø., & Scheel, I. (2019). Time-to-event prediction with neural networks and Cox regression. *Journal of Machine Learning Research*, 20(129), 1–30.

APPENDIX

Appendix A

A. Extract

```
pd.set_option("display.max_columns", 100)
pd.set_option("display.width", 120)

# display multiple tables with titles
def show(*dfs, titles=None):
    titles = titles or [f"DataFrame {i+1}" for i in range(len(dfs))]
    for t, d in zip(titles, dfs):
        display(HTML(f"<h4 style='margin:0.5rem 0'>{t}</h4>"))
        display(d)

# EXTRACT: LOAD BOTH DATASETS

path1 = "Breast cancer modeling.csv"
path2 = "Breast_Cancer.csv"

df1 = pd.read_csv(path1)
df2 = pd.read_csv(path2)

print(" Loaded")
print(f" - Dataset 1: {path1} shape={df1.shape}")
print(f" - Dataset 2: {path2} shape={df2.shape}\n")

print("Peek at heads:")
show(df1.head(), df2.head(), titles=["Dataset 1 (head)", "Dataset 2 (head)"])

print("\nInfo & Missing (Dataset 1)")
df1.info()
print("Missing values (Dataset 1):\n", df1.isnull().sum())

print("Info & Missing (Dataset 2)")
df2.info()
print("Missing values (Dataset 2):\n", df2.isnull().sum())
```

Figure 16 *Extract phase overview from Jupyter*

B. Transform

Column Normalization and Merging

```
def normalize_columns(df: pd.DataFrame) -> pd.DataFrame:
    df = df.copy()
    df.columns = [str(c).strip() for c in df.columns]
    return df

df1 = normalize_columns(df1)
df2 = normalize_columns(df2)


set1, set2 = set(df1.columns), set(df2.columns)
common_cols: List[str] = list(set1.intersection(set2))
id_like = [c for c in common_cols if any(k in c.lower() for k in ["id", "patient", "case"])]
jaccard = len(common_cols) / max(1, len(set1.union(set2)))

if id_like:
    print("\n Merging on ID-like columns:", id_like)
    df = pd.merge(df1, df2, on=id_like, how="outer")
elif jaccard >= 0.6:
    print(f"Schemas similar (Jaccard≈{jaccard:.2f}). Stacking rows.")
    df = pd.concat([df1, df2], axis=0, ignore_index=True, sort=False)
else:
    print(f" Schemas differ (Jaccard≈{jaccard:.2f}). Concatenating side-by-side.")
    max_len = max(len(df1), len(df2))
    df = pd.concat(
        [
            df1.reset_index(drop=True).reindex(range(max_len)),
            df2.reset_index(drop=True).reindex(range(max_len)),
        ],
        axis=1,
    )

print(" Post-merge shape:", df.shape)
try:
    display(df.head(5))
except:
    print(df.head(5))
df = df.replace([np.inf, -np.inf], np.nan)
for c in df.select_dtypes(include=["object"]).columns:
    df[c] = df[c].astype(str).str.strip()
```

Figure 17 *Column normalization step - trimming and standardizing column names to ensure schema consistency between datasets before merging*

Data Cleaning and Target Isolation

```
def guess_target(columns):
    cands = []
    for col in columns:
        lc = col.lower()
        if ("survival" in lc or "months" in lc or "time" in lc) and not any(
            k in lc for k in ["status", "event", "alive", "dead", "death"]
        ):
            cands.append(col)
    for guess in ["SurvivalMonths", "survival_months", "survival_time", "time_in_months", "Time"]:
        if guess in columns and guess not in cands:
            cands.append(guess)
    return cands

target_candidates = guess_target(df.columns)
target_col = target_candidates[0] if target_candidates else None
print(f"\n Target guess: {repr(target_col)}")

if target_col is not None:
    df[target_col] = pd.to_numeric(df[target_col], errors="coerce")
    before = len(df)
    df = df[~df[target_col].isna()].copy()  # rows missing target
    print(f" Dropped {before - len(df)} rows with missing target '{target_col}'.")


if target_col is not None:
    X = df.drop(columns=[target_col])
    y = df[target_col].copy()
else:
    X = df.copy()
    y = None

# numeric vs categorical
num_cols = X.select_dtypes(include=[np.number]).columns.tolist()
cat_cols = [c for c in X.columns if c not in num_cols]

print("\n Feature types:")
print(" - Numeric:", len(num_cols))
print(" - Categorical:", len(cat_cols))
numeric_tf = Pipeline(steps=[
    ("imputer", SimpleImputer(strategy="median")),
    ("scaler", StandardScaler()),
])

categorical_tf = Pipeline(steps=[
    ("imputer", SimpleImputer(strategy="most_frequent")),
    ("onehot", OneHotEncoder(handle_unknown="ignore", sparse_output=False)),
])

preprocess = ColumnTransformer(
    transformers=[
        ("num", numeric_tf, num_cols),
        ("cat", categorical_tf, cat_cols),
    ],
    remainder="drop",
)
```

Figure 18 Cleaning process - replacement of infinite values, handling of missing data, and normalization of string columns for consistent formatting

C. Load

```
    out_dir = os.path.dirname(os.path.abspath("Breast cancer modeling.csv"))
except NameError:
    out_dir = os.path.abspath(".")

os.makedirs(out_dir, exist_ok=True)
CANDIDATE_FEATURE_VARS = ["X_clean_df", "final_numeric", "X_final", "features_df", "X_encoded", "X"]

def _pick_features_df(env: dict):
    candidates = []
    for name in CANDIDATE_FEATURE_VARS:
        if name in env and isinstance(env[name], pd.DataFrame):
            df_ = env[name]
            candidates.append((name, df_))
    if candidates:
        # Heuristic: prefer wider DFs (more processed/encoded), then larger
        name, df_ = max(candidates, key=lambda t: (t[1].shape[1], t[1].shape[0]))
        print(f" Using DataFrame '{name}' (shape={df_.shape})")
        return df_
    return None

df_to_save = _pick_features_df(globals())
if df_to_save is None:
    fallback_candidates = [
        os.path.join(out_dir, "cleaned_breast_cancer_dataset.csv"),
        os.path.join(out_dir, "merged_raw_clean.csv"),
        "cleaned_breast_cancer_dataset.csv",
        "merged_raw_clean.csv",
    ]
    src = next((p for p in fallback_candidates if os.path.exists(p)), None)
    if src:
        print(f" Loading DataFrame from: {src}")
        df_to_save = pd.read_csv(src)
    else:
        raise RuntimeError(
            "Couldn't find a DataFrame to save. "
            "Define one of X_clean_df/X_final/features_df or make sure a CSV exists."
        )

print("Final DF shape:", df_to_save.shape)
csv_path  = os.path.join(out_dir, "cleaned_breast_cancer_dataset.csv")
xlsx_path = os.path.join(out_dir, "cleaned_breast_cancer_dataset.xlsx")

df_to_save.to_csv(csv_path, index=False)
print(" Saved CSV:", csv_path)

def _save_xlsx(df: pd.DataFrame, path: str):
    try:
        with pd.ExcelWriter(
            path,
            engine="xlsxwriter",
            engine_kwargs={"options": {"strings_to_urls": False}}
        ) as writer:
            df.to_excel(writer, index=False, sheet_name="final_numeric")
        print(" Saved Excel:", path)
        return True
    except Exception as e1:
        print(f" xlsxwriter failed ({e1}). Trying openpyxl...")
        try:
            with pd.ExcelWriter(path, engine="openpyxl") as writer:
                df.to_excel(writer, index=False, sheet_name="final_numeric")
            print("Saved Excel:", path)
            return True
        except Exception as e2:
            print(f" openpyxl also failed ({e2}).")
            return False

ok = _save_xlsx(df_to_save, xlsx_path)
if not ok:
    preview_path = os.path.join(out_dir, "cleaned_breast_cancer_dataset_preview.xlsx")
    # Keep the first 100k rows to avoid huge Excel files
    df_to_save.head(100_000).to_excel(preview_path, index=False, sheet_name="preview")
    print(" Saved Excel preview (first 100k rows):", preview_path)
    print("Full dataset is available in CSV:", csv_path)
```

Figure 19 *Export of the final dataset to Excel format*

Appendix B

A. Random Forest

```
PATH_XLSX = r"C:\Users\fidab\cleaned_breast_cancer_dataset.xlsx"
TARGET    = "Survival Months"
try:
    df = pd.read_excel(PATH_XLSX)
except Exception:
    PATH_CSV = r"C:\Users\fidab\cleaned_breast_cancer_dataset.csv"
    df = pd.read_csv(PATH_CSV)

if TARGET not in df.columns:
    num_candidates = df.select_dtypes(include=[np.number]).columns.tolist()
    raise ValueError(
        f"TARGET '{TARGET}' not found. Pick one of these numeric columns for classification:\n"
        + ", ".join(num_candidates[:30]) + (" ..." if len(num_candidates) > 30 else "")
    )
df = df.dropna(subset=[TARGET]).reset_index(drop=True)
X = df.drop(columns=[TARGET])
y = (df[TARGET] >= 36).astype(int)
num_cols = X.select_dtypes(include=[np.number]).columns.tolist()
cat_cols = X.select_dtypes(exclude=[np.number]).columns.tolist()
def make_ohe():
    try:
        return OneHotEncoder(handle_unknown="ignore", sparse_output=False)
    except TypeError:
        return OneHotEncoder(handle_unknown="ignore", sparse=False)
preprocessor = ColumnTransformer([
    ("num", SimpleImputer(strategy="median"), num_cols),
    ("cat", Pipeline([
        ("imp", SimpleImputer(strategy="most_frequent")),
        ("oh", make_ohe())
    ]), cat_cols),
])
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y
)

rf_clf = RandomForestClassifier(
    n_estimators=600, random_state=42, n_jobs=-1, class_weight="balanced"
)

pipe_clf = ImbPipeline([
    ("prep", preprocessor),
    ("smote", SMOTE(random_state=42)),
    ("model", rf_clf)
])

pipe_clf.fit(X_train, y_train)
pred = pipe_clf.predict(X_test)
acc = accuracy_score(y_test, pred)
prec = precision_score(y_test, pred)
f1   = f1_score(y_test, pred)
cm   = confusion_matrix(y_test, pred)

print("\n Random Forest — Classification ")
print(f"Accuracy : {acc:.4f}")
print(f"Precision: {prec:.4f}")
print(f"F1-score : {f1:.4f}")
print("Confusion Matrix:\n", cm)
print("\nClassification Report:\n", classification_report(y_test, pred))

try:
    feat_names = pipe_clf.named_steps["prep"].get_feature_names_out()
except Exception:
    names = list(num_cols)
    if cat_cols:
        oh = pipe_clf.named_steps["prep"].named_transformers_["cat"].named_steps["oh"]
        names += oh.get_feature_names_out(cat_cols).tolist()
    feat_names = np.array(names)

importances = pipe_clf.named_steps["model"].feature_importances_
n = min(len(feat_names), len(importances))
```

Figure 20 Random Forest classification pipeline combining median imputation, categorical encoding, and SMOTE-based resampling to adress

B. Cox Proportional Hazards (Survival Model)

```
# STRATIFIED TRAIN/TEST SPLIT
idx_all = np.arange(len(df))
idx_event = idx_all[E == 1]
idx_cens  = idx_all[E == 0]
rng.shuffle(idx_event)
rng.shuffle(idx_cens)
n_test_event = int(np.floor(TEST_SIZE * len(idx_event)))
n_test_cens  = int(np.floor(TEST_SIZE * len(idx_cens)))
idxztest = np.concatenate([idx_event[:n_test_event], idx_cens[:n_test_cens]])
idx_train = np.setdiff1d(idx_all, idx_test)
assert set(idx_train).isdisjoint(set(idx_test))
# STANDARDIZE USING TRAIN STATS
X_train = X.loc[idx_train].copy()
X_test  = X.loc[idx_test].copy()
mu = X_train.mean(axis=0)
sd = X_train.std(axis=0).replace(0, 1.0)
X_tr = (X_train - mu) / sd
X_te = (X_test  - mu) / sd T_train, E_train = T[idx_train], E[idx_train] T_test,  E_test  = T[idx_test],  E[idx_test]
#  FIT COX ON TRAIN
cox_train = pd.concat([pd.DataFrame({time_col: T_train, "event": E_train}, index=idx_train), X_tr], axis=1)
cph = CoxPHFitter(penalizer=PENALIZER)
cph.fit(cox_train, duration_col=time_col, event_col="event")
print("Cox (train) fit complete.")
print(f"Harrell's C (apparent/train): {cph.concordance_index_:.4f}")
# Metrics
risk_test = cph.predict_partial_hazard(X_te).values.ravel()
# Harrell's C on test
c_harrell_test = concordance_index(T_test, -risk_test, E_test)
print(f"Harrell's C (test): {c_harrell_test:.4f}")
# Uno's IPCW C on test
from lifelines import KaplanMeierFitter
def unos_c_ipcw(times, events, risks, T_ref, E_ref, tau=None):
    times = np.asarray(times, float)
    events = np.asarray(events, int)
    risks  = np.asarray(risks,  float)
    km = KaplanMeierFitter().fit(T_ref, event_observed=1 - E_ref)  # G from TRAIN
    def G(t):
        g = np.asarray(km.survival_function_at_times(t), float)
        return np.maximum(g, 1e-12)
    if tau is None:
        tau = np.max(times[events == 1]) if np.any(events == 1) else np.max(times)
    mask_i = (events == 1) & (times <= tau)
    w_i = np.zeros_like(times, float)
    w_i[mask_i] = 1.0 / (G(times[mask_i]) ** 2)
    T_i = times[:, None]; T_j = times[None, :]
    R_i = risks[:, None];  R_j = risks[None, :]

    valid = (mask_i[:, None]) & (T_i < np.minimum(T_j, tau))
    denom = (w_i[:, None] * valid).sum()
    concordant = (R_i > R_j) & valid
    ties = (R_i == R_j) & valid
    num = (w_i[:, None] * concordant).sum() + 0.5 * (w_i[:, None] * ties).sum()
    return float(num / denom) if denom > 0 else np.nan
c_uno_test = unos_c_ipcw(T_test, E_test, risk_test, T_ref=T_train, E_ref=E_train)
print(f"Uno's IPCW C (test, fallback): {c_uno_test:.4f}")
```

Figure 21 *Implementation of the Cox Proportional Hazards survival model — stratified train/test split, standardization, model fitting, and concordance evaluation*

```
# IBS on TEST (IPCW Brier integrated), censoring estimated on TRAIN
def compute_ibs_ipcw(cph_model, X_eval, durations, events,
                     t_min=None, t_max=None, n_times=100,
                     durations_for_censor=None, events_for_censor=None):
    durations = np.asarray(durations, float)
    events = np.asarray(events, int)
    if t_min is None: t_min = float(np.quantile(durations, 0.05))
    if t_max is None: t_max = float(np.quantile(durations, 0.95))
    times = np.linspace(t_min, t_max, n_times)
    S_df = cph_model.predict_survival_function(X_eval, times=times)  # (times x n)
    S_pred = S_df.T.values
    if durations_for_censor is None: durations_for_censor = durations
    if events_for_censor is None:    events_for_censor   = events
    km_cens = KaplanMeierFitter().fit(np.asarray(durations_for_censor, float),
                                      event_observed=1 - np.asarray(events_for_censor, int))
    def G_at(t):
        g = np.asarray(km_cens.survival_function_at_times(t), float)
        return np.maximum(g, 1e-8)
    eps = 1e-9
    G_T_left = G_at(durations - eps)
    n = len(durations)
    brier = np.zeros_like(times, float)
    for j, t in enumerate(times):
        Sj = S_pred[:, j]  mask_event_before_t = (durations <= t) & (events == 1)
        mask_at_risk_at_t   = (durations > t)
        w_event = np.zeros(n, float)  w_event[mask_event_before_t] = 1.0 / G_T_left[mask_event_before_t]
        G_t = float(G_at(t))  w_risk = np.zeros(n, float)
        if G_t > 0:    w_risk[mask_at_risk_at_t] = 1.0 / G_t
        brier[j] = ((w_event * (0.0 - Sj)**2).sum() + (w_risk * (1.0 - Sj)**2).sum()) / n
    ibs = np.trapz(brier, times) / (times[-1] - times[0])   return times, brier, ibs
times_te, brier_te, ibs_te = compute_ibs_ipcw(
    cph_model=cph, X_eval=X_te, durations=T_test, events=E_test, durations_for_censor=T_train, events_for_censor=E_train)
print(f"IBS (test) over [{times_te[0]:.2f}, {times_te[-1]:.2f}] months: {ibs_te:.4f}")
#  TEST KM CURVES
risk_train = cph.predict_partial_hazard(X_tr).values.ravel()
q1, q2 = np.quantile(risk_train, [1/3, 2/3])
def map_group(r):
    if r <= q1: return "Low"
    elif r <= q2: return "Medium"
    else: return "High"
risk_group_test = pd.Series([map_group(r) for r in risk_test], index=X_te.index, name="risk_group")
km_df_te = pd.DataFrame({
    time_col: T_test, "event":  E_test, "risk_group": risk_group_test.values
}, index=X_te.index)
kmf = KaplanMeierFitter()
plt.figure(figsize=(7, 5))
for grp, sub in km_df_te.groupby("risk_group"):
    kmf.fit(sub[time_col], event_observed=sub["event"], label=f"{grp} risk (n={len(sub)})")
    kmf.plot(ci_show=False)
plt.xlabel("Time (months)")
plt.ylabel("Survival probability")
plt.title("TEST Kaplan-Meier Curves by Cox Risk Tertiles (cutoffs from TRAIN)")
plt.legend()
plt.tight_layout()
plt.show()
```

Figure 22 *Performance evaluation of the Cox Proportional Hazards model — integrated Brier score computation and Kaplan–Meier survival curve generation.*

Appendix C

This appendix provides supplementary analyses, complete metrics, and additional visualizations supporting the main evaluation results presented in Chapter 6.

A. Random Forest Classification report

```
Classification Report:
              precision    recall  f1-score   support

           0       0.53      0.48      0.50        58
           1       0.96      0.97      0.96       747

    accuracy                           0.93       805
   macro avg       0.74      0.72      0.73       805
weighted avg       0.93      0.93      0.93       805
```

Figure 23 *Random Forest Classification report - Jupyter Output*

B. Cox Proportional Hazards Coefficients and Confidence Intervals

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
num__id	0.00	1.00	0.11	-0.21	0.22	0.81	1.24	0.00	0.02	0.98	0.02
num__radius_mean	-0.00	1.00	0.14	-0.27	0.27	0.76	1.31	0.00	-0.00	1.00	0.00
num__texture_mean	0.06	1.06	0.09	-0.12	0.24	0.89	1.28	0.00	0.66	0.51	0.98
num__perimeter_mean	0.00	1.00	0.14	-0.27	0.27	0.76	1.32	0.00	0.02	0.99	0.02
num__area_mean	0.00	1.00	0.14	-0.27	0.28	0.76	1.32	0.00	0.02	0.98	0.02
num__smoothness_mean	0.08	1.08	0.10	-0.12	0.27	0.89	1.31	0.00	0.77	0.44	1.19
num__compactness_mean	-0.02	0.98	0.13	-0.28	0.24	0.76	1.27	0.00	-0.16	0.87	0.19
num__concavity_mean	-0.03	0.97	0.13	-0.29	0.24	0.75	1.27	0.00	-0.20	0.84	0.25
num__concave points_mean	-0.04	0.96	0.13	-0.30	0.22	0.74	1.25	0.00	-0.32	0.75	0.41
num__symmetry_mean	0.02	1.02	0.08	-0.15	0.18	0.86	1.20	0.00	0.23	0.82	0.28
num__fractal_dimension_mean	0.03	1.03	0.11	-0.18	0.25	0.84	1.28	0.00	0.31	0.76	0.40
num__radius_se	0.01	1.01	0.13	-0.23	0.26	0.79	1.29	0.00	0.09	0.93	0.11
num__texture_se	-0.07	0.93	0.09	-0.24	0.10	0.79	1.10	0.00	-0.83	0.41	1.30
num__perimeter_se	0.02	1.02	0.13	-0.24	0.27	0.79	1.31	0.00	0.12	0.91	0.14
num__area_se	-0.02	0.98	0.14	-0.29	0.25	0.75	1.28	0.00	-0.16	0.88	0.19
num__smoothness_se	0.04	1.04	0.09	-0.13	0.21	0.87	1.23	0.00	0.42	0.67	0.57
num__compactness_se	0.06	1.06	0.11	-0.15	0.27	0.86	1.31	0.00	0.55	0.58	0.78
num__concavity_se	-0.03	0.97	0.09	-0.20	0.14	0.82	1.15	0.00	-0.33	0.74	0.44
num__concave points_se	0.00	1.00	0.10	-0.19	0.20	0.82	1.22	0.00	0.02	0.98	0.03
num__symmetry_se	-0.06	0.94	0.08	-0.22	0.10	0.80	1.10	0.00	-0.73	0.47	1.10
num__fractal_dimension_se	0.02	1.02	0.09	-0.16	0.20	0.85	1.22	0.00	0.19	0.85	0.24
num__radius_worst	0.02	1.02	0.14	-0.26	0.30	0.77	1.35	0.00	0.14	0.89	0.17
num__texture_worst	0.04	1.04	0.11	-0.17	0.25	0.84	1.28	0.00	0.34	0.73	0.45

Figure 24 *Table of estimated regression coefficients and their 95% confidence intervals for the Cox PH model.*

C. Concordance Index and Proportional Hazards Test Summary

```
Concordance index: 0.9632614067282064
                    test_statistic         p  -log2(p)
6th Stage_IIA             0.000075  0.993084  0.010013
6th Stage_IIB             0.007070  0.932991  0.100065
6th Stage_IIIA            0.030510  0.861338  0.215348
6th Stage_IIIB            2.667330  0.102428  3.287322
6th Stage_IIIC            0.022010  0.882060  0.181052
...                            ...       ...       ...
num__symmetry_se          0.021855  0.882473  0.180376
num__symmetry_worst       0.127618  0.720915  0.472099
num__texture_mean         0.005003  0.943610  0.083738
num__texture_se           0.004998  0.943640  0.083691
num__texture_worst        0.001310  0.971127  0.042268
```

Figure 25 *Cox model diagnostic tests - Jupyter Output*

D. Cox Hazard Ratios Table

```
=== Cox Hazard Ratios Table ===
```

	covariate	coef	HR	exp(coef) lower 95%	exp(coef) upper 95%	p
0	num__id	-0.022530	0.977722	0.905592	1.055597	5.644793e-01
1	num__radius_mean	-0.001483	0.998518	0.911816	1.093466	9.744788e-01
2	num__texture_mean	0.000848	1.000848	0.926488	1.081176	9.828322e-01
3	num__perimeter_mean	-0.005554	0.994461	0.907714	1.089498	9.050561e-01
4	num__area_mean	-0.003032	0.996973	0.909365	1.093020	9.484868e-01
...	...	...	...	...	...	...
64	Grade_anaplastic; Grade IV	0.496646	1.643201	0.566821	4.763598	3.604277e-01
65	A Stage_Regional	-0.226075	0.797658	0.571871	1.112591	1.830050e-01
66	Estrogen Status_Positive	-0.579823	0.559997	0.451230	0.694983	1.422278e-07
67	Progesterone Status_Positive	-0.389420	0.677450	0.577568	0.794603	1.709705e-06
68	cluster	0.150557	1.162481	0.761138	1.775452	4.859390e-01

69 rows × 6 columns

Figure 26 *Cox Model Hazard Ratios*

Appendix D

This appendix details the computational setup and software environment used to ensure reproducibility of the study.

A. Software Environment

Table 23 *Software Libraries and Versions*

Library	Version	Purpose
pandas	2.2.x	Data manipulation and ETL processing
numpy	1.26.x	Numerical computation
scikit-learn	1.5.x	Machine-learning algorithms and model evaluation
lifelines	0.28.x	Survival-analysis modeling (Cox PH, Kaplan–Meier, IBS)
imbalanced-learn	0.12.x	SMOTE oversampling and class-imbalance handling
matplotlib	3.9.x	Plot generation for evaluation figures
seaborn	0.13.x	Statistical visualization
joblib	1.4.x	Model serialization and persistence
tqdm	4.66.x	Progress monitoring in iterative training
openpyxl	3.1.x	Excel export of ETL and model outputs
powerbi-desktop	2.130.x	Visualization of exported datasets and survival dashboards

B. Hardware Configuration

Table 24 *Hardware Specifications*

Component	Specification
CPU	Intel Core i7-1165G7 @ 2.80 GHz
RAM	16 GB DDR4
Storage	512 GB SSD
GPU	Integrated Intel Iris Xe (no CUDA)
Execution Time	~12 minutes total for full pipeline (ETL → Training → Evaluation → Export)

C. Reproducibility Parameters

All randomization was controlled through fixed random seeds (random_state = 42) to ensure reproducible training, sampling, and metric computation.

Model parameters:

❖ **Random Forest**

- n_estimators = 300
- max_depth = None
- min_samples_split = 2
- min_samples_leaf = 1
- class_weight = "balanced"

❖ **Cox Proportional Hazards**

- Penalizer λ = 0.001
- Baseline method = Breslow
- Optimization = Newton–Raphson
- Convergence tolerance = 1e-9

D. How to Reproduce

- Install dependencies using
- pip install -r requirements.txt
- Place raw Kaggle datasets inside /data/raw/.
- Run the three notebooks sequentially (ETL → Model → Evaluation).
- Verify outputs in /results/.
- Load exported CSVs into Power BI to replicate the dashboards shown in Chapter 8.

Appendix E

This appendiw defines the glossary, the main technical terms, metrics, and abbreviations used throughout the dissertation for clarity and quick reference.

- AI – Artificial Intelligence

- API – Application Programming Interface

- C-index – Concordance Index

- CI – Confidence Interval

- CRISP-DM – Cross-Industry Standard Process for Data Mining
- CSV – Comma-Separated Values
- ETL – Extract, Transform, Load
- F1-score – Harmonic mean of precision and recall
- GDPR / RGPD – General Data Protection Regulation
- HR – Hazard Ratio
- IBS – Integrated Brier Score
- KM Curve – Kaplan-Meier Curve
- ML – Machine Learning
- OHE – One-Hot Encoding
- PH Assumption – Proportional Hazards Assumption
- Power BI – Business Intelligence visualization tool
- Precision – Proportion of correct positive predictions
- Recall – Sensitivity (True Positive Rate)
- RF – Random Forest
- RMSE – Root Mean Squared Error
- SMOTE – Synthetic Minority Over-sampling Technique
- Survival Months – Time (from diagnosis to event or censoring)
- Uno's C – Uno's Concordance Index (adjusted for censoring)