INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Predictive Model for Heart Failure Decompensation

Francisca Dias Passos

Master in Business Analytics

Supervisors:
Doctor Raul Manuel Silva Laureano, Associate Professor, ISCTE Business School

Medical Doctor David Cabrita Roque, Cardiology Department, Hospital Prof. Doutor Fernando Fonseca, Unidade Local de Saúde Amadora/Sintra

October 2024

Department of Quantitative Methods for Management and Economics

Predictive Model for Heart Failure Decompensation

Francisca Dias Passos

Master in Business Analytics

Supervisors:
Doctor Raul Manuel Silva Laureano, Associate Professor, ISCTE Business School

Medical Doctor David Cabrita Roque, Cardiology Department, Hospital Prof. Doutor Fernando Fonseca, Unidade Local de Saúde Amadora/Sintra

October 2024

Acknowledgements

This investigation would not have been possible without the unwavering support of those who stood by me every step of the way, and to them, I extend my deepest gratitude.

To my parents, who always made every effort to enable my academic pursuits, fostering my commitment to learning and always having the right words to say when something seemed not to be going well. And to my little brother, who brings me immense joy as I watch him grow into a kind and intelligent young person.

A special thanks to my sister, who, for as long as I can remember, has always helped me with everything, especially encouraging me to push harder and strive to become the best version of myself.

To all my friends, who made this journey so much more enjoyable.

I extend heartfelt appreciation to the cardiology department for welcoming me each day and creating a nurturing environment.

Special thanks to Dr. David Roque, a proactive thinker who inspired this study and always made time to guide me.

I am especially grateful to Professor Raul Laureano, who undertook this investigation with me and whose expertise was invaluable for my academic growth throughout my master's studies, particularly in this final step.

To everyone who stood by my side, I am profoundly grateful.

Abstract

Heart failure (HF) ranks among the most prevalent and growing chronic health conditions worldwide. Characterized by a progressive course, HF often presents phases of symptomatic stability interspersed with episodes of worsening. These episodes of decompensation frequently necessitate unplanned healthcare visits and/or hospital admissions, increasing mortality risk and significantly reducing quality of life. In this way, HF imposes considerable healthcare-related costs, placing a heavy burden on patients, healthcare systems, professionals, and society at large. Accurately predicting decompensation episodes could enable timely interventions, potentially reducing hospital visits, lowering healthcare expenses, and improving quality of life.

This study aims to address this gap by developing an interpretable machine learning model to predict HF decompensation within a 30-day period. This study is based on data from 584 HF patients, followed over a period of 3 years and 4 months, resulting in a dataset of 2,008 consultations from which over 600 variables were generated and ultimately refined to 25 key predictors. The final model, combining the strengths of XGBoost and a Random Tree model, achieved a recall of 81.40% and an AUC of 0.96. To improve interpretability, a C5.0 algorithm was used to provide a global explanation, along with sensitivity analysis and logistic regression.

This study advances knowledge in HF decompensation prediction, showcasing the integration of business analytics in the health domain. These findings form a foundation for future studies and practical applications aimed at improving patient care and optimizing healthcare resource allocation.

Keywords: Heart Failure, Decompensation, Predictive Model, Machine Learning

JEL Classification System: C53, I12

Resumo

A insuficiência cardíaca (IC) é uma das doenças crónicas mais prevalentes e em crescimento globalmente. Caracteriza-se por um percurso progressivo, com fases de estabilidade intercaladas por episódios de agravamento. Estes episódios de descompensação frequentemente exigem hospitalizações e/ou visitas não programadas, aumentando o risco de mortalidade e diminuindo a qualidade de vida dos doentes. Assim, a IC representa custos significativos para a saúde, sobrecarregando doentes, sistemas de saúde, profissionais e a sociedade no geral. A antecipação destes episódios possibilita uma atuação mais rápida, reduzindo hospitalizações, custos em saúde e melhorando a qualidade de vida dos doentes.

Este estudo procura colmatar esta lacuna desenvolvendo um modelo de *machine learning* interpretável para prever descompensações de IC num período de 30 dias. Dados de 584 doentes, recolhidos ao longo de 3 anos e 4 meses, resultaram num total de 2.008 consultas e na criação de mais de 600 variáveis, das quais 25 preditores foram selecionados. O modelo final, que combina XGBoost e Random Tree, obteve uma sensibilidade de 81,40% e uma AUC de 0,96. Para aumentar a interpretabilidade, foi utilizado o algoritmo C5.0 para explicação global, complementado por uma análise de sensibilidade e regressão logística.

Este estudo avança o conhecimento na previsão de descompensações em IC, demonstrando a aplicação de *business analytics* no setor da saúde. Estes resultados estabelecem uma base para estudos futuros e aplicações práticas, visando a melhoria dos cuidados aos doentes e a otimização de recursos nos sistemas de saúde.

Palavras-Chave: Insuficiência cardíaca, Descompensação, Modelo Preditivo, Machine Learning

Classificação JEL: C53, I12

Index

Index of Table

Index of Figure

List of Acronyms and Abbreviations

AHEAD - Atrial fibrillation, Hemoglobin, Elderly, Abnormal renal parameters, Diabetes mellitus

AUC - Area under the ROC curve

BIOSTAT-CHF - BIOlogy Study to TAilored Treatment in Chronic Heart Failure

CEP - Composite Endpoint

CHAID - Chi-squared Automatic Interaction Detection

Cr - Creatinine

CRISP-DM - Cross-Industry Standard Process for Data Mining

CRT - Cardiac resynchronization therapy

DH - Day Hospital

DM - Diabetes mellitus

DT - Decision Tree

ECG - Echocardiogram

EHMRG - Emergency Heart Failure Mortality Risk Grade

EN - Elastic net

EPICA - Epidemiologia de Insuficiência Cardíaca e Aprendizagem

ES - Emergency Service

GB - Gradient Boosting

GDRP - General Data Protection Regulation

GFR - Glomerular filtration rate

Hb - Hemoglobin

HF - Heart Failure

HOSPITAL - Hemoglobin, discharge from an Oncology service, Sodium level, Procedure during the index admission, Index Type of admission (urgent), number of Admissions during the last 12 months, and Length of stay

Htc - Hematocrit

HTN – Hypertension

ICD - Implantable Cardioverter-Defibrillator

ICD-10 - International Classification of Diseases, Tenth Revision

IT - Information Technology

Lab - Laboratory

LaCE - Length of stay, Age, Comorbidity, and Emergency visit

LACE - Length of stay, Acuity of admission, Comorbidity, and Emergency visit

LIME - Local Interpretable Model-agnostic Explanations

LOS - Length of stay

LR - Logistic Regression

MAGGIC - Meta-Analysis Global Group in Chronic Heart Failure

MARKER-HF - Machine learning Assessment of RisK and EaRly mortality in Heart Failure

ML - Machine Learning

MLP - Multilayer Perceptron

Na - Sodium

NN - Neural Networks

NT-proBNP - N-terminal pro B-type natriuretic peptide

NYHA - New York Heart Association Classification

PND - Paroxysmal nocturnal dyspnea

PORTHOS - PORTuguese Heart failure Observational Study

PRISMA - Preferred Reporting Items for Systematic Reviews and Meta-Analyses

RBC - Red blood cells

RF - Random Forest

ROC - Receiver Operating Characteristic

RT - Random Trees

SHAP - SHapley Additive Explanations

SHFM - Seattle Heart Failure Model

SLR - Systematic Literature Review

SOB - Shortness of Breath

SVM - Support Vector Machine

TTE - Electrocardiogram

ULSASI - Unidade Local de Saúde Amadora/Sintra

VIF - Variance Inflation Factor

WBC - White blood cells

WoS - Web of Science

XGBoost - Extreme Gradient Boosting

1. Introduction

This introduction provides an overview of the research topic, explaining the concepts of Heart Failure (HF) and its decompensation, while emphasizing the importance of predicting such episodes. Furthermore, it delineates the problem statement and research question, outlines the objectives, and describes the methodological approach adopted. Finally, it concludes by presenting the structure and organization of the dissertation.

1.1. Theme and its Importance

HF is a clinical syndrome characterized by symptoms like breathlessness, fatigue and ankle swelling, accompanied by signs such as elevated jugular venous pressure and peripheral edema. This syndrome is the result of structural and/or functional abnormalities in the heart, that causes elevated filling pressures and/or inadequate cardiac output (McDonagh et al., 2021). Often, HF patients progress to a stabilized chronic phase with medical and device therapy. However, this fragile stability is often disrupted by decompensation episodes (Figure 1.1) (Fonseca et al., 2017; Greene et al., 2023).

The definition of decompensation in HF, also referred to as acute HF, acute decompensated HF, and worsening HF, evolved over time and is still not universally standardized (Greene et al., 2023). Traditionally, it has been defined by worsening HF signs and symptoms requiring urgent or emergent care for intravenous or invasive therapies (Butler et al., 2014). However, with the improvement of HF therapies, contemporary perspectives recognize decompensation as a broader concept occurring both before and beyond hospitalization episodes, since the treatment can be offered in the outpatient setting, eliminating the need for hospitalization (Bozkurt, 2023).

Decompensation episodes can culminate in several outcomes, two of them critical in the disease trajectory: readmission and/or mortality. Mortality is defined as death, regardless of the cause, while readmission is characterized by the unplanned hospital visit due to decompensated HF (Tong et al., 2023). Considering the complexities of patient outcomes, researchers often study a composite endpoint (CEP) that consolidates multiple individual events or endpoints into a unified assessment (Baracaldo-Santamaría et al., 2023). In this study, CEP is defined as a composite of mortality or readmission.

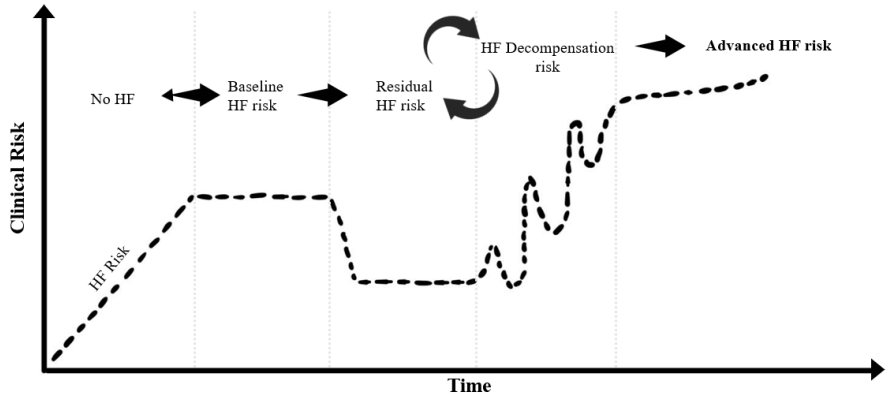


Figure 1.1: Progression of HF Clinical Risk Over Time

Source: Adapted Greene et al. (2023, p.419)

HF is among the most prevalent chronic health conditions globally. This syndrome affects up to 64 million individuals worldwide, and its prevalence is likely to increase as populations age and diagnostic technologies improve (Norhammar et al., 2023). In Portugal, the "Epidemiology of Heart Failure and Learning"[1] (EPICA) study published in 1998, established a baseline HF prevalence of 4.36% among individuals older than 25 on the Portuguese mainland (Ceia et al., 2002). However, the recent presented results of the "Portuguese Heart Failure Prevalence Observational Study" (PORTHOS), found a higher prevalence of 16% among individuals over 50 years old, placing Portugal as one of the countries with the highest reported HF prevalence (Baptista, 2024).

Because of its high prevalence and natural history characterized by frequent health care visits, HF imposes significant healthcare-related costs, placing a substantial effort on patients, national healthcare systems, healthcare professionals, and society as a whole (Baptista et al., 2023). In high-income nations, HF accounts for 1 to 2% of overall healthcare spending, with annual care expenses per patient ranging from 5,000 to 15,000 EUR in Europe and 17,000 to 30,000 USD in the United States (Hessel, 2021). Gouveia et al. (2019, 2020) conducted two critical studies on HF in Portugal, projecting its societal, health-related, and economic implications, between 2014 and 2036. One forecasts a significant rise in HF-related deaths by 2036, reaching 8,112 fatalities, along with a 27.9% increase in disability-adjusted life years lost, totaling 27,059 years (Gouveia et al., 2019). The second study projects that total HF-associated costs in Portugal will reach 503 million EUR in the year 2036, marking a 24% increase from expenditure levels of 2014 (Gouveia

[1] Translation of "*Epidemiologia de Insuficiência Cardíaca e Aprendizagem*"

et al., 2020). Both studies emphasize the urgent need for proactive strategies to tackle the growing challenges posed by HF in Portugal.

Furthermore, the increasing number of initiatives in the healthcare domain, such as the "Transforming Health and Care Systems" aimed at promoting research and innovation in healthcare, underscores the growing significance of this subject (https://www.thcspartnership.eu/).

1.2. Research Problem

There has been an increasing number of studies on machine learning (ML) techniques applied to the population with HF. A preliminary search conducted on Web of Science (WoS) and PubMed using the query ("Heart Failure" AND "Machine Learning") yielded a total of 396 and 248 articles, of which 70% and 90.7%, respectively, were published within the last four years. These ML models designed for HF encompass a diverse array of targets/outcomes. For instance, in Błaziak et al. (2022), several outcomes were considered, including mortality, rehospitalization, response to treatment, and medication adherence, without specific information on predictive models for decompensation.

The potential benefits of integrating ML in the healthcare industry are widely recognized (e.g., Johnson et al., 2018; Obermeyer & Emanuel, 2016; Shameer et al., 2018). However, significant barriers are delaying its broader adoption. One major obstacle arises from the opaque nature of many ML techniques, especially concerning in healthcare decision-making due to the potential costs of inaccurately predicted outcomes (Ahmad et al., 2018). Usually, there is a trade-off between the performance and interpretability of ML models. While simpler models are more interpretable, they often perform worse than complex models like deep learning models (ElShawi et al., 2021). This has led to criticism of using complex models in the medical field despite their higher accuracy. Recognizing these concerns, the European Union has taken action through the General Data Protection Regulation (GDPR), which mandates a "right to explanation" (GDPR Article 22). This regulation aims to provide individuals with meaningful information about the logic and consequences of automated decisions, emphasizing transparency and accountability (Goodman & Flaxman, 2017; Selbst & Powles, 2017). However, ongoing debates persist regarding the practical implementation of this initiative, due to the inherent vagueness of the document and the divergent interpretations of interpretability (ElShawi et al., 2021; Vellido, 2020).

The concept of interpretability in ML has been approached through various terms like transparency, fidelity, and trust. However, existing definitions often overlook end-user needs, resulting in models and explanations that may not adequately address user requirements (Lipton, 2016; Vellido, 2020). Interpretability should ensure the transparency of ML systems, making algorithmic processes understandable and graspable by end users. While interpretable ML models like decision trees (DT) and logistic regressions (LR) offer explanations, others, such as neural networks (NN) or support vector machines (SVM), require post-hoc methods to extract explanations, necessitating careful consideration of medico-legal and ethical requirements in healthcare. One way to achieve this interpretability is by using the simpler structures of the first models to replicate the behavior of less interpretable ones. By attaining comparable metrics, it is possible to offer global explanations that balance accuracy and interpretability, aiding users in assessing and potentially accepting or rejecting predictions before taking action, which is especially crucial in clinical settings.

However, ML models face additional significant challenges, often struggling with imbalanced data (Ahsan & Siddique, 2022), and despite the development of numerous predictive models (e.g., Guo et al., 2021; Kerexeta et al., 2023; Sohrabi et al., 2019), few have undergone external testing. This raises concerns about their generalizability and emphasizes the need for broader validation (Błaziak et al., 2022). Unlike other domains, the medical sector faces strict privacy and legal regulations, leading to smaller datasets that can result in biased ML models and increase the risk of overfitting by focusing on noise rather than general patterns (Althnian et al., 2021; Ying, 2019) .

To address all these concerns and gaps in the literature, this study aims to answer the question: “How can a ML model for HF decompensation prediction enhance the management of HF patients? **“**

1.3. Objectives and Contributions

Considering the research question, the main goal of this study is to develop an effective and interpretable ML model for predicting the occurrence of HF decompensation. To achieve this, four specific objectives are defined: O1) Characterize and define the clinical event of decompensation in the context of HF; O2) Develop a predictive model for HF decompensation; O3) Identify the key factors influencing decompensation events; O4) Determine patient profiles prone to experiencing decompensation events and those less susceptible.

By achieving these objectives and thereby addressing the research question, this study contributes to the development of interpretable predictive models for HF decompensation, which are almost non-existent at the current moment.

Unplanned admissions due to HF decompensation pose a substantial burden on healthcare systems globally (Hessel, 2021; Norhammar et al., 2023; Savarese et al., 2023). Creating an effective interpretable predictive model for decompensation episodes offers multifaceted advantages, benefiting patients, medical professionals, and all society (Moreno-Sánchez, 2023; Sharma et al., 2022a). Patients, as the primary beneficiaries, stand to gain by breaking free from the vicious cycle of declining health, frequent hospital visits, and readmissions, which not only deteriorate their quality of life but also reduce survival (Gouveia et al., 2019). Healthcare providers, on the other hand, can leverage the model for more precise and targeted care interventions, leading to a reduction in complications and readmissions, which ultimately facilitate and reduce their work burden (Rahman et al., 2023). At a societal level, there is potential for substantial cost savings (Savarese et al., 2023). Moreover, this research contributes to advancing medical understanding and promotes the integration of technology into healthcare practices, promising improved patient outcomes across a spectrum of medical scenarios.

1.4. Methodological approach

Given the research problem and its intrinsic objectives, this case study employs a mixed methods design (Johnson et al., 2007), with objective O1 being addressed using a qualitative method, while objectives O2 to O4 are approached quantitatively.

In terms of methodology, the study adopts the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology (Chapman et al., 1999) as its framework. This approach, recognized for its flexibility and iterative nature, enables interactive engagement with physicians and clinical experts, thereby enhancing the attainment of the specified objectives.

This study employs a retrospective methodology, analyzing secondary data from medical records of 584 patients followed by the HF Team in the Cardiology Department at Unidade Local de Saúde de Amadora/Sintra (ULSASI) between 2020 and April of 2024. This data encompasses various types of medical records, including written consultations and laboratory results. However, these records comprise free text lacking a cohesive structure and are stored within the hospital system, requiring assistance from the

Information Technology (IT) team to retrieve them, followed by extensive data preparation.

The present study adhered to all international ethical standards, including the "World Medical Association Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Subjects", was approved by the Ethics Committee of ULSASI, follows the guidelines of ISCTE, and the regulations of the GDPR, demonstrating a strong commitment to the protection of personal data. The clinical investigation unit of ULSASI permitted the study to proceed without signed informed consent, given the academic value of the retrospective study. Patient data were anonymized through pseudonymization, ensuring data processing without names or hospital record numbers, so that no information could be traced back to any specific individual.

1.5. Dissertation structure

To effectively predict HF decompensation, this thesis is structured into five chapters, each representing a different stage of the research process.

The first chapter provides an overview of the study, presenting the theme and its significance, along with the research problem, objectives, primary contributions, and the methodological approach employed. The second chapter reflects the current state-of-art regarding the prediction of HF decompensation through ML techniques. It encompasses the protocol of the systematic literature review (SLR) and subsequent systematic article analyses across four main domains: the context and scope of the studies, methodology, results, and their impacts. The subsequent chapter provides a detailed narrative of the adopted methodology, covering the entire research process from data collection and processing to data analyses, including the metrics for model evaluation. The results, along with their discussion, are presented in the fourth chapter. The fifth and final chapter presents the conclusions, emphasizing the contributions of this study, discussing its limitations, and providing recommendations for future research.

2. Systematic Literature Review

This chapter seeks to systematize the scientific understanding of HF decompensation with a specific focus on its prediction through the application of ML techniques.

A widely accepted definition of ML is that it encompasses models capable of autonomously learning from data, eliminating the need for explicit input from modelers. Including a diverse set of algorithms that empower computers to discern patterns, make predictions, and decisions based on underlying data structures (Benedetto et al., 2022; Christodoulou et al., 2019; Shameer et al., 2018). Expanding upon the previous definition, Beam and Kohane (2018) advocate for viewing ML as a spectrum. They propose that whether a predictive algorithm qualifies as ML depends on the degree to which its structure or parameters are preordained by humans. ML approaches are broadly categorized as either unsupervised or supervised learning, each serving distinct purposes. In this dichotomy, unsupervised learning focuses on uncovering inherent relationships among variables, while supervised learning involves the classification of observations into specific categories or outcomes based on a dataset containing predictor variables and labeled outcomes (Johnson et al., 2018).

In critical fields like healthcare, methodological rigor is crucial for informed decision-making, especially for conditions like HF decompensation, where accurate prediction and management are key. Within this context, a SLR plays a crucial role in gathering, evaluating, and synthesizing existing knowledge.

A SLR, also known as a systematic review (Kitchenham, 2004), is a methodological approach that formulates precise research inquiries (Ahsan & Siddique, 2022) and uses systematic, explicit, and reproducible techniques. It provides a framework for identifying, evaluating, and synthesizing existing research (Kitchenham, 2004; Okoli & Schabram, 2015), offering a comprehensive overview of the state-of-the-art, highlighting research gaps, and suggesting directions for future studies (Paul et al., 2023; Paul & Criado, 2020).The research and evaluation methodology for the SLR is based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework (Moher et al., 2009), which enhances transparency and completeness in the process of writing systematic reviews. PRISMA is one of the most widely adopted protocols (Paul et al., 2023), with updates in 2020 (Page et al., 2021), reinforcing its importance in modern research practices.

2.1. Protocol

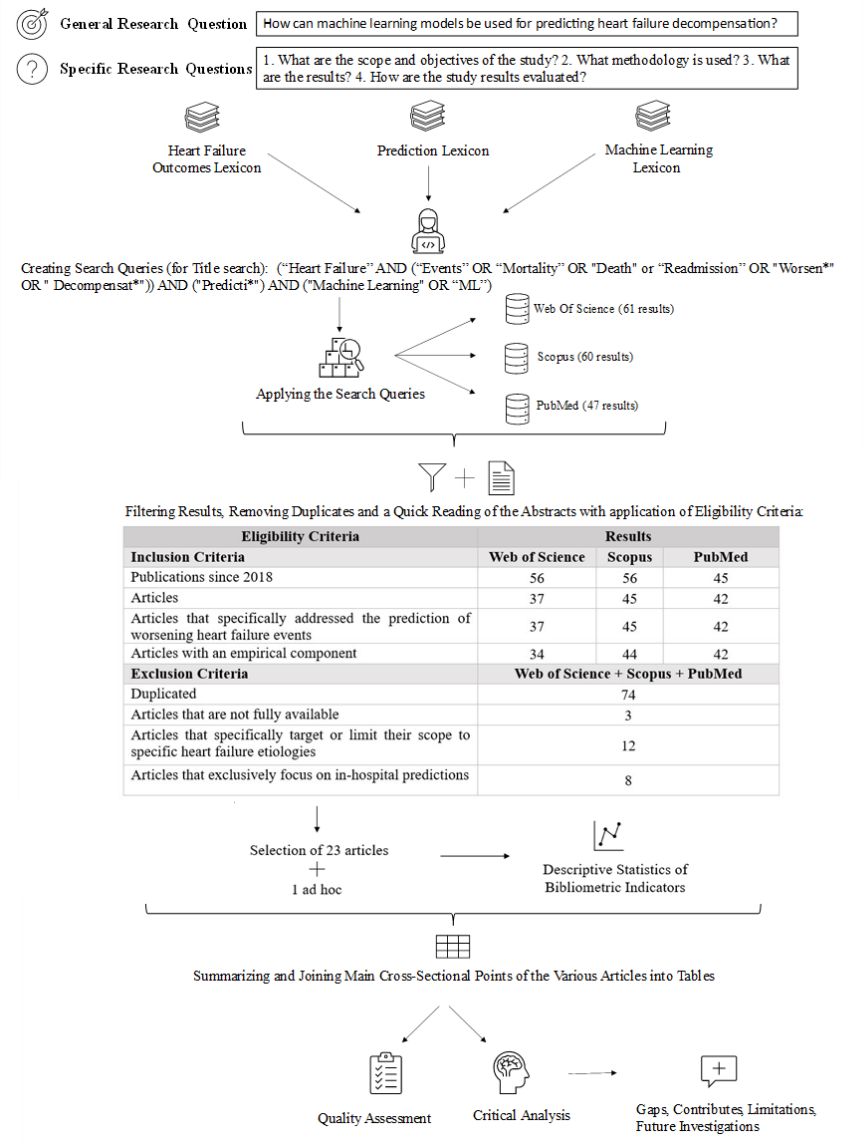
Given the research objective, this literature review seeks to address the following question: How can ML models be used for predicting HF decompensation? To answer this question four additional questions were necessary: 1) “What are the scope and objectives of the study?”, 2) “What methodology is used?”, 3) “What are the study results?”, 4) “How are the study results evaluated?”.

The articles included in this SLR are chosen from multiple databases comprising various scientific publications, considering the title, and inclusion/exclusion criteria. WoS (www.webofknowledge.com), Scopus (https://www.scopus.com), and PubMed (https://PubMed.ncbi.nlm.nih.gov/) have been selected for their renowned comprehensive coverage of high-impact journals across the fields of medicine, health sciences, and technology (Li et al., 2018; Paul et al., 2021; Pranckutė, 2021).

A Boolean search is conducted across multiple databases, focusing exclusively on the title field. The chosen keywords are derived from the literature and include terms associated with HF outcomes, prediction, and ML. The formulated query is as follows: ((“Heart Failure” OR HF) AND (“Events” OR “Mortality” OR "Death" OR “Readmission” OR "Worsen*" OR "Decompensat*")) AND ("Predicti*") AND ("Machine Learning" OR “ML”). The validity of this query is confirmed by two subject matter experts, a scholar and a cardiologist. This rigorous process results in the identification of one hundred and sixty-eight articles on the specified topic, plus one additional article identified through an ad hoc search.

Initially, filters are applied based on two criteria: publication years and document type. Articles from 2018 onward are included, aligning with the surge in interest and growth within the field, notably highlighted by the increasing number of ML models for HF populations, as observed in the study by Błaziak et al. (2022). After removing duplicates and articles that are not fully available, a review of abstracts results in the exclusion of articles not aligned with the research objectives (Paul et al., 2023) , particularly those focused on specific HF etiologies (e.g., Dai et al. (2022), which primarily targets sarcoidosis patients) and those exclusively centered on in-hospital predictions (e.g., Chen et al. (2023), which focuses solely on predicting in-hospital mortality). The inclusion criteria are concurrently applied, resulting in the selection of articles that specifically address the prediction of worsening HF events and those incorporating an empirical component.

Following the application of the eligibility criteria, a total of twenty-four articles are selected for review. The article by Kerexeta et al. (2023), which focuses on predicting HF decompensation through AI models, is not found through the query and is therefore added through ad hoc research.



Eligibility Criteria	Results		
Inclusion Criteria	**Web of Science**	**Scopus**	**PubMed**
Publications since 2018	56	56	45
Articles	37	45	42
Articles that specifically addressed the prediction of worsening heart failure events	37	45	42
Articles with an empirical component	34	44	42
Exclusion Criteria	**Web of Science + Scopus + PubMed**		
Duplicated	74		
Articles that are not fully available	3		
Articles that specifically target or limit their scope to specific heart failure etiologies	12		
Articles that exclusively focus on in-hospital predictions	8		

Figure 2.1: Article selection process

Chronologically arranged, the Table 2.1 features articles from 2019 to 2023 with a notable rise starting in 2021 (n=6) and peaking at ten articles in 2023. This trend highlights the growing effort to develop predictive models over the past five years, which is five times higher than in 2019, suggesting a promising trajectory for future expansion. The articles are spread across twenty different journals, with most (n=19) indexed in all three major scientific databases, Scopus being the primary contributor. Additionally, only two articles have authors in common (i.e., Awan et al., 2019a; Awan, et al., 2019b).

Table 2.1: Articles under study

ID	Year	Title	Journal	Authors	Source
1	2023	Machine Learning Based Readmission and Mortality Prediction in Heart Failure Patients	*Scientific Reports*	Sabouri, M.; Rajabi, A.B.; Hajianfar, G.; Gharibi, O.; Mohebi, M.; Avval, A.H.; Naderi, N.; Shiri, I.	W; S; P
2	2023	Heart Failure Emergency Readmission Prediction Using Stacking Machine Learning Model	*Diagnostics*	Rahman, M.S.; Rahman, H.R.; Prithula, J.; Chowdhury, M.E.H.; Ahmed, M.U.; Kumar, J.; Murugappan, M.; Khan, M.S.	W; S; P
3	2023	Comparing Machine Learning Classifiers for Predicting Hospital Readmission of Heart Failure Patients in Rwanda	*Journal of Personalized Medicine*	Rizinde, T.; Ngaruye, I.; Cahill, N.D.	W; S; P
4	2023	Comparison Of Linear and Non-Linear Machine Learning Models for Time-Dependent Readmission or Mortality Prediction Among Hospitalized Heart Failure Patients	*Heliyon*	Tong, R.; Zhu, Z.S.; Ling, J.	W; S; P
5	2023	The Price of Explainability in Machine Learning Models For 100-Day Readmission Prediction In Heart Failure: Retrospective, Comparative, Machine Learning Study	*Journal of Medical Internet Research*	Soliman, A.; Agvall, B.; Etminani, K.; Hamed, O.; Lingman, M.	W; S; P
6	2023	Interpretable Prediction Of 3-Year All-Cause Mortality in Patients with Chronic Heart Failure Based on Machine Learning	*BMC Medical Informatics and Decision Making*	Xu, C.G.; Li, H.X.; Yang, J.P.; Peng, Y.Z.; Cai, H.Y.; Zhou, J.; Gu, W.Y.; Chen, L.X.	W; S; P
7	2023	Predicting Mortality and Re-Hospitalization for Heart Failure: A Machine-Learning and Cluster Analysis on Frailty and Comorbidity	*Aging Clinical and Experimental Research*	Okoye, C.; Mazzarone, T.; Niccolai, F.; Bencivenga, L.; Pescatore, G.; Bianco, M.G.; Guerrini, C.; Giusti, A.; Guarino, D.; Virdis, A.	W; S; P
8	2023	Prediction and Analysis of Heart Failure Decompensation Events Based on Telemonitored Data and Artificial Intelligence Methods	*Journal of Cardiovascular Development and Disease*	Kerexeta, J.; Larburu, N.; Escolar, V.; Lozano-Bahamonde, A.; Macía, I.; Beristain Iraola, A.; Graña, M.	W; S; P
9	2023	Predicting Six-Month Re-Admission Risk in Heart Failure Patients Using Multiple Machine Learning Methods: A Study Based on The Chinese Heart Failure Population Database	*Journal of Clinical Medicine*	Chen, S.; Hu, W.; Yang, Y.; Cai, J.; Luo, Y.; Gong, L.; L,i Y.; Si, A.; Zhang, Y.; Liu, S.; Mi, B.; Pei, L.; Zhao, Y.; Chen, F.	P
10	2023	Mortality Prediction in Patients with or Without Heart Failure Using a Machine Learning Model	*JACC: Advances*	Jang, S.Y.; Park, J.J.; Adler, E.; Eshraghian, E.; Ahmad, F.S.; Campagnari, C.; Yagil, A.; Greenberg, B.	S
11	2022	A Comparative Study on Prediction of Survival Event of Heart Failure Patients Using Machine Learning Algorithms	*Neural Computing & Applications*	Özbay, Karakuş M.; Er, O.	S
12	2022	Predicting 30-Day Readmissions in Patients with Heart Failure Using Administrative Data: A Machine Learning Approach	*Journal of Cardiac Failure*	Sharma, V.; Kulkarni, V.; Mcalister, F.; Eurich, D.E.A.N.; Keshwani, S.; Simpson, S.H.; Voaklander, D.O.N.; Samanani, S.	S
13	2022	Machine Learning and LACE Index for Predicting 30-Day Readmissions After Heart Failure Hospitalization in Elderly Patients	*Internal and Emergency Medicine*	Polo Friz, H.; Esposito, V.; Marano, G.; Primitz L.; Bovio, A.; Delgrossi, G.; Bombelli, M.; Grignaffini, G.; Monza, G.; Boracchi, P.	S
14	2022	Predicting Long-Term Mortality in Patients with Acute Heart Failure by Using Machine Learning	*Journal of Cardiac Failure*	Park, J; Hwang, IC; Yoon, YE; Park, JB; Park, JH; Cho, GY	W; S; P
15	2022	Comparison Of Machine Learning and The Regression-Based EHMRG Model for Predicting Early Mortality in Acute Heart Failure	*International Journal of Cardiology*	Austin, D.E.; Lee, D.S.; Wang, C.X.; Ma, S.H.; Wang, X.S.; Porter, J.; Wang, B.	W; S; P
16	2021	Predicting 90 Day Acute Heart Failure Readmission and Death Using Machine Learning-Supported Decision Analysis	*Clinical Cardiology*	Sarijaloo, F.; Park, J.; Zhong, X.; Wokhlu, A.	W; S; P
17	2021	Prediction Model Using Machine Learning for Mortality in Patients with Heart Failure	*American Journal of Cardiology*	Negassa, A.; Ahmed, S.; Zolty, R.; Patel, S.R.	W; S; P
18	2021	Predicting Hospital Readmission in Heart Failure Patients in Iran: A Comparison of Various Machine Learning Methods	*Healthcare Informatics Research*	Najafi-Vosough, R.; Faradmal, J.; Hosseini, S.K.; Moghimbeigi, A.; Mahjub, H.	W; S; P
19	2021	Machine Learning-Based Model for Predicting 1 Year Mortality of Hospitalized Patients with Heart Failure	*ESC Heart Failure*	Tohyama, T.; Ide, T.; Ikeda, M.; Kaku, H.; Enzan, N.; Matsushima, S.; Funakoshi, K.; Kishimoto, J.; Todaka, K.; Tsutsui, H.	W; S; P
20	2021	Derivation Of an Electronic Frailty Index for Predicting Short-Term Mortality in Heart Failure: A Machine Learning Approach	*ESC Heart Failure*	Ju, C.S.; Zhou, J.D.; Lee, S.R.; Tan, M.S.; Liu, T.; Bazoukis, G.; Jeevaratnam, K.; Chan, E.W.Y.; Wong, I.C.K.; Wei, L.; Zhang, Q.P.; Tse, G.	W; S; P
21	2021	Prediction of Long-Term Hospitalisation and All-Cause Mortality in Patients with Chronic Heart Failure on Dutch Claims Data: A Machine Learning Approach	*BMC Medical Informatics and Decision Making*	van der Galiën, O.P.; Hoekstra, R.C.; Gürgöze, M.T.; Manintveld, O.C.; van den Bunt, M.R.; Veenman, C.J.; Boersma, E.	W; S; P
22	2020	Utilizing Electronic Health Data and Machine Learning for The Prediction Of 30-Day Unplanned Readmission or All-Cause Mortality in Heart Failure	*Cardiovascular Digital Health Journal*	Beecy, A.N.; Gummalla, M.; Sholle, E.; Xu, Z.; Zhang, Y.; Michalak, K.; Dolan, K.; Hussain, Y.; Lee, B.C.; Zhang, Y.; Goyal, P.; Campion, T.R., Jr.; Shaw, L.J.; Baskaran, L.; Al'Aref, S.J.	W; S; P
23	2019	Feature Selection and Transformation by Machine Learning Reduce Variable Numbers and Improve Prediction for Heart Failure Readmission or Death	*PLOS ONE*	Awan, S.E.; Bennamoun, M.; Sohel, F.; Sanfilippo, F.M.; Chow, B.J.; Dwivedi, G.	W; S; P
24	2019	Machine Learning-Based Prediction of Heart Failure Readmission or Death: Implications of Choosing the Right Model and the Right Metrics	*ESC Heart Failure*	Awan, S.E.; Bennamoun, M.; Sohel, F.; Sanfilippo, F.M.; Dwivedi, G.	W; S; P

Notes: P: PubMed; S: Scopus; W: Web of Science

To validate the alignment of the selected articles with the objectives of the SLR, a word cloud is generated (Figure 2.2), highlighting the most frequent terms in the abstracts. As expected, terms related to HF, HF patients, ML, and predictive models dominate. Interestingly, while events like readmission are present, mortality emerges as the most common target.

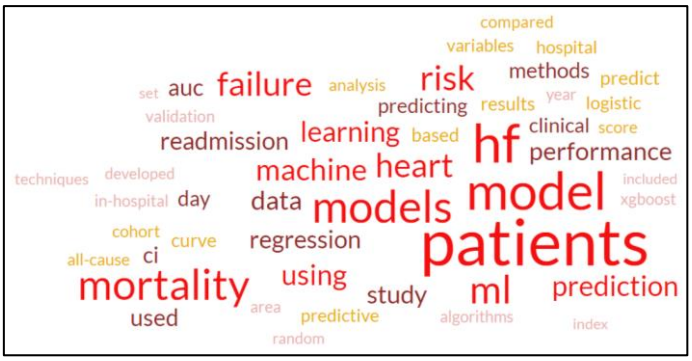


Figure 2.2: Word could of the abstracts

After reviewing the selected articles, a quality assessment is undertaken to evaluate their contribution to the research objective and specifically address the four specific questions. For each research question, evaluation criteria are defined and translated into operational questions. The articles are then evaluated against these criteria, with each criterion assigned to a numerical value based on the capability of the article to answer the research questions (Błaziak et al., 2022): 1 for full capability, 0.5 for partial capability, and 0 if the criteria were not met (Table 2.2).

Table 2.2: Quality criteria for article evaluation

	Id	Quality Criteria
What are the scope and objectives?	1.1	Does the article clearly describe and justify the objective of applying the model?
	1.2	Does the article clearly describe the scope of the study?
What methodology is used?	2.1	Is the sample utilized in the article relevant and clearly contextualized?
	2.2	Does the article provide a detailed description of the data collection and preparation phase?
	2.3	Does the article present and justify the variables used?
	2.4	Does the article provide, describe, and justify the employed techniques/algorithms?
	2.5	Does the article compare different predictive models?
What are the results?	3.1	Does the article clearly apply validation methods and various evaluation metrics?
	3.2	Does the article clearly present the results of the evaluation metrics?
	3.3	Does the article clearly identify and justify the best-performing model?
	3.4	Does the article clearly identify the most important variables?
	3.5	Does the article provide a detailed discussion of the results, including insights gained from the evaluation metrics?
What are the study impacts?	4.1	Does the article clearly describe the contributions of the study?
	4.2	Does the article clearly describe the limitations of the study?
	4.3	Does the article provide and justify the future recommendations?

2.2. Critical Synthesis of the Literature

The review process commences with an in-depth examination of each selected article. To systematically address the four research questions, relevant data is meticulously extracted and documented in dedicated Excel tables, where each column represents a partial answer contributing to the comprehensive response that the entire table aims to provide (Paul & Menzies, 2023).

The first table characterizes the scope and objectives of each article, including details like the country, study period, and data origin to provide an initial understanding. The following tables focus on the methodological aspects. One outlines the outcomes each study aimed to predict and their follow-up periods, while another details model characteristics such as sample size (noting if techniques to correct sample imbalance were applied), algorithms used, feature selection methods, and the number of independent variables. The evaluation process is divided into four tables, each focused on a different outcome, detailing the evaluation method, metrics used, and best-performing algorithm. Traditional models are also included if applied for comparison. The most significant variables are presented separately in another table, divided into nine dimensions, including demographics and clinical history. Lastly, two tables critically evaluate the study results, highlighting significant contributions and key stakeholders in one, and identifying limitations with recommendations for future investigations in the other.

Upon concluding the data extraction phase, the accumulated information is subjected to thorough interpretation. Employing both statistical and descriptive analyses, each table undergoes critical examination to reveal discernible patterns, identify valuable insights, and explore potential gaps within the existing literature (Paul et al., 2023).

2.2.1. Scope and Objectives

The selected studies cover a wide range of objectives and scopes, applied to a variety of demographic cohorts (Table 2.3).

Regarding the scope, the primary focuses were mortality (n=8) and readmission (n=7), each similarly represented. Some studies addressed both outcomes separately (n=2) (Sabouri et al., 2023; van der Galiën et al., 2021), while others aimed to predict a combined endpoint (CEP) (n=6) (e.g., Okoye et al., 2023), and only one study specifically addressed decompensation (Kerexeta et al., 2023).

The studies span various follow-up periods, with mortality predictions ranging from in-hospital (Sabouri et al., 2023) to death within 7 days (Austin et al., 2022) up to 3 years

(e.g., Park et al., 2022; Xu et al., 2023). Readmission predictions extend from 20 days (Rizinde et al., 2023) up to 3 years (van der Galiën et al., 2021), while CEP outcomes are between 30 days (e.g., Awan et al., 2019a) and 6 months (e.g., Beecy et al., 2020). Mortality studies, by comparison, often include longer follow-up periods, though 30 days is a frequent benchmark across outcomes. Decompensation is uniquely predicted within a 7-day period (Kerexeta et al., 2023). Notably, only one study applied a time-dependent prognosis approach, which does not define a specific time frame but instead focuses on understanding the evolving risk profile over time without a fixed endpoint (Tong et al., 2023).

While all the studies revolve around prediction models, not all of them have the construction of predictive models for mortality and readmission as their final objective. Some aim to identify the risk factors associated with a specific outcome (e.g., Awan et al., 2019b; Tohyama et al., 2021), others seek to create clusters according to the risk of the outcome (Okoye et al., 2023), and others want to compare the model performance with more traditional ML models (e.g., Negassa et al., 2021) or preexisting cardiologic scores/models already widely used (e.g., Austin et al., 2022; Sharma et al., 2022). Additionally, some studies aim to provide explainability of the models, offering insights into why certain predictions are made, more focused on the predictive value (e.g., (Soliman et al., 2023; Xu et al., 2023).

Most reviewed studies used a single-center investigation approach, with a subset (n=10) opting for a multi-center methodology (e.g., Austin et al., 2022; Tohyama et al., 2021), i.e. the data was collected from various hospitals within the same country. However, a wide range of time-periods was observed, from eight months (Karakuş & Er, 2022) to 11 years (Rizinde et al., 2023). Most studies (n=16) had collection periods of 3 years or higher, indicating a significant focus on long-term data collection.

Previous studies show that HF has different prevalences across racial or ethnic groups (Guo et al., 2021; Lewsey & Breathett, 2021), highlighting the importance of cultural diversity in a study of this nature. The selected articles span 13 countries across five continents, each contributing unique healthcare systems, patient demographics, and research capacities to HF outcome prediction. China stands out with the highest number of studies (n=5).

Table 2.3: Context and scope of the studies

ID	Scope	Study Objective	Data Origin	Period	Country
1	In-hospital and 6-month mortality; 30- and 90-day readmission	Predict mortality and readmission, based on conventional features.	SC	2012-2018	Iran
2	6-month readmission	Predict emergency readmissions by leveraging ML models to develop a stacking ML model based on EHR data.	SC	2016-2019	China
3	20-day readmission	Compare ML models to evaluate their performance in predicting which HF patients are at high risk for readmission after discharge.	MC (7)	2008-2019	Rwanda
4	Readmission or mortality	Predict readmission or mortality, using conventional statistics and ML to improve time-dependent prognosis understanding.	SC	2016-2019	China
5	100-day readmission	Compare deep learning with traditional ML models in the prediction of readmission risk for HF patients, while providing explanations for the predictions.	MC (-)	2017-2019	Sweden
6	3-year all-cause mortality	Evaluate ML models for predicting mortality and develop an interpretable model with individualized risk assessments and explanations.	SC	2017-2020	China
7	6-month mortality or readmission	Identify risk factors for re-hospitalization or death and determine high-risk clinical phenotypes in older patients using unsupervised ML.	SC	2018-2019	Italy
8	7-day decompensation	Develop an AI model to predict decompensation using telemonitoring data and analyze variable significance for HF progression.	SC	2014-2022	Spain
9	6-month readmission	Develop an interpretable predictive model to forecast readmission in the HF Chinese population.	SC	2016-2019	China
10	1-year mortality	Assess the ability of MARKER-HF to predict mortality in patients both with and without HF, in patients with and without cardiovascular disease, and in sub-groups with various common medical conditions.	SC	2013-2020	South Korea
11	4 to 285 day mortality	Determine the life-threatening risk associated with HF in patients using classification-based ML techniques.	SC	2015	Pakistan
12	30-day readmission	Develop and validate ML models to predict unplanned readmissions in a HF cohort using administrative health data and compare their performance with the LaCE score.	MC (-)	2012-2019	Canada
13	30-day readmission	Assess the ability of ML algorithms to predict readmissions and compare their prognostic performance with the LACE index.	SC	2010-2019	Italy
14	3-year mortality	Develop a ML- based risk-prediction model for mortality in patients with acute HF.	MC (3)	2009-2016	South Korea
15	7- and 30-day mortality	Compare ML algorithms with the regression-based EHMRG model in predicting mortality	MC (86)	2004-2007	Canada
16	90-day readmission or mortality	Develop a ML-based model integrating various EHR risk factors to predict readmission or all-cause mortality.	SC	2011-2019	USA
17	30-day mortality	Develop a ML ensemble model to predict mortality, comparing its effectiveness against the logistic model	SC	2001-2010	USA
18	Readmission	Compare the performance of six ML methods for predicting hospital readmission in HF patients.	SC	2015-2019	Iran
19	1-year mortality	Predict mortality using ML algorithms and administrative claim data, comparing performance with conventional risk models and identifying top predictors for accurate ML-based prediction.	MC (128)	2013	Japan
20	30-day and 90-day mortality	Improve short-term mortality prediction by integrating NLR and PNI into an electronic frailty index using ML.	MC (9)	2013-2017	China
21	1- and 3-year readmission; 1- and 3-year all-cause mortality	Assess the predictive value of Insurance claim data for hospitalization and mortality, utilizing both ML and traditional statistical techniques.	SC	2012-2014	Netherlands
22	30-day readmission or mortality; 6-month readmission or mortality	Create a method for predicting readmissions and mortality by combining clinical and physiological data from EHR.	SC	2008-2018	USA
23	30-day readmission or mortality	Apply ML to identify key variables for predicting readmission or death, using feature extraction to simplify and improve performance.	MC -	2003-2008	Australia
24	30-day readmission or mortality	Evaluate ML models for predicting readmission or death using administrative data, addressing class imbalance and improving accuracy over regression methods	MC -	2003-2008	Australia

Notes: AI: Artificial intelligence; EHMRG: Emergency Heart Failure Mortality Risk Grade; EHR: Electronic Health Record; HF: Heart Failure; LaCE: Length of stay, Age, Comorbidity, and Emergency visit; ; LACE: Length of stay, Acuity of admission, Comorbidity, and Emergency visit; MARKER-HF: Machine learning Assessment of RisK and EaRly mortality in Heart Failure; MC: Multicenter; ML: Machine Learning; NLR: Neutrophil-to-Lymphocyte Ratio; PNI: Prognostic Nutritional Index; SC: Single-center.; USA: United States of America

2.2.2. Methodology

Understanding the methodology employed is crucial for assessing the reliability of research findings and enables replication by other researchers, promoting transparency and advancing scientific knowledge.

The vast majority of studies (91.7%) present a binary dependent variable, where 1 indicates if mortality/readmission/CEP/decompensation occurs during the follow-up period, and 0 indicates otherwise. These variables are reflected in Table 2.4. The studies that did not have a binary target are the ones that opted for a survival analysis investigation. The study conducted by Tong et al. (2023), utilized survival time as the dependent variable, representing the number of days from the index hospital admission to the occurrence of the event, which could be either death or readmission. In contrast, Jang et al. (2023) employed a composite outcome approach, where one category indicated death within 90 days, while the other represented survival time beyond a period of 2 years. Exposing the algorithm to the two extreme outcomes represented an intermediate step in constructing the MARKER-HF [2] risk score (Adler et al., 2020).

Table 2.4: Dependent variables for binary outcome studies

Dependent Variables		**Study ID**																					
	Follow-up Period	1	2	3	5	6	7	8	9	11	12	13	14	15	16	17	18	19	20	21	22	23	24
D	7-day							✓															
Mortality	In-hospital	✓																					
	7-day													✓									
	30-day													✓		✓			✓				
	90-day																		✓				
	4-285 day									✓													
	6-month	✓																					
	1-year																	✓		✓			
	3-year					✓							✓							✓			
Readmission	Und.																✓						
	20-day			✓																			
	30-day	✓									✓	✓											
	90-day	✓																					
	100-day				✓																		
	6-month		✓						✓														
	1-year																			✓			
	3-year																			✓			
CEP	30-day																				✓	✓	✓
	90-day														✓								
	6-month						✓														✓		

Note: D: Decompensation; Und.: Undefined

[2] In medical research, these models are commonly recognized by their acronyms, therefore, the full names are provided only in the acronym list.

In Table 2.5**Table *2.5***, an overview of the methodology applied to predict these outcomes is presented. The predominant focus of the studies lies in classification tasks (n=21), except for two dedicated to survival analysis (Jang et al., 2023; Tong et al., 2023) and another aimed at cluster creation (Okoye et al., 2023).

An examination of dataset sizes reveals a range from 299 (Karakuş & Er, 2022) to 41749 (Jang et al., 2023) HF patients, with an average of 7612 individuals per study. Despite the prevalent challenge of imbalanced data in the medical field (Tasci et al., 2022), only 37.5% of the articles implemented strategies to mitigate this issue. Among the approaches employed, oversampling was the most common method with three of the four articles that resourced to this methos opting for the utilization of the synthetic minority over-sampling technique (SMOTE) to augment the minority class (e.g., Rizinde et al., 2023; Sabouri et al., 2023). The remaining studies opted for a weight adjustment (Awan et al., 2019a; Sharma et al., 2022; van der Galiën et al., 2021), with two going with an under-sampling technique (Kerexeta et al., 2023; Polo Friz et al., 2022). Interestingly, some studies only applied these techniques to specific algorithms due to of their properties (e.g., Austin et al. (2022) only applied random oversampling to the neural network algorithm).

A diverse range of algorithms has been utilized across the studies, with only six of them opting not to compare different ML algorithms for their final model. However, these studies still conducted various other types of comparisons. For instance, some studies compared algorithms as variable selection options (e.g., Okoye et al., 2023; Sarijaloo et al., 2021), while others assessed model performance against the pre-existing models (e.g., Park et al., 2022). Additionally, some studies investigated variations in algorithm performance under different circumstances. For example, examined the same algorithm with different feature selection techniques (e.g., Awan et al., 2019), explored its performance with different independent variables (e.g., Beecy et al., 2020), and analyzed its performance across different patient cohorts (e.g., Jang et al., 2023). These additional comparisons were also part of some of the other studies that did have comparisons between different algorithms. Regarding the specific algorithms that were employed, random forest (RF) was the most used algorithm appearing in 15 studies (e.g., Rahman et al., 2023; Sabouri et al., 2023) followed closely by LR (n=12).

Other risk prediction scores and models have been developed to assist clinicians in assessing patient outcomes across different medical conditions and are highlighted in Table 2.5 in bold to distinguish them from the other algorithms. The LACE (Van Walraven et al., 2010), LaCE (Au et al., 2012), and HOSPITAL (Donze et al., 2013) scores are versatile tools used to predict hospital readmission risk across a range of patients and conditions. In contrast, models

like MARKER-HF, BIOSTAT-CHF (Voors et al., 2017), SHFM (Levy et al., 2006), AHEAD (Spinar et al., 2016), EHMRG (Lee et al., 2012) and MAGGIC (Pocock et al., 2013) risk scores are tailored for HF patients, focusing on the prediction of mortality and/or readmission outcomes. Several studies (n=8) aimed to compare their findings with these established and validated scores, while Jang et al. (2023) exclusively utilized the MARKER-HF model to assess its performance as a universal scoring system applicable beyond HF patients.

In over half of the studies (n=15), feature selection techniques were employed (e.g., Awan et al., 2019b; Sabouri et al., 2023). The remainder either relied on existing literature (Sharma et al., 2022a; Xu et al., 2023) and clinical expertise (Soliman et al., 2023), or simply utilized variables available in the dataset (Karakuş & Er, 2022; Najafi-Vosough et al., 2021). For those studies utilizing pre-existing scores, variables were pre-defined (Jang et al., 2023). Once again RF emerges as the most employed technique, underscoring its versatility and effectiveness in data analysis.

Naturally, the number of independent variables varies from study to study, ranging from only 8 (Jang et al., 2023) variables up to 2032 (Beecy et al., 2020). The variables encompass demographics, clinical history, vital signs, physical exam findings, including electrocardiogram (TTE) and echocardiogram (ECG) results, treatments, and comorbidities. Unlike the rest, two studies adjusted the number of independent variables based on the specific model being applied (Austin et al., 2022; Beecy et al., 2020).

.

Table 2.5: Methodology of the studies

ID	Sample*	Problem	Algorithm	Feature Selection Techniques	Ind. Var.
1	737 (Y)	C	*KNN; LR; MLP; NB; QDA; RF; SVM; XGBoost*	RFE; MRMR; Boruta	34
2	2008 (N)	C	*MLP; LDA; GB; RF; LR; SVM; ET; AdaBoost; KNN; CatBoost; LGB; EN; Stacking Model (XGBoost)*	XGBoost; RF; ExtraTrees	166
3	4083 (Y)	C	*MLP; LR; DT; KNN; RF; SVM*	ExtraTrees	75
4	1976 (N)	SA	*Multivariate CR; LASSO CR; RF (SA); GB (SA)*	Univariate CR; LASSO; Feautre Importance (RF, GB)	146
5	15612 (N)	C	*CatBoost; LSTM;* **LACE**	Clinical expertise; data science knowledge	40
6	626 (N)	C	*LR; KNN; RF; NB; DT; XGBoost*	Literature	45
7	571 (N)	Clust.	-	RF; Multivariable CR	ND
8	488 (Y)	C	*LR; Bagging; XGBoost; SVM; RF; ET; Ada Boost; GB*	Boruta	42
9	2002 (N)	C	*LR; DT (CART); XGBoost; NB; SVM; RF*	Combination of single-factor and multi-factor regression with FDR correction; LASSO; RF	12
10	41749 (N)	SA	*Boosted DT;* **MARKER-HF**	Pre-defined	8
11	299 (N)	C	*LR; NB; LinSVM; CubSVM; QuadSVM; FG-SVM; MG-SVM; CG-SVM; FKNN; MKNN; CKNN; CosKNN; CubKNN; WKNN; SKNN; BT; BagT; FT; MT; CT; RUSBT; SubD; MLNN(1–2–3HL)*	Available variables	12
12	9845 (Y)	C	*XGBoost; GB; AdaBoost; CatBoost; LGB; LinSVC; Gaussian NB; RF; DT; LR (L1); NN; LSTM;* **LaCE**	Literature	+/- 160
13	3079 (Y)	C	*AdaBoost; GB; XGBoost; RF;* **LACE**	Lasso	
14	4312 (N)	C	*CoxBoost;* **BIOSTAT-CHF; AHEAD Score**	RF minimal depth evaluation	27
15	12608 (Y)	C	*NN; RF; XGBoost; LASSO;* **LR (EHMRG)**	RFE (XGBoost; RF); Feature importance (EN; RF)	111/10
16	3189 (N)	C	*LASSO-LR*	SVM; RF; GB; LASSO	98
17	7516 (N)	C	*LR, LR with VS, LASSO, RF, MARS, Mean, Bayesian Logistic, Boosting, Bagging, SuperLearner*	Literature	34
18	1856 (Y)	C	*SVM; LS-SVM; Bagging; RF; AdaBoost; NB*	Available variables	46
19	10175 (N)	C	*LASSO; SVM, RF; GBDT; VC; NN;* **SHFM; MAGGIC**	Permutation feature importance	89
20	8893 (N)	C	*GB; DT; LR*	Literature	16
21	25776 (Y)	C	*LR; EN; RF; NN*	Prevalence prioritization; LASSO	150
22	3774 (N)	C	*XGBoost;* **HOSPITAL Score**	Goodman-Kruskal Tau; Chi-Square test	796/2032
23	10575 (N)	C	*MLP*	t-test; Chi-Squared test; SFS; SBS; mRMR; PCA	47
24	10575 (Y)	C	*MLP, LR; RF; DT; SVM (weigthed RF, DT, SVM);* **LACE**	-	47

Notes: AdaBoost: Adaptive Boosting; Diabetes mellitus; BagT: Bagging Trees; BT: Boosting Trees; C: Classification; CART: Classification and Regression Trees; CatBoost: Categorical Boosting; CG-SVM: Coarse Gaussian Support Vector Machine; Clust.: Clustering; CKNN: Coarse K-Nearest Neighbors; CosKNN: Cosine K-Nearest Neighbors; CR: Cox Regression; CubKNN: Cubic K-Nearest Neighbors; CubSVM: Cubic Support Vector Machine; DT: Decision Tree; EN: Elastic Net; ET: Extra Trees; FDR: False Discovery Rate; FG-SVM: Fine Gaussian Support Vector Machine; FT: Fine Trees; GB: Gradient Boosting; GBDT: Gradient Boosting Decision Tree; KNN: K-Nearest Neighbors; LASSO: Least Absolute Shrinkage and Selection Operator; LDA: Linear Discriminant Analysis; LGB: Light Gradient Boosting; LR: Logistic Regression; LS-SVM: Least Squares Support Vector Machine; LSTM: Long Short-Term Memory; LinSVC: Linear Support Vector Classifier; LinSVM: Linear Support Vector Machine; MARS: Multivariate Adaptive Regression Splines; MLP: Multilayer Perceptron; MLNN (1–2–3HL): Multilayer Neural Network with 1 to 3 Hidden Layers; MG-SVM: Medium Gaussian Support Vector Machine; MKNN: Medium K-Nearest Neighbors; MRMR: Minimum Redundancy Maximum Relevant; MT: Medium Tree; NB: Naive Bayes; NN: Neural Network; QDA: Quadratic Discriminant Analysis; RUSBT: Random Under Sampling Boosted Trees; RF: Random Forest; RFE: Recursive Feature Elimination; RSF: Random Survival Forest; SA: Survival analysis; SBS Sequential backward selection; SFS: Sequential forward selection; SKNN: Subspace K-Nearest Neighbors; SVM: Support Vector Machine; SubD: Subspace Discriminant; VC: Voting Classifier; VS: Variable Selection; WKNN: Weighted K-Nearest Neighbors; XGBoost: Extreme Gradient Boosting; * Y indicates if the study applied techniques to address class imbalance; N indicates if not

2.2.3. Results

For results evaluation, four tables were devised. These tables categorized the models according to outcomes, including decompensation (Table 2.6), mortality (Table 2.7), readmission (Table 2.8), and CEP (Table 2.9**Table *2.9***). Additionally, the models were arranged based on their prediction period, which provides a structured framework for presenting the data, contributing to improved organization and readability.

The predominant validation method (n=9) involved a combination of hold out, which partitions the dataset into training and test sets, and k-fold cross-validation, which divides the dataset into k subsets, using each as a validation set while training on the rest (e.g., Özbay Karakuş & Er, 2022; Tong et al., 2023). Additionally, six studies exclusively used hold-out validation, with only two studies incorporating a separate validation set in addition to the training and test sets (Awan et al., 2019a; Awan, et al., 2019b). Among the k-fold cross-validation techniques, the 5-fold cross-validation method was the most commonly employed (n=7). Notably, two articles did not mention any validation method (Jang et al., 2023; Sarijaloo et al., 2021).

Most studies included various evaluation metrics (n=19) with AUC (Area under the ROC curve), also referred to as c-statistics (e.g.,Tohyama et al., 2021) or c-index (Negassa et al., 2021; Tong et al., 2023), being the most reported metric (n=21), with values ranging from 0.59 (Sabouri et al., 2023) to 1 (Karakuş & Er, 2022) across the various outcomes. On average, the models for mortality prediction achieved a higher AUC of 0.82, followed by readmission with 0.74, and CEP with 0.68. Additionally, four studies also incorporate calibration metrics, such as the Brier Score (Austin et al., 2022; Tohyama et al., 2021) and its variations, that is Integrated Brier score (Tong et al., 2023) and Scaled Brier Index (Negassa et al., 2021), among others. Calibration in predictive modelling assesses the accuracy of predicted probabilities relative to observed outcomes, distinct from discrimination metrics, which estimate the ability of the model to differentiate between outcomes (Jiang et al., 2012). In contrast, six studies were noted for their absence of evaluation graphs (e.g., Tohyama et al., 2021; van der Galiën et al., 2021), limiting the visual representation of their findings.

For additional insight, in articles that compared ML models with pre-existing ones, the evaluation metrics of both were considered. In cases where studies report multiple top-performing models, the model emphasized as the best is highlighted in bold. Only in one instance was the ML model surpassed, specifically by the EHMRG score for 30-day mortality prediction (Austin et al., 2022). Notably, Extreme Gradient Boosting (XGBoost) was

highlighted as the top-performing model in six studies, three of which focused on readmission prediction (e.g., Polo Friz et al., 2022; Sharma et al., 2022) and the only study on decompensation (Kerexeta et al., 2023). For the other two outcomes, there appears to be a tie, with multiple models emerging as top performers. For mortality, LR, elastic net (EN), and gradient boosting (GB) each achieve the best performance twice, as shown by Sabouri et al. (2023), van der Galiën et al. (2021) and Ju et al. (2021), respectively. Similarly, for the CEP, RF, XGBoost, and Multilayer Perceptron (MLP) each stand out twice, as seen in Okoye et al. (2023), Beecy et al. (2020) and Awan et al. (2019b), respectively.

Table 2.6: Evaluation of the models used for decompensation prediction

Follow-up Period	ID	Evaluation Method	Evaluation Metrics	Comparison Metrics	Evaluation Graphs	Best Performance
7-day	8	Hold out:80/20; 10-fold CV	AUC=0,694	-	✓	XGBoost

Notes: AUC: Area under the curve; CV: Cross-validation; XGBoost: eXtreme Gradient Boosting

Table 2.7: Evaluation of the models used for mortality prediction

Follow-up Period	ID	Evaluation Method	Evaluation Metrics	Comparison Metrics	Evaluation Graphs	Best Performance
In- Hospital	1	Hold out:70/30; 10-fold CV	AUC=0.90-0.91; A=0.84-0.85; S=0.84-0.85; R=0.83		✓	RFE-LR; Boruta-LR
7-days	15	stratified random sampling; Hold-out 80/20	AUC=0,794; BS=0.021; CIL=0.017; CS=0.954	LASSO: AUC=0,774 BS=0.021; CIL=-0.264; CS=0.882	✓	LASSO; **EHMRG**
30-days	15	stratified random sampling; Hold-out 80/20	AUC=0,759; BS=0.063; CIL=-0.863; CS=1.025	EHMRG: AUC=0,755; BS=0.056; CIL=-0.026; CS=0.964	✓	**XGBoost;** EHMRG
	17	Hold out: 55.7/44.3; 10-fold CV	AUC=0.83; SBI=0.084; RP=0.000-0.171; IDI=0.024;	LR: AUC=0.79; R2=0.168; SBI=0.074; RP=0.003-0.162; IDI=0.022	✓	Super Learner
	20	5-fold CV;	AUC=0.90; P=0.91; R=0.89		×	GB
90-days	20	5-fold CV;	AUC=0.92; P=0.94; R=0.93		×	GB
4 to285-days	11	Hold out: 70/30; k-fold (3-fold,5-fold) CV	AUC=0.87-1; A=58.5%-100%		✓	MLNN (2 HL); FG-SVM; FKNN; WKNN; SKNN; BT; BagT
6-Months	1	Hold out:70/30; 10-fold CV;	AUC=0.59-0.61; A=0.61-0.63; S=0.44-0.48; R=0.63-0.66		✓	MRMR-LR; MRMR-NB
1-year	19	hold out 80:20, 5-fold CV	AUC=0.777; A=70.7%; R=71.1%; S=70.6%; BS=0.121	SHFM: AUC=0.713; A=80.5% R=26.6%; S=91.7%; BS=0.139 MAGGIC: AUC=0.726%; A=70.4%; R=58.7%; S=72.9%; BS=0.130%	✓	VC
	21	CV	AUC=0.7866; R=74.2; S=69.7;		×	EN
	10	ND	AUC =0.770		✓	MARKER-HF (patients without HF)
3-years	6	5-fold CV;	AUC=0.82; A=78.96%; P=98%; R=99.44%		✓	RF
	14	Hold out: 62.7/37.3; 5-fold CV	AUC=0,760-0,761	BIOSTAT-CHF: AUC=0.714-0.715; AHEAD: AUC=0.656-0.681	✓	CoxBoost
	21	CV	AUC=0.7911; R=64.4; S=78.1		×	EN

Notes: A: Accuracy; AHEAD: Atrial fibrillation, Hemoglobin, Elderly , Abnormal renal parameters, Diabetes mellitus; AUC: Area Under the Curve; BagT: Bagging Trees; BIOSTAT-CHF: BIOlogy Study to TAilored Treatment in Chronic Heart Failure; BS: Brier Score; BT: Boosted Trees; CIL: Calibration-in-the-large; CS: Calibration Slope; CV: Cross-Validation; EHMRG: Emergency Heart Failure Mortality Risk Grade; EN: Elastic Net; FG-SVM: Fine Gaussian Support Vector Machine; FKNN: Fuzzy k-Nearest Neighbors; GB: Gradient Boosting; IDI:

Integrated discrimination index; LASSO: Least Absolute Shrinkage and Selection Operator; LR: Logistic Regression; MAGGIC: Meta-Analysis Global Group in Chronic Heart Failure; MLNN (2 HL): Multilayer Neural Network with 2 Hidden Layers; MRMR: Minimum Redundancy Maximum Relevant; NB: Naive Bayes; P: Precision; R: Recall; RF: Random Forest; RP: Range of prediction ; S: Specificity; SBI: Scaled Brier Index; SHFM: Seattle Heart Failure Model; SKNN: Soft k-Nearest Neighbors; SVM: Support Vector Machine; VC: Voting Classifier; WKNN: Weighted K-Nearest Neighbors

Table 2.8: Evaluation of the models used for readmission prediction

Follow-up Period	ID	Evaluation Method	Evaluation Metrics	Comparison Metrics	Evaluation Graphs	Best Performance
Undefined	18	CV	A=0.90-0.91; R=0.69-0.72; S=0.94-0.95; PPV=0.73-0.78; NPV=0.93-0.94		×	RF (median imputation method; multiple imputation method)
20-days	3	Hold out:80/20	AUC=94%; A=87%; P=84%; R=89%; F1=87%		✓	RF
30-days	1	Hold out:70/30; 10-fold CV;	AUC=0.71-0.73; A=0.68-0.81; S=0.69-0.85; R=0.50-0.63		✓	Boruta-SVM; MRMR-LR
	12	Hold out:80/20; 10-fold CV	AUC=0.654-0.685; PPV=0.21-0.62; LR+=1.00-6.12	LaCE: AUC=0.570; PPV=0.21-0.24; LR+=1.00-1.20	✓	XGBoost
	13	Hold out:80/20	AUC=0.803; R=0.78; S=0.75; A=0.77	LACE: AUC=0.504; R=0.54; S=0.59; A=0.0.59	✓	XGBoost
90-days	1	Hold out:70/30; 10-fold CV;	AUC=0.60; A=0.63; S=0.66; R=0.53		✓	MRMR-KNN
100-days	5	Hold out:88.39/11.61; 10-fold CV	AUC=0.66; R=0.83; S=0.50; F1=0.60; AUPRC=0.68	LACE: R=0.35; S=0.78; AUC=0.56; F1=0.39; AUPRC=0.51	×	Catboost
6-months	2	CV	AUC=0.861- 0.881; A=84.9%-89.41%; P=84.92%-90.10%; R=84.9%-89.41%; S=84.07%-87.83%; F1=84.84%-89.28%,		✓	Catboost; Adaboost; GB; **Stacking Model (XGBoost)**
	9	Hold out: 70/30	AUC=0.634; A=0.652; R=32.4%; S=84.4%		✓	LR
1-year	21	CV	AUC=0.7320; R=71.0; S=66.3		×	EN
3-years	21	CV	AUC=0.7330; R=67.8; S=67.7		×	EN

Notes: A: Accuracy; AdaBoost: Adaptive Boosting; AUC: Area under the ROC Curve; AUPRC: Area under the precision-recall curve; CatBoost: Categorical Boosting; CV: Cross-validation; EN: Elastic Net; F1: F1-score; GB: Gradient Boosting; KNN: K-Nearest Neighbors; LACE: Length of stay, Acuity level, Comorbidity, and Emergency visit; LaCE: Length of stay, Age, Comorbidity, and Emergency visit; LR: Logistic Regression; LR+: Positive likelihood ratio; MRMR: Minimum Redundancy Maximum Relevant; NPV: Negative predictive value; P: Precision; R: Recall; RF: Random Forest; PPV: Positive Predictive Value; S: Specificity; SVM: Support Vector Machine;

Table 2.9: Evaluation of the models used for CEP prediction

Follow-up Period	ID	Evaluation Method	Evaluation Metrics	Comparison Metrics	Evaluation Graphs	Best Performance
Undefined	4	5-fold CV;	AUC=0.587-0.61; TDAUC=0.617-0.641; IBS=0.164-0.166;		✓	RSF; GB
30-days	22	Hold out:90/10; 5-fold CV;	AUC=0.756	HOSPITAL score: AUC=0.666	✓	XGBoost (feature-aggregated)
	23	Hold out: 70/15/15	AUC= 0.62; R=58.7%; S=60.6% (original model with 47 var. + PCA= 0.66)		✓	MLP + mRMR
	24	Hold out: 70/15/15	AUC= 0.628; AUPRC=0.461; A=64.93%; R=48.42%; S=70.01	LACE: AUC=0.551; 0.448; A=59.85; R=45.54; S=64.80	×	MLP
90-days	16	N.D.	AUC=0.760; R=83%; S=56%; A=63%; PPV=38% (minimum cost threshold)		✓	LASSO+LR
6-Months	7	Hold out: 70/30	N.D.		×	RF
	22	Hold out: 90/10; 5-fold CV;	AUC= 0.701	HOSPITAL score AUC=0.654	✓	XGBoost (discharged index)

Notes: A: Accuracy; AUC: Area under the ROC Curve; AUPRC: Area under the precision-recall curve; CV: Cross-validation; GB:Gradient Boosting ; HOSPITAL: Hemoglobin, discharge from Oncology service, Sodium level, Procedure during the index admission, Index Type of admission, number of Admissions, and Length of stay; IBS: Integrated Brier score; LACE: Length of stay, Acuity level, Comorbidity, and Emergency visit; ; LASSO: Least Absolute Shrinkage and Selection Operator; LR: Logistic Regression; MLP: Multilayer Perceptron; MRMR:

Minimum Redundancy Maximum Relevant; ND: Not Disclosed R: Recall; RF: Random Forest; RSF: Random Survival Forest; PPV: Positive Predictive Value; S:Specificity; TDAUC: Time-Dependent AUC; XGBoost: Extreme Gradient Boosting

Over 83% of the articles identified key variables that enhance predictive ability and provide insights into the dependent variable. Focusing on the 20 articles that explicitly identified these variables, 87 are identified and organized into 8 groups: demographics (n=9), clinical history (n=15), medical status (n=9), lab results (n=23), vital signs (n=3), symptoms (n=4), treatments (n=4), TTE/ECG (n=12), and comorbidities (n=8)[3]. Table 2.10 displays the most important variables mentioned in at least two of these articles.

Table 2.10: Most important variables identified across the studies

Most Important Variables	Study ID																			Total
	1	2	3	4	5	6	7	8	9	12	13	14	15	16	18	19	20	21	22	23	-
Demographics																					-
Age	✓		✓		✓	✓	✓					✓			✓	✓	✓	✓		✓	11
Sex			✓													✓		✓			3
Clinical History																					
AT		✓																		✓	2
LOS			✓		✓				✓							✓					4
Nº Hospitalizations					✓	✓				✓	✓										4
Medical Status																					-
Type of HF		✓							✓												2
CCI		✓			✓		✓														3
NYHA						✓			✓												2
Lab Results																					-
BNP						✓	✓														2
BUN															✓				✓		2
Cr	✓						✓								✓				✓		4
GFR						✓						✓									2
Hb	✓	✓								✓									✓		4
Htc									✓										✓		2
NLR		✓															✓				2
NT-proBNP					✓							✓		✓							3
RBC				✓															✓		2
Na												✓	✓	✓	✓						4
Uric acid	✓			✓					✓												3
Vital signs																					-
HR			✓								✓			✓							3
DBP			✓			✓						✓		✓							4
SBP	✓		✓			✓								✓							4
Symptoms																					-
Edema	✓							✓													1
Treatments																					-
Inotropic Support	✓													✓							2
Diuretic use														✓		✓			✓		3
TTE/ECG																					-
LVEF														✓	✓						2
Comorbidities																					-
DM									✓									✓			2
CKD	✓				✓											✓				✓	4

Notes: AT: Admission type; BNP: B-type Natriuretic Peptide; BUN: Blood Urea Nitrogen; CCI: Charlson Comorbidity Index; CKD: Chronic Kidney Disease; Cr: Creatinine ; DBP: Diastolic Blood Pressure; DM: Diabetes Mellitus; Electro/Echo: Electrocardiogram/Echocardiogram; GFR: Glomerular Filtration Rate; Hb: Hemoglobin; HF: Heart Failure; HR: Heart Rate; Htc: Hematocrit; LOS: Length of stay; LVEF: Left Ventricle Ejection Fraction; NLR: Neutrophil To Lymphocyte Ratio; NT-proBNP: Na: Sodium; N-Terminal pro–B-type Natriuretic Peptide; NYHA: New York Heart Association Functional Classification; RBC: Red Blood Cell Count; SBP: Systolic Blood Pressure; TTE/ECG: Electrocardiogram/ Echocardiogram.

[3] Appendix A lists all variables identified as important in the reviewed studies

Notably, age predominates as the most frequently cited variable, highlighted in 55% of the studies (e.g., Rahman et al., 2023; Sabouri et al., 2023; Soliman et al., 2023), with no other variable reaching a comparable frequency. In terms of categories, lab results emerged with the highest number of features (n=11), highlighting creatinine (Cr) (including serum creatinine and discharged creatinine), hemoglobin (Hb), and sodium (Na) as the top predictors. In contrast, the TTE/ECG and symptoms categories contain only one variable meeting the minimum frequency requirement: left ventricle ejection fraction and edema, respectively. For the remaining categories, clinical history, medical status, and vital signs each encompass three variables, with Length of stay (LOS) and number of hospitalizations, Charlson Comorbidity Index, diastolic blood pressure, and systolic blood pressure identified as the top predictors, respectively. In contrast, treatments and comorbidities feature only two variables each, with diuretic use and chronic kidney disease, being the most frequently mentioned in each category.

2.2.4. Impacts: Contributions, Future recommendations, Stakeholders, Limitations

To comprehend the impacts of the studies included in the SLR, their results are thoroughly examined. Table 2.11 systematizes the contributions of the studies and underlined stakeholders, while Table 2.12 summarizes their limitations, and suggestions for future research.

Many studies make significant advances in both scientific understanding and practical applications by demonstrating the efficacy of ML algorithms in various HF outcomes (e.g., Polo Friz et al., 2022; Sabouri et al., 2023). They accomplish this by comparing these algorithms with commonly used pre-existing models (e.g.,Sharma et al., 2022), identifying previously overlooked important variables (e.g.,Tohyama et al., 2021), validating existing models (e.g., Austin et al., 2022; Jang et al., 2023), and offering solutions to common challenges in applying ML in medicine, such as interpretability (e.g., Soliman et al., 2023; Xu et al., 2023), imbalanced data, and missing values (e.g., Najafi-Vosough et al., 2021).

The primary stakeholders emphasized in the studies are healthcare providers and HF patients (e.g., Negassa et al., 2021; Rizinde et al., 2023). Furthermore, certain studies delineate patient cohorts, like pre-fail and chronic HF patients, or focus on specific populations, such as Chinese patients. Additionally, some studies emphasize the involvement of healthcare administrators, researchers, government bodies, and insurance companies.

Among the studies reviewed, only two did not acknowledge any limitations (Austin et al., 2022; Xu et al., 2023). The primary challenge highlighted across most studies is the absence of external validation (e.g., PARK et al., 2022; Sarijaloo et al., 2021). Additionally, issues such as missing data (e.g., Sharma et al., 2022), selection bias (e.g., Jang et al., 2023), incomplete

variable inclusion (e.g., laboratory or non-clinical biomarkers) (e.g., Karakuş & Er, 2022), dataset size limitations (e.g., Okoye et al., 2023), class imbalance (e.g., Sabouri et al., 2023) and interpretability concerns (e.g., Polo Friz et al., 2022) are commonly recognized. Future recommendations often echo these limitations, advocating the inclusion of additional variables (e.g., NT-proBNP, Troponin) (e.g., Ju et al., 2021; Negassa et al., 2021) , inclusion of additional endpoints like quality of life and functionality (Sabouri et al., 2023), the exploration of new interventions like the SGLT-2 inhibitors (Soliman et al., 2023), and the thorough validation of their findings (e.g., Tohyama et al., 2021).

Table 2.11: Contributions and highlighted stakeholders of the studies

ID	Contributions	Highlighted Stakeholder
1	Effective ML models for in-hospital mortality and 30-day readmission; Conventional features combined with ML methods;	Healthcare providers; Healthcare administrators
2	Demonstrates the effectiveness of stacking models in improving prediction accuracy, especially in complex data structures; Addresses a gap in understanding multi-feature-based ML models.	Healthcare providers; Government bodies
3	Provides a comparative analysis of six ML models for HF management in Rwanda, identification of key predictors for hospital readmission.	Healthcare providers; HF patients
4	Use of non-linear algorithms in survival analysis.	Healthcare providers; Researchers
5	Shallow models offer comparable performance and better interpretability, deep models, despite their complexity, require surrogate models for explanation	Healthcare providers
6	ML interpretable tools like permutation importance, PDP plots, and SHAP values identify key mortality predictors (e.g., number of hospitalizations, age, GFR, BNP, DBP, SBP, and NYHA)	Healthcare providers
7	Integrates frailty and comorbidity assessments with ML	Healthcare providers; Pre-frail patients,
8	Enhances HF decompensation predictions using XGBoost and LR; improves clinician trust through model explainability; aids in patient triaging based on risk assessments; identifies symptoms as key predictors (e.g., edema, orthopnea)	Healthcare providers
9	Provides visually intuitive nomogram for clinicians to identify high-risk patients for six-month re-admission.	Healthcare providers; Chinese HF Patients
10	Presents a universal mortality prediction model, the MARKER-HF risk score, applicable across diverse medical conditions, including heart failure.	General medical population (with and without HF)
11	Achieves maximum performance of Fine Gaussian SVM and KNN methods for mortality-survival outcomes	Healthcare providers; Researchers in Informatics Science and Healthcare Data Analytics
12	Provides a ML classifier that surpasses the LaCE score in informativeness, includes feature importance and impact through SHAP plots.	Healthcare system administrators
13	Shows superiority of XGBoost, over the conventional LACE index. Utilizes SHAP analysis to elucidate influential features.	Care teams
14	ML model surpasses traditional risk scores, shows the significance of echocardiographic data.	Healthcare providers
15	EHMRG matches ML in discrimination and surpasses in calibration, importance of calibration metrics	Healthcare researchers
16	Identifies 18 crucial risk factors for HF	Healthcare providers; HF patients
17	Impact of clinical variables on improving short-term risk prediction compared to comorbidities, clinical, psychosocial variables, and discharge medications.	Healthcare providers; HF patients
18	Compares six ML algorithms for predicting hospital readmission with RF as the most effective. Addresses class imbalance and missing data	Healthcare providers; HF patients; Healthcare administrators
19	Barthel index in ACD-VC as a key indicator of frailty; Develops the SMART-HF model (https://hfriskcalculator.herokuapp.com/)	Healthcare providers; Non-healthcare providers, Government bodies; Insurance companies
20	electronic frailty index that includes inflammatory and nutritional indices	Healthcare providers
21	Integrates healthcare data from hospitals, GPs, and pharmaceutical claims.	CHF Patients
22	ML outperforms the HOSPITAL score, with social, clinical, and physiological data integration enhancing accuracy and providing valuable insights.	Healthcare administrators
23	Uses ML and feature selection to identify key variables; simplified models with selected variables match full-set performance, leveraging techniques like PCA for better accuracy.	ND
24	Introduces an MLP model to address class imbalance, achieving higher accuracy than the LACE score; underscores the importance of model and metric selection in imbalanced data.	ND

Notes: ACD-VC: Administrative Claim Data - Voting Classifier; BNP: B-type Natriuretic Peptide; DBP: Diastolic Blood Pressure; GFR: Glomerular Filtration Rate; GP: General Practitioner; HF: Heart Failure; KNN: k-Nearest Neighbors LR: Logistic Regression; ML: Machine Learning; MLP: Multilayer Perceptron; ND: Not Defined; NYHA: New York Heart Association Functional Classification; PCA: Principal Component Analysis; PDP: Partial Dependence Plot; RF: Random Forest; SBP: Systolic Blood Pressure; SHAP: SHapley Additive exPlanations; SVM: Support Vector Machine; XGBoost: Extreme Gradient Boosting.

Table 2.12: Future recommendations and limitations of the studies

ID	Future Recommendations	Limitations
1	Refine prediction models for 3-month readmission and 6-month mortality; Investigate additional endpoints; Enhance model applicability with external data and boost performance using DL methods.	Small sample; Gender Imbalance; SC; Exclusion of non-clinical factors; Lack of external data; Limited consideration of additional endpoints
2	Investigate important HF biomarkers; Fine-tune model on specific population data; Expand research by accumulating multiple datasets from different regions	SC; Lack of time series data for entire hospitalization; Exclusion of important HF biomarkers; Geographical constrains; Small sample; RN
3	Create tailored predictive models by addressing data quality and adapting to local healthcare systems	Cultural influences on healthcare practices and data collection
4	Incorporate multimodal data using NN and DL for better SA; Explore automated ML methods to streamline the modelling process	SC; LEV; RN; MV in key predictors; Simple imputation strategy introduce bias; Non-linear models pose interpretability and ethical challenges.
5	Enhance ML model interpretability; Improve feature engineering with clinical factors; Investigate new interventions; Assess cost-efficiency of readmission reduction strategies.	Insufficient data on post-hospital and primary care; Incomplete assessment of drug-readmission correlation; Limited data on patients using SGLT-2 inhibitors
6	ND	ND
7	Validate CFS and BNP levels clustering in multicentre ADHF studies; Perform external validation in larger elderly HF cohorts	SC; Small to medium sample; Lack of consideration for certain variables such as BMI and the etiology of hospitalization;
8	Enhance predictive performance by using models that leverage temporal data, such as recurrent neural networks and attention-based models	Exclusion of baseline variables (e.g., age, diagnosis duration); Discarding customized patient alerts to avoid clinician conflicts; Missing records of home treatments over the last four year
9	ND	RN; SC; Lack of ECG Features; Sensitivity <50%; Unidentified causes of readmission; Limited applicability
10	Validate MARKER-HF in diverse non-HF populations, including Caucasian and Black patients; Expand applicability to broader clinical contexts, beyond CVD.	SC; Only East Asian patients; Exclusion of patients with missing data introducing selection bias; Underestimation of mortality risk in patients with malignancy
11	Use feature engineering to address limitations related to the dataset's size and imbalance. Augment patient information to capture additional CVD risk factors.	Small sample size; Imbalance dataset; Lack of certain patient information
12	Improve ML model accuracy by incorporating diverse data types, such as social factors and clinical notes through natural NLP.	MV in lab data; Requirement of 2 echos; Limited clinical prediction ability; Barriers to data access and sharing; Assumption of hospitalization independence; Inability to measure significant readmission predictors.
13	ND	RN; SC; LEV; Selection bias; Lack of post-discharge data; CCI may underreport chronic conditions.
14	Confirm ML model applicability across diverse populations; Evaluate treatment efficacy based on ML-predicted mortality risk levels.	RN; Institutional-based train and test cohorts may introduce bias due to institutional differences; LEV
15	Compare ML and conventional statistical models for prognosis-based research, focusing on discrimination and calibration.	ND
16	Validate in larger multi-center or prospective cohorts	SC; RN; Applicability uncertainty for patients without baseline echo; Missing data from lab variables
17	Validate findings on broader database for enhanced external validity; Incorporate additional covariates regarding health status and QOL information	SC; Difficulty in determining cause of death; Lack of patient information; Conservative findings due to exclusion of relevant factors and HF-specific mortality
18	ND	Lack of medication and psychosocial information
19	Further validation of SMART-HF using other databases with different populations and settings.	Median age of 80; Missing data bias; BI lacks literature support for assessing frailty.; Omission of vital signs, echocardiography, lab and treatment data; Reliance on Japanese medical standards
20	Conduct prospective studies to validate the developed model by integrating additional inflammatory, nutritional, and frailty assessment tools for clinical application.	LEV; Lack of treatment data during acute and post-admission phases; Lack of HF subtype differentiation; Limited generalizability; Interpretability challenges due to opaque nature of ML
21	Enrich Health insurance claims data with clinical data; Explore causal relationships of important features; Enhance risk stratification and prognosis with advanced ML with clinical data.	Absence of key clinical features; Data history limited to 7 years; Limited interpretability of ML models, especially NN; Possible exclusion of relevant features due to feature reduction methods
22	Integrate disease-specific predictive models into EHR; Assess the impact of the integration on patient care and resource utilization.	RN; SC; Reliance on billing data may miss HF progression; ML methods unable to infer variable-outcome direction; LR modelling may oversimplify complex relationships.
23	ND	Age Limitation (65+); Lack of Clinical Data
24	Investigate the utilization of performance measures that consider the class imbalance problem during parameter tuning.	RN; Age limitation (65+); Lack of clinical data; Lack of HF subtype differentiation; Hyperparameter optimization limited to traditional stepwise range values.

Notes: ADHF: Acute Decompensated Heart Failure; BI: Barthel Index; BMI: Body Mass Index; BNP: B-type Natriuretic Peptide; CCI: Charlson Comorbidity Index; CFS: Clinical Frailty Scale; CVD: Cardiovascular Disease; DL: Deep Learning; ECG: Echocardiogram; EHR: Electronic Health Record; HF: Heart Failure; LEV: Limited External Validity; LR: Logistic Regression; ML: Machine Learning; MV: Missing Values; NN: Neural Network; NLP: Natural Language Processing; QOL: Quality of Life; RN: Retrospective Nature; RF: Random Forest; SA: Survival Analysis; SC: Single Center; SGLT-2: Sodium-glucose cotransporter-2;

2.3. Quality Assessment

To ensure rigor and reliability in the findings and streamline the research process, articles are evaluated on their effectiveness in addressing the research questions. This evaluation involves assessing articles across four dimensions outlined in Table 2.2. Each criterion is translated into a specific question, and a score is assigned accordingly: 0 if the question is not addressed, 0.5 if partially addressed, and 1 if fully addressed. The Table 2.13 displays the quality assessment of the 24 articles, indicating their scores for each evaluation item and dimension.

Although none of the articles reached the maximum score of 15, two studies attained a score of 14.5 for quality (Ju et al., 2021; Soliman et al., 2023). These exemplary studies should be considered as benchmarks for future research endeavors in the domain of HF prediction models. The average score across all articles was 12.44, indicating a relatively high level of quality.

Among the specific research questions, “What are the scope and objectives?”, received the highest evaluation across the studies, with an average score of 22.75 out of a maximum of 24. Notably, within this dimension, criterion Q1.1, which pertains to the clear description and justification of objectives, stood out as the best criterion overall, with only one article falling short of the total score (Jang et al., 2023). In contrast, the dimension "What methodology is used?" not only received the lowest average score of 19 but also housed the worst quality criterion (Q2.2). Across the studies, there is a lack of information regarding the data collection and preparation process, with only 54.2% (n=13) of articles achieving the maximum mark in this critical aspect. Additionally, in this dimension is also possible to see an insufficient explanation and justification for the variable selection (Q2.3).

The third dimension, focused on the study results, closely followed the methodology dimension with an average score of 19.3, resulting in its placement as the second worst. Within the results dimension, there appears to be difficulty in clearly identifying the best-performing model (Q3.3) and justifying the key variables (Q3.4). In the fourth and final dimension, which assesses the impacts of the studies and ranks as the second highest, an average score of 20.5 is observed. Within this dimension lies the second-best quality criterion (Q4.1). This indicates that articles present their contributions in a clear way. Despite its significance, the quality criterion Q4.3, regarding future recommendations,

obtained one of the lowest scores, with 20.8% (n=5) of articles lacking any such recommendations.

To summarize, for comprehensive insights into the scope, objectives, and study impacts, recommended readings include studies achieving the maximum score in these dimensions, as well as those with an overall high score. For the methodology dimension and in general, Tong et al. (2023) is particularly informative, while Ju et al. (2021) and Rahman et al. (2023) stand out for their rigorous result assessments. For researchers focused on HF decompensation prediction, Kerexeta et al. (2023), is an essential read, as it is the only study addressing this topic, which is the main concern of this investigation.

Table 2.13: Assessment of Article Quality

ID	What are the scope and objectives?		What methodology is used?					What are the results?					What are the study impacts?			Total
	Q1.1	Q1.2	Q2.1	Q2.2	Q2.3	Q2.4	Q2.5	Q3.1	Q3.2	Q3.3	Q3.4	Q3.5	Q4.1	Q4.2	Q4.3	-
1	1	1	1	1	1	0.5	1	1	1	0.5	1	0.5	1	1	1	13.5
2	1	0.5	1	1	0.5	1	1	1	1	1	1	1	1	1	1	14
3	1	1	0.5	0.5	0.5	0.5	1	1	1	1	1	1	1	1	1	13
4	1	1	1	1	1	0.5	1	1	1	1	1	0.5	1	1	1	14
5	1	1	1	1	1	1	1	1	1	1	0.5	1	1	1	1	14.5
6	1	1	0.5	1	1	1	1	1	1	1	1	1	1	0	0	12.5
7	1	1	0.5	0.5	1	1	0	0	0	0	1	0.5	1	1	1	9.5
8	1	1	1	1	1	0.5	1	0.5	1	1	1	1	1	0.5	1	13.5
9	1	1	0.5	1	1	1	1	1	1	1	1	1	1	0.5	0	13
10	0.5	1	1	0	1	1	0	0	1	0.5	0	0.5	1	1	1	9.5
11	1	0.5	0.5	0.5	0.5	1	1	1	0.5	0.5	0	0.5	1	1	1	10.5
12	1	1	1	1	0.5	1	1	1	1	1	1	1	0.5	0.5	1	13.5
13	1	1	0.5	1	0.5	0.5	1	1	1	1	0.5	1	0.5	1	0	11.5
14	1	0.5	1	0.5	0.5	1	0.5	0.5	0.5	0.5	1	0.5	1	1	1	11
15	1	1	1	1	0.5	1	1	1	1	0.5	0.5	0.5	1	0	1	12
16	1	1	1	0.5	1	0.5	0	0.5	1	0	1	0.5	1	1	0.5	10.5
17	1	1	1	0.5	1	0.5	1	1	1	1	0.5	0.5	1	1	0.5	12.5
18	1	0.5	1	1	0.5	1	1	1	1	1	1	1	1	1	0	13
19	1	1	1	0.5	0.5	0.5	1	0.5	1	0.5	1	1	1	1	1	12.5
20	1	1	1	0.5	1	1	1	1	1	1	1	1	1	1	1	14.5
21	1	1	1	1	0.5	1	1	1	1	1	0.5	1	1	1	1	14
22	1	1	1	1	1	1	0.5	0.5	1	0.5	0.5	1	1	1	1	13
23	1	1	0.5	0	1	0.5	0.5	1	1	0.5	1	0.5	1	1	0	10.5
24	1	1	0.5	0.5	0.5	1	1	1	1	1	0	1	1	1	1	12.5
Total	23.5	22	20	17.5	18.5	19.5	19.5	19.5	22	18	18	19	23	20.5	18	-

2.4. Summary and Implications of Literature

The SLR aims to advance scientific knowledge by identifying gaps, guiding future research, and shaping discussions on HF management, especially for HF decompensation prediction through ML. Despite ongoing research, only one model for HF decompensation prediction exists, highlighting the need for a robust, clinically interpretable model to enhance real-world applicability and prevent adverse outcomes.

It is worth noting that previous studies have already developed highly predictive models for various HF outcomes, specifically for readmission (Rahman et al., 2023; Rizinde et al., 2023) and for mortality (Ju et al., 2021; Karakuş & Er, 2022). A smaller number of studies focused on constructing models for a CEP, yet they did not attain comparable high metrics. Unfortunately, the only article addressing decompensation reported only the AUC, which was relatively low (0.69), leaving this study without references for other important metrics. Nevertheless, there is room for improvement by incorporating symptoms, as seen in Kerexeta et al. (2023), such as like shortness of breath (SOB) and paroxysmal nocturnal dyspnea (PND), along with non-clinical variables, such as autonomy and medication adherence, as recommended by Sarijaloo et al. (2021), which require more complex data collection. Awan et al. (2019a) and Najafi-Vosough et al. (2021) emphasize the importance of addressing class imbalance in clinical research, while attention to sample size is also widely acknowledged across studies as a critical limitation. Equally important is a careful and well-reported feature selection process, with this SLR showcasing diverse techniques for researchers to consider, as emphasized by (Awan et al. (2019b).

XGBoost demonstrated promising performance in various studies (e.g., Kerexeta et al., 2023; Polo Friz et al., 2022), however, its interpretability is comparatively lower than other ML models. Integrating interpretability techniques is crucial, with some authors already employing them (e.g., Xu et al., 2023). External validation remains a critical gap, as few studies test models across diverse settings, leaving questions about their applicability and generalizability (e.g., Ju et al., 2021).

Future efforts should focus on incorporating diverse data types, including clinical notes, to retrieve and integrate the previously mentioned variables. Transparent documentation of data collection and careful variable selection can enhance reproducibility, reliability, and practical utility, ultimately supporting external validation and advancing predictive model development in clinical practice.

3. Methodology

This chapter outlines the procedures undertaken in the empirical part of the study, guided by the CRISP-DM methodology. The focus of this research is to develop a predictive model for HF decompensation, ensuring transparency, reproducibility, and scientific rigor through the systematic application of CRISP-DM.

CRISP-DM is a widely used methodology in data mining. Its structured and iterative approach is designed to handle the complexities of analytics projects (Chapman et al., 1999), making it suitable for a variety of industries, including healthcare applications (Felix et al., 2021; Martins et al., 2021). The methodology is organized into six key phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. Each phase ensures that the analysis and model-building process remains aligned with the goals of this study.

3.1. Business Understanding

As outlined in the introduction, HF decompensation presents significant challenges in patient management and places increasing demands on healthcare systems (Bozkurt, 2023; Greene et al., 2023; McDonagh et al., 2021). This research builds a case study from a sample of HF patients managed by the HF team at ULSASI, spanning from January 2020 to April 2024. The ULSASI serves the cities of Amadora and Sintra, two of the most populous regions in Portugal, covering 6% of the national population.

Table 3.1 provides details on the evolution of the number of followed patients and their respective appointments, as well as their outcomes over time, including day hospital (DH) admissions, emergency service (ES) visits, hospitalizations, and deaths. Over this period, the HF team, consisting of five cardiologists, managed 639 HF patients, starting with 139 in 2020 and growing by an average of 140 patients per year. The numbers from 2020 and 2021 were affected by COVID-19 pandemic. As the impact of the pandemic lessened, 2022 and 2023 showed a clearer rise in patient numbers and healthcare use. DH admissions grew significantly (32% in 2022 and 82% in 2023), while ES visits increased by 11% in 2022 before declining by 4.7% in 2023, reflecting a greater tendency to manage decompensation cases without hospitalization criteria in the outpatient clinic instead of admitting all cases, regardless of their severity, as well as a more direct contact with the team without resorting to the emergency department. Hospitalizations remained stable,

while deaths increased, reaching the peak of 17 deaths in 2023. By 2024, 27.7% of the consultations and 76.5% of deaths recorded in 2023 have already been achieved in just four months, indicating that 2024 could surpass previous years in these areas.

Table 3.1: Summary of HF Patient Metrics (January 2020 – 30 April 2024)

	1 Jan 2020	2021	2022	2023	30 April 2024	Total
Patients	139	226 (+133)	309 (+153)	391 (+147)	236 (+67)	639
Consultations	325	669	998	1 175	326	3 493
DH Visits	127	217	287	523	214	1 468
ES Visits	564	686	763	727	291	3 031
Hospitalizations	134	237	243	242	92	948
Deaths	6	9	16	17	13	61

Note: The values in parentheses indicate the number of new patients that entered in each respective year.

These trends at ULSASI are representative of what is occurring in similar healthcare settings worldwide. In the United States, over 1 million hospitalizations occur annually, accounting for 6.5 million hospital days (Ambrosy et al., 2014), and in 2020, HF contributed to 415,922 deaths (Bozkurt et al., 2023). In Germany, HF leads to approximately 440,000 hospitalizations each year and is the most common primary inpatient diagnosis (Dörr et al., 2021).

To address these challenges, the primary business objective is to reduce hospitalizations and emergency visits by 20% through early intervention after the first indicators of imminent HF decompensation, a target set based on the recommendation of cardiologists, to improve patient health and reduce costs. With this goal, the research focuses on four key analytical objectives: O1) characterizing and defining HF decompensation which enables the creation of the prediction target; O2) developing a predictive model for HF decompensation; O3) identifying key factors influencing decompensation; O4) assessing patient profiles to identify those more or less prone to these events. To achieve the final two objectives, the model must be interpretable, providing clear insights into the factors driving decompensation risk.

The first objective is considered successfully completed once consensus is reached among consulted cardiologists on the definition of HF decompensation, establishing a clear foundation for the prediction target. To ensure accurate identification of at-risk individuals, success is then defined by achieving a recall of at least 60 percent and an

AUC of 0.65 or higher. As the only available study on decompensation provided only the AUC (Kerexeta et al., 2023), these thresholds are set to ensure adequate detection of at-risk patients while maintaining a balance in performance. Recall measures the model ability to identify patients who decompensated within 30 days after the consultation, while AUC measures the model overall performance in distinguishing between those who decompensated and those who did not. Prioritizing recall ensures that patients at risk of decompensation are identified in time for intervention, even if it means identifying some patients who are not actually at risk. The third objective is complete when the most important variables of the best-performing model are identified, providing the necessary explainability. The fourth and final objective is achieved when patient profiles more and less prone to decompensation are identified with confidence levels above 80% and support of at least 50 cases.

Is important to note that, in the final dataset, each row represents a consultation rather than an individual patient. This approach treats each consultation as a distinct event, capturing the variability in patient condition, symptoms, and background over time (Bayés-Genís et al., 2005; Bettencourt et al., 2004). By structuring the data this way, the number of cases increases from 639 to over 2,500, allowing the model to learn from multiple events per patient and providing a richer dataset for predicting diverse clinical scenarios (Kerexeta et al., 2023; Sharma et al., 2022).The inclusion criteria specify that patients must have a confirmed HF diagnosis and laboratory results available before each consultation, with further details provided in the following chapter. To ensure clinical rigor, two cardiologists provide guidance throughout the investigation, validating the clinical relevance of the data and findings. Supporting this analysis, the investigation utilizes Excel, SPSS Statistics (version 29.0), and SPSS Modeler (version 18.5).

3.1.1. Defining HF Decompensation and Establishing the Target Variable

As discussed in the introduction, the definition of HF decompensation has evolved over time and is not universally agreed upon. Beyond defining decompensation, it is crucial to understand how to characterize and identify it. To fulfill the first objective and, by doing so, create a reliable target variable, two cardiologists are consulted: the first provides initial insights on how decompensation could be assessed through the patient records, and the second expert is engaged to confirm and refine these viewpoints, given the complexity and controversy surrounding the topic.

Both experts agreed with the most recent definition of decompensation as presented by Bozkurt (2023) and Greene et al. (2023). According to this definition, decompensation is characterized by worsening of HF signs, symptoms and/or functional capacity requiring urgent medical attention, which includes not only unplanned hospital visits (in-patient admission, ES or DH visits), but also oral diuretic optimization and HF guideline directed medical therapy adjustment in a scheduled, although urgent, appointment, or outpatient admission for endovenous diuretic or initiation of chronic inotropic therapy support. Hospitalizations for planned complementary exams or procedures are excluded. Diuretic therapy optimization, indicating decompensation/worsening HF, is identified in the consultation records as the initiation or increase in the dosage of furosemide. Outpatient inotropic therapy, indicating decompensation/worsening HF, is identified by the initiation of intermittent pulses of levosimendan. These are outlined procedures in the clinical guidelines to treat decompensated HF episodes (McDonagh et al., 2021).

The final target variable is determined by a conditional formula that confirms a decompensation event if any of these criteria are met within the 30-day follow-up period, thereby classifying the patient as decompensated. The target is defined as binary: 1 if the patient is decompensated within 30 days, and 0 if stable. A 30-day follow-up period is selected to ensure timely detection and intervention, consistent with guidelines for monitoring decompensated HF patients (McDonagh et al., 2021). This timeframe is also the most used for assessing various outcomes in this domain, as observed in the SLR.

Thus, the analytical objectives translate into a classification problem in which the aim is to classify patients as being prone to HF decompensation or not. To address this, supervised learning algorithms are applied to learn from labeled data, identifying patterns and relationships that can generalize to new, unseen cases (Johnson et al., 2018).

3.2. Data Understanding and Preparation

This chapter covers the second and third phases of the CRISP-DM methodology. Data understanding involves exploring, describing, and assessing the quality of the data, while data preparation focuses on selecting, cleaning, and transforming it into a single dataset for modeling. These phases are combined because the iterative nature of CRISP-DM allows revisiting data understanding after preparation, ensuring the analysis is aligned with the study objectives.

The HF team uses several platforms for patient management, with Soarian being the main system for storing patient data and the sole source used for this analysis. Due to the design of the software, which primarily supports individual patient management, collaboration with the hospital IT team is necessary to collect data from multiple patients simultaneously.

For this analysis, three Microsoft Excel files are extracted, each containing information on patients who were followed in cardiology outpatient consultations. The inclusion criteria are defined as patients aged 18 years or older, with a diagnosis of HF according to the "International Classification of Diseases, Tenth Revision" (ICD-10) main diagnosed code defined in the cardiology appointment [4] (World Health Organization, 2019). The data from these files is summarized in Table 3.2, which outlines the different datasets and their corresponding row counts. The Table 3.3 summarizes the different data sources used in this analysis, highlighting the key fields, the type of information provided by each source, and the encountered challenges. Additionally, all tables contain various identifiers such as "is primary," "code," "doctor code," "service," and "visit type," which are not relevant to the analysis and are discarded.

Table 3.2: Summary of Datasets and Corresponding Row Counts

Category	Dataset Description	Rows
Consultations	Annual consultation counts per patient	639
	Patient details per consultation	3,493
	Consultation Records	55,596
Hospital Admissions	Discharge summaries	948
	ES and DH Visits per Patient	609
	ES summaries by visit	6,460
	DH summaries by visit	61
Lab Results	Laboratory results per semester	277,027 - 636,143

[4] Appendix B lists the ICD-10 codes for HF diagnoses, characterizing the patient sample used in this study.

Table 3.3: Key Information and Challenges Across Data Sources

Data Source	Key Fields	Type of information	Challenges
Consultation Records	Process ID, Sex, Birth date, Date, Registration, Evaluation	Patient past history, hospitalizations, prior lab results, symptoms, NYHA classification, treatment updates, follow-up plan	Registration and evaluation are composed by unstructured texts with inconsistent terminology and misspellings
Hospital Admissions	Process ID, Discharge type, Dates (from ES, DH, and discharge records), discharge clinical notes	HF-related admissions, discharge outcomes, hospitalization patterns, death occurrences	Discharge clinical notes are unstructured text; Discharge records include unrelated procedures as hospitalizations (e.g., device implantation)
Laboratory Results	Process ID, Date, Test name, Test value	52 different lab test values, trends over time	Each row represents a single lab result

To ensure data quality, it is crucial to confirm that all patients, identified by the process ID, have both consultation and lab records, as these are essential for modeling. Two discrepancies are identified: 1) a mismatch between the number of rows in the "Patient details per consultation" and "Consultation Records" tables, due to the extraction including all cardiology patients (13,000) instead of just the 639 HF patients; and (2) out of the 639 intended patients, only 584 have associated lab results.

To maximize data quality, two expert cardiologists help identify key information from consultations and compile a dictionary of terminology, acronyms, and abbreviations. This guides the extraction and structuring of key data into organized formats, ensuring accurate data capture for analysis. Despite these efforts, some critical information remains difficult to capture due to variations in its recording. For example, patient weight and furosemide dosage are too inconsistent in format (e.g., “lost 3 kg” vs. total weight) and vague terms (e.g., “1 pill/ 2xday”, “1+1/2”) to be reliably extracted with Excel text manipulation formulas. Additionally, due to the challenges presented by the “discharged clinical notes” field (Table 3.3), hospitalizations are tracked using ES records to verify whether the discharge destination indicates hospitalization, ensuring that hospitalization for routine procedures are excluded. Lastly, duplicate records are removed, terminology is standardized, and the reasons behind missing data are investigated. Verification ensures that the data only concerns the time frame from January 2020 to April 30, 2024, for consistency across the different datasets. These steps are

essential for maintaining dataset integrity, ensuring it remains reliable for accurate analysis and decision-making.

The data preparation process is divided into three categories of information, each cleaned and organized in separate Excel files before being compiled into a consolidated table for analysis. Various types of formulas (e.g., logical combined with text functions) are created to structure data from consultation records and hospital admissions, while Excel Power Query is used for laboratory results (Table 3.8 provides examples of the formulas employed in this process).

From the data understanding phase, 25 final columns are created from consultation records, capturing various aspects of patient information. These include demographic factors (e.g., race), lifestyle factors (e.g., autonomy, smoker status), implantable devices (e.g., implantable cardioverter-defibrillator (ICD) device), comorbidities (e.g., diabetes mellitus (DM), hypertension (HTN)), symptoms (e.g., Shortness of Breath (SOB), paroxysmal nocturnal dyspnea (PND)), and functional status (New York Heart Association (NYHA) classification). Additionally, the medications column tracks changes such as starting, stopping, or adjusting furosemide and levosimendan, while consultation history records the number of consultations in the last 6 months (Consult_6m) and in the last 12 months (Consult_12m) (McDonagh et al., 2021).

The creation of variables from unstructured consultation text presents several challenges, particularly when dealing with inconsistent terminology, acronyms, and abbreviations. For instance, extracting race requires multiple steps to account for misspellings, indirect references, and formatting inconsistencies (e.g., "light-skinned," "Caucasian," or references like "native of"). Symptoms are similarly challenging, as the text often references the absence of symptoms rather than their presence. For example, doctors may note "no dyspnea" or "without orthopnea," and formulas must detect these phrases accurately. When multiple symptoms are listed as absent, it becomes challenging to differentiate them accurately. In these cases, ChatGPT is used to confirm whether the text refers to the presence or absence of symptoms. To ensure reliability, a cardiologist reviews a random sample of 30 cases to confirm the accuracy of the extracted variables, finding that 29 out of 30 cases were correct. Lastly, variables are created to capture variations from previous consultations, such as changes in symptoms and sings, resulting

in the addition of 12 past-related variables [5]. The variables collected in this phase and included in the modeling process are shown in Table 3.4.

Table 3.4: Excerpt from the data dictionary of consultation records variables for modeling

Variable	Description	Type	Units	Distinct Values NR	Descriptive
Sex	Patient gender	N	-	V:2 NR:0	HF: Male (61.3%); LF: Female (38.7%)
Smoker	Smoker status	N	-	V:3 NR:0	HF: Never smoked (62.0%); LF: Smoker (16.5%)
ICD	Implantable Cardioverter-Defibrillator	N	-	V:2 NR:0	HF: No (89.1%); LF: Yes (10.9%)
NYHA_cod	New York Heart Association Functional Classification	O	-	V:7 NR:0	HF: 2 (41.5%); LF: 4 (1%)
Symptoms	Number of symptoms	DQ	Symptoms	V:4 NR:0	Min: 0; Mx: 4; M: 0.24; SD: 0.593
Consult_6m	Number of consultations in the past 6 months.	DQ	Consultations	V:8 NR:0	Min: 0; Max: 7; M: 0.78; SD: 0.982

Notes: DQ: Discrete Qualitative; HF: Highest Frequency; LH: Lower Frequency; M: Mean; Max: Maximum; Min: Minimum; N: Nominal; NR: No replies; O: Ordinal; SD: Standard Deviation; V: Number of distinct values.

From the hospital admissions data, 10 variables are calculated to create a historical record of patient visits and hospitalizations. These calculations focus on key fields like admission dates to capture important metrics, such as the total number of visits over various time periods (e.g., number of ES from January 2020 until the consultation (ES_2020_Now) or in the last six months (ES_6m)), whether the patient has ever been hospitalized (Previously_Hospitalized), the length of stay for the most recent hospitalization (LOS_Last), and the average number of days hospitalized (Avg_Time_Hosp) , among others (Table 3.5)

[5] Examples of these variables are presented in appendix C.

Table 3.5: Data dictionary of hospital admissions variables for modeling

Variable	Description	Type	Unit	#Values NR	Descriptive
ES_2020_Now	Emergency visits from 2020 to consultation day	CQ	ES visits	#2008 NR:0	Min: 0; Mx: 29; M: 3.84; SD: 4.27
DH_2020_Now	Day Hospital visits from 2020 to consultation day	CQ	DH visits	#2008 NR:0	Min: 0; Mx: 104; M: 1.59; SD: 5.87
ES_6m	Emergency visits in the last 6 months	CQ	ES visits	#2008 NR:0	Min: 0; Mx: 12; M: 1.16; SD: 1.66
DH_6m	Day Hospital visits in the last 6 months	CQ	DH visits	#2008 NR:0	Min: 0; Mx: 26; M: 0.46; SD: 1.89
Hosp_2020_Now	Hospitalizations from 2020 to consultation day	CQ	Hospitalizations	#2008 NR:0	Min: 0; Mx: 9; M: 1.13; SD: 1.34
Hosp_6m	Hospitalizations in the last 6 months	CQ	Hospitalizations	#2008 NR:0	Min: 0; Mx: 5; M: 0.41; SD: 0.69
Previsously_Hospitalized_12m	Whether the patient was previously hospitalized in the last 12 months	N	-	#2008 NR:0	HF: No (53.5%); LF: Yes (46.5%)
Previsously_Hospitalized	Whether the patient was ever hospitalized	N	-	#2008 NR:0	HF: Yes (63.9%); LF: No (36.1%)
Avg_Time_Hosp	Average duration of hospitalizations	CQ	Days	#2008 NR:0	Min: 0; Mx: 78; M: 4.50; SD: 8.00
LOS_Last	Most recent hospitalization length of stay	CQ	Days	#2008 NR:0	Min: 0; Mx: 92; M: 4.35; SD: 8.29

Notes: CQ: Continuous Quantitative; HF: Highest Frequency; LH: Lower Frequency; M: Mean; Max: Maximum; Min: Minimum; N: Nominal; NR: No replies; SD: Standard Deviation; # Number of valid replies

The laboratory results are divided into six Excel sheets due to the volume of data, and the columns are not initially organized for analysis. To address this, Power Query in Excel is used to consolidate and restructure the data by patient and date of analysis, expanding it to include 52 different lab results as separate columns. Key adjustments, such as transposing columns, removing duplicates, changing data types, and handling null values (e.g., “sample not collected” or “not performed”), are applied consistently across all tables before merging them into a single dataset.

To build a comprehensive historical record, additional variables are created from the lab results. This includes capturing values from up to four previous analyses for each test and calculating variations between them. Metrics such as maximum, minimum, average, standard deviation, and coefficient of variation are computed for each lab result [6]. Binary variables are also introduced to indicate clinically significant variations, such as

[6] Examples of these variables are presented in appendix D

whether the reduction in NT-proBNP is less than 30% (Bayés-Genís et al., 2005; Bettencourt et al., 2004), and for creatinine, whether the value has doubled or increased by 0.3 mg/dL or more (KDIGO, 2024). Additionally, since glomerular filtration rate (GFR) values were only available from May 2021, they are calculated using the CKD-EPI Creatinine Equation (2021) (Inker et al., 2021) for use in the final analysis. In total, 579 variables are created from the laboratory data, allowing for a detailed historical perspective and the evaluation of clinically relevant trends over time. Table 3.6 provides the data dictionary for laboratory results, focusing only on variables selected later for modeling.

Table 3.6: Excerpt from the data dictionary of laboratory results variables for modeling

Variable	Description	Type	Units	#Values NR	Descriptive
Hb	Hemoglobin	CQ	g/L	#1902 NR:106	Min: 7; Mx: 20.3; M: 13.31; SD: 2.00
RBC	Reed Blood Count	CQ	X 10^12/L	#1902 NR:106	Min: 1.77; Mx: 7.03; M: 4.42; SD: 0.70
Hct	Hematocrit	CQ	%	#1902 NR:106	Min: 21.9; Mx: 61.3; M: 40.94; SD: 5.79
WBC	White Blood Count	CQ	X 10^9 / L	#1902 NR:106	Min: 2; Mx: 30.5; M: 7.59; SD: 2.56
NT-proBNP	N-terminal pro B-type Natriuretic Peptide	CQ	pg/ml	#1266 NR:742	Min: 101; Mx: 72585; M: 3646.09; SD: 6060.91
Na	Sodium	CQ	mmol/L	#1869 NR:139	Min: 124.7; Mx: 151.1; M: 139.68; SD: 3.23
K	Potassium	CQ	mmol/L	#1834 NR:174	Min: 2.68; Mx: 5.99; M: 4.62; SD: 0.52
Urea	Urea	CQ	mg/dL	#1815 NR:193	Min: 11.9; Mx: 293.6; M: 59.36; SD: 35.53
GFR_calc	Glomerular Filtration Rate	CQ	mL/min./1	#1906 NR:102	Min: 6; Mx: 132; M: 64.48; SD: 26.04

Notes: CQ: Continuous Quantitative; HF: Highest Frequency; LH: Lower Frequency; M: Mean; Max: Maximum; Min: Minimum; NR: No replies; SD: Standard Deviation; # Number of valid replies

The target variable is created to identify decompensation events, as previously mentioned, which occur when a patient requires an unplanned medical visit or when therapeutic adjustments are prescribed in a scheduled visit. These events are indicated by hospital admissions, including ES visits, DH visits, or hospitalizations, as well as therapy optimization, such as starting or increasing furosemide or initiating levosimendan cycles. To capture these events, 5 columns are created to track hospital admissions and therapy changes within 30 days after each appointment. The target variable (Table 3.7) is defined using the formula provided in Table 3.8, where a “yes” (indicating any decompensation

event within 30 days) is coded as 1, and a “no” (indicating no decompensation event) is coded as 0.

Table 3.7: Target variable characterization

Variable	Description	Type	Nº Categories	Descriptive
D_30d	Whether the patient decompensated within 30 days	N	#2 NR:0	HF: No (82,9%); LF: Yes (17,1%)

Table 3.8: Examples of formulas used for structuring data and creating variables

Variable	Formula
Smoker Status	=IF(OR(ISNUMBER(SEARCH("ex-fumador";[@[Registo_Consulta]])); ISNUMBER(SEARCH("fumador";[@[Registo_Consulta]])); ISNUMBER(SEARCH("tabagismo";[@[Registo_Consulta]]))); 1; 0)
DM	=IF(OR(ISNUMBER(SEARCH("DM";[@[Registo_Consulta]])); ISNUMBER(SEARCH("Diabetes tipo 2";[@[Registo_Consulta]])); ISNUMBER(SEARCH("Diabetes mellitus";[@[Registo_Consulta]]))); 1; 0)
Furosemide	=IF(OR(ISNUMBER(SEARCH("aumentar furo";[@Avaliacao2])); ISNUMBER(SEARCH("aumenta furo";[@Avaliacao2])); ISNUMBER(SEARCH("iniciar furo";[@Avaliacao2])); ISNUMBER(SEARCH("reiniciar furo";[@Avaliacao2])); ISNUMBER(SEARCH("aumentar lasix";[@Avaliacao2])); ISNUMBER(SEARCH("iniciar lasix";[@Avaliacao2])); ISNUMBER(SEARCH("reiniciar lasix";[@Avaliacao2]))); 1; IF(OR(ISNUMBER(SEARCH("reduz furo";[@Avaliacao2])); ISNUMBER(SEARCH("mantém furo";[@Avaliacao2])); ISNUMBER(SEARCH("suspende furo";[@Avaliacao2])); ISNUMBER(SEARCH("reduz lasix";[@Avaliacao2])); ISNUMBER(SEARCH("mantém lasix";[@Avaliacao2])); ISNUMBER(SEARCH("suspende lasix";[@Avaliacao2]))); 2; 0))
ES_2020_Now	=COUNT(FILTER(Hospital_Admissions.xlsx!ES[Data_admissao]; (Hospital_Admissions.xlsx!ES[N_processo]=[@[Process_ID]])*(Hospital_Admissions.xlsx!ES[Data_admissao]<[@Date]); ""))
Hosp_2020_Now	=COUNT(FILTER(Hospital_Admissions.xlsx!ES[Date_admission]; (Hospital_Admissions.xlsx!ES[processo_ID]=[@[Process_ID]])*(Hospital_Admissions.xlsx!ES[Date_admission]<=[@Date])*(Hospital_Admissions.xlsx!ES[Type_discharge]="Serviço de Internamento"); ""))
D_30d	=IF(OR([@[Es_Hosp_30d]]>0;[@[ES_30d]]>0;[@[DH_30d]]>0;[@[TO_Furo_30d]]>0;[@[TO_Levo_30d]]>0);1;0)

The data preparation process integrates all relevant variables into a single dataset, resulting in a data model with 626 columns and 2,008 rows. Given this extensive number of variables, a feature selection process is necessary to reduce complexity and maintain a manageable model.

3.2.1. Feature Selection

To improve model performance and focus on the most relevant variables, a feature selection process is applied. Given the large number of variables, they are divided into three dimensions: Patient Profile, Clinical Presentation, and Laboratory Results to clarify

the role of each variable in the analysis. This process involves exploratory data analysis, using bivariate techniques, including correlation analysis and LR, and multivariate model-based techniques.

A bivariate analysis is performed using LR to get familiar with the primary variables and understand their impact on HF decompensation (Rahman et al., 2023). LR models the probability of an outcome based on predictor variables. It uses the logistic function to constrain probabilities between 0 and 1 and is valued for its interpretability and ability to estimate the effect of each variable on the outcome, making it especially useful for binary classification tasks (Cox, 1958). Variables are considered significant at the 0.1 level, as this facilitates detecting additional relationships. Pearson correlation and the Variance Inflation Factor (VIF) are used to check for multicollinearity [7]. Table 3.9 shows the significant variables from the LR.

Table 3.9: Bivariate LR results for target (D_30d): significant variables across patient profile, clinical presentation, and lab results

	Patient Profile		Clinical Presentation		Lab Results	
Independent Variables	Sig	Exp(B)	Sig	Exp(B)	Sig	Exp(B)
Sex: Female (Male=0)	0.017	1.332	-	-	-	-
Autonomy: Independent (Dependent=0)	0.038	0.635	-	-	-	-
Smoker: Smoker (Never smoked=0)	0.052	0.714	-	-	-	-
Smoker: Ex-Smoker (Never smoked=0)	0.046	0.735	-	-	-	-
ICD: Yes (No=0)	0.047	1.416	-	-	-	-
CRT-P: Yes (No=0)	0.064	0.38	-	-	-	-
ES_2020_Now	<0.001	1.073	-	-	-	-
ES_6m	<0.001	1.301	-	-	-	-
DH_6m	<0.001	1.343	-	-	-	-
DH_2020_Now	<0.001	1.073	-	-	-	-
Hosp_2020_Now	0.025	1.096	-	-	-	-
Hosp_6m	<0.001	1.55	-	-	-	-
Previsously_Hospitalized_12m: Yes (No=0)	0.055	1.255	-	-	-	-
Avg_Time_Hosp	0.092	1.011	-	-	-	-
LOS_Last	0.073	1.012	-	-	-	-
Consult_6m	0.03	1.13	-	-	-	-
JVD: Yes (No=0)	-	-	0.012	2.738	-	-
PND: Yes (No=0)	-	-	0.032	1.951	-	-
Orthopnea: Yes (No=0)	-	-	0.023	1.621	-	-
Symptoms	-	-	0.006	1.279	-	-
NYHA_Cod: 3 (1=0)	-	-	0.019	1.752	-	-
NYHA_Cod: 4 (1=0)	-	-	0.003	5.304	-	-
Hb	-	-	-	-	<0.001	0.815
RBC	-	-	-	-	<0.001	0.638
Hct	-	-	-	-	<0.001	0.936
WBC	-	-	-	-	0.049	0.952
N-TproBNP	-	-	-	-	<0.001	1
Na	-	-	-	-	0.002	0.943
Var3_30_NTproBNP: Reduction < 30% (Reduction >=30% =0)	-	-	-	-	0.004	1.651
Var1_0.3_Cr: Increase >= 0.3 (Increase <0.3 =0)	-	-	-	-	0.079	1.463

[7] See appendix E

In addition, Principal Component Analysis (PCA) is performed to explore the possibility of reducing the dimensionality of the dataset (Wold et al., 1987). However, the reduction in dimensionality is not substantial enough to justify the loss in interpretability. Therefore, PCA is not applied further [8].

The Figure 3.1 shows the process of feature selection for multivariate modeling. First, models are tested to determine whether to use all variables or selected ones from each dimension. Laboratory Results are guided by the input of the cardiologists and SLR findings, while Patient Profile and Clinical Presentation use algorithmic feature selection methods. Once the variables are chosen, historical data, such as previous symptoms and descriptive statistics from laboratory results (e.g., max_Hb, SD_Htc), are tested to see if their addition improves model performance. If not, the previously chosen variables are retained. While recall and AUC are prioritized, simplicity is also important, and fewer variables are chosen when performance is similar to avoid adding unnecessary complexity that could affect the model interpretability.

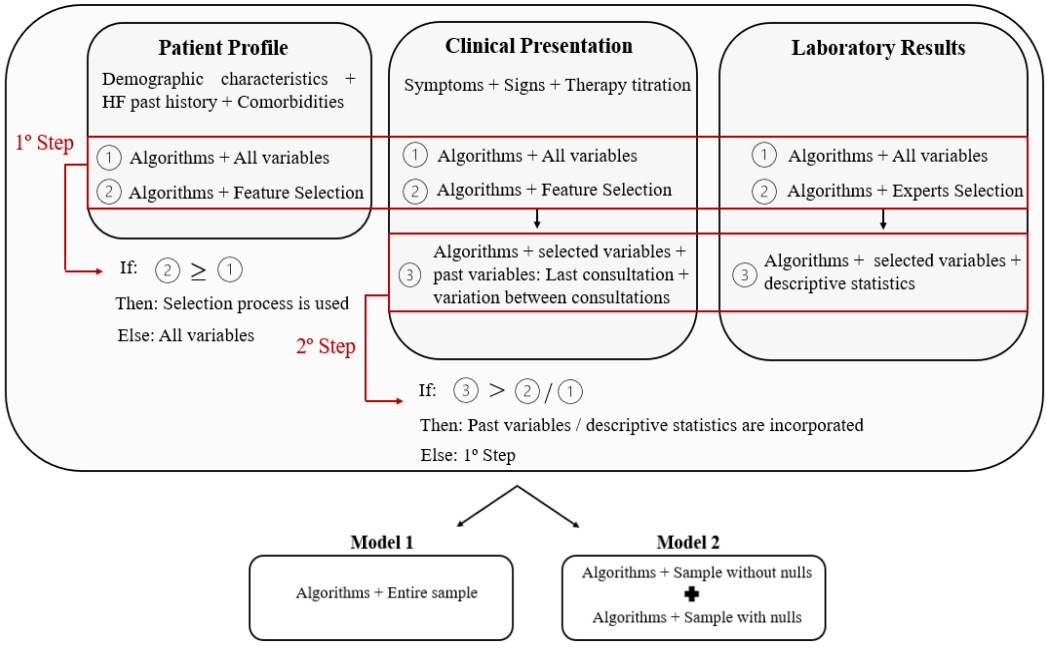


Figure 3.1: Feature Selection Process for Predicting HF Decompensation

ML algorithms, such as DT-based models, XGBoost, and NN, are used for feature selection. These models are capable of handling correlated variables and help uncover new insights, selecting the most relevant predictors for HF decompensation (Chowdhury

[8] The outputs of the PCA with the best results are provided in appendix F.

et al., 2021; De Veaux & Ungar, 1994). The descriptions of each algorithm are presented in Table 3.10.

Table 3.10: Description of algorithms applied in the study (feature selection, modeling, and interpretability)

Algorithm	Description
XGBoost	A gradient boosting algorithm that builds decision trees iteratively, optimizing the model by correcting errors of previous trees. It is highly efficient due to the use of second-order gradients and regularization, making it powerful for classification tasks (Chen & Guestrin, 2016).
CHAID	A DT algorithm that uses chi-square tests to determine the best splits in the data. It identifies the most statistically significant associations between the target variable and predictors by iteratively partitioning the data into subgroups. CHAID does not require pruning, as it stops splitting when no significant association can be found (Kass, 1980).
RT	An ensemble learning method that builds multiple decision trees using random subsets of data and features. It aggregates the predictions of individual trees to improve accuracy and reduce overfitting, making it robust for classification tasks (Breiman, 2001).
C5.0	An improved version of the C4.5 decision tree, designed for classification. It builds the tree by splitting data based on the most informative features. It is robust to missing values and large input features and supports boosting, pruning, and handling noisy data. In addition, the model rules are straightforward and easy to interpret (Quinlan, 1993; IBM, n.d.).
NN	Modeled after the human brain, consisting of layers of neurons that learn complex patterns. Each neuron processes inputs and passes the result to the next layer. Through backpropagation, the network adjusts weights to minimize errors, making it useful for classification tasks (Bishop, 2006).

To improve the performance of the selected models, several parametrization strategies are applied. Bagging and boosting, both ensemble techniques, are used to enhance predictive accuracy. Bagging improves model stability and generalization by creating multiple subsets of the data through bootstrapping (random sampling with replacement) and averaging predictions to reduce variance and prevent overfitting (Breiman, 1996). Boosting, in contrast, builds models sequentially, with each new model focusing on correcting the errors of the previous one. By assigning more weight to misclassified instances, boosting reduces bias and increases overall accuracy (Ganaie et al., 2022).

Additionally, parameters like maximum depth and number of cases in parent and child nodes are adjusted to optimize performance. Maximum depth controls the length of the decision tree, with deeper trees making the model more complex but potentially overfitting the data. The number of parent and child nodes controls the growth of the tree, which consequently impacts the complexity and overfitting of the model. In XGBoost, parameters like scale pos weight help address class imbalance, giving more weight to the

minority class (IBM, n.d.). For NN algorithms, the number of layers is adjusted, where increasing layers allows the network to capture more complex patterns (Austin et al., 2022).

In this context, misclassification costs are specifically applied to improve recall, assigning higher costs to false negatives. This ensures that the model prioritizes minimizing false negatives, which are critical in health-related tasks. Since the number of patients who suffer decompensation (class 1) is lower and the sample size is relatively small, a balancing technique using boost (Rizinde et al., 2023; IBM, n.d.) is applied to ensure the model does not favor the majority class, improving generalization and accuracy for minority class predictions.

The Auto Classifier (IBM, n.d.), which has built-in cross-validation, is used to automatically test various models and compare them. It allows for adjustments to model parameters; however, for quicker performance, the parameters are left in their default mode. The output displays model details, including different parameterizations, chosen inputs, and predictor importance, ranking models based on performance and saving the best ones for further analysis or scoring. Detailed performance specifications for the best models of each algorithm across all dimensions are presented in Table 3.11 and Table 3.12.

Table 3.11: Feature selection: parameterization of the best models for each algorithm by dimension (Part 1)

	PP	LR	CP	PP	LR	CP	PP	LR	CP
Algorithm	XGBoost			CHAID			RT		
Ensembles	-	-	-	Bagging	Bagging	Bagging	-	-	-
Maximum tree depth	18	18	18	5	12	3	10	10	5
Records in parent branch	-	-	-	2	30	2	-	-	-
Records in child branch	-	-	-	1	15	1	10	10	1
Misclassification costs	-	-	-	Yes	Yes	No	Yes	Yes	No

Table 3.12: Feature selection: parameterization of the best models for each algorithm by Dimension (Part 2)

	PP	LR	CP	PP	LR	CP
Algorithm	C5.0			NN		
Ensembles	Boosting	-	-	-	Bagging	Boosting
Maximum tree depth	-	-	-	-	-	-
Records in parent branch	-	-	-	-	-	-
Records in child branch	1	-	-	-	-	-
Misclassification costs	Yes	Yes	No	-	-	-

Recalling the selection process, 1 includes all variables, 2 applies feature selection, and 3 adds past variables for Clinical Presentation and Lab Results. Comparing 1 and 2 for the Patient Profile, 2 shows better overall performance, with even the worst model in 2 performing better than the worst in 1. Although XGBoost drops in performance, models CHAID (AUC 0.610 to 0.653, recall 53.78% to 58.82%), RT (AUC 0.611 to 0.654, recall 50.42% to 56.30%), and C5.0 (AUC 0.611 to 0.628, recall 50.42% to 62.18%) show clear improvements. In the Clinical Presentation dimension, both the worst and best models in 2 show modest improvements over 1, while 3 results in worse performance. For Lab Results, the best model in 2 performs similarly to 1, with a slightly higher recall, making 2 the preferred choice for its simplicity. In 3, recall falls below 50%, confirming that adding extra variables does not improve performance.

Table 3.13: Feature selection: results of best and worst algorithms (Part 1)

Dimensions		Patient Profile				Clinical Presentation					
Selection		1		2		1		2		3	
Metrics		Worst RT	Best XGBoost	Worst C5.0	Best XGBoost	Worst RT	Best Auto	Worst CHAID	Best Auto	Worst RT	Best Auto
Training	Recall	85,78%	100%	91,67%	99,54%	41,33%	69,98%	42,71%	74,07%	55,25%	54,10%
	AUC	0,893	0,999	0,901	0,991	0,568	0,570	0,579	0,570	0,645	0,658
Test	Recall	50,42%	100%	62,18%	95,80%	36,13%	73,95%	38,66%	73,95%	34,45%	35,29%
	AUC	0,611	0,995	0,628	0,946	0,523	0,542	0,555	0,558	0,492	0,548

Table 3.14: Feature selection: results of best and worst algorithms (Part 2)

Dimensions		Lab Results					
Selection		1		2		3	
Metrics		Worst NN	Best XGBoost	Worst C5.0	Best XGBoost	Worst RT	Best XGBoost
Training	Recall	74,82%	100%	90,88%	100%	95,16%	100%
	AUC	0,768	0,997	0,855	0,998	0,945	0,987
Test	Recall	53,78%	100%	54,62%	100%	48,74%	100%
	AUC	0,600	0,973	0,573	0,973	0,586	0,923

In total, 25 variables were selected, with 13 from the Patient Profile, 10 from Lab Results, and only 2 from the Clinical Presentation. After the selection process, the most important variables from each dimension are identified, with the top 5 presented in Table 3.15 based on the feature importance of the best-performing algorithms. Interestingly, in the Patient Profile, the top 5 variables are primarily related to hospital interactions and the presence of an ICD. For the Clinical Presentation, only two variables were selected,

NYHA classification and number of symptoms, making them the most important for this dimension. As for Lab Results, key biomarkers like hemoglobin (Hb), hematocrit (Hct), and NT-proBNP, which are often associated with heart function and overall health status, stood out as the most important variables.

Table 3.15: Top 5 most important variables by dimension

Dimension	Model	Most Important Variables (Top 5)				
Patient Profile	C5.0	Consult_6m	ICD	ES_6m	DH_2020_Now	DH_6m
Clinical Profile	Auto (RT)	NYHA_cod	Symptoms	-	-	-
Lab results	CHAID	Hb	Hct	RBC	NT-proBNP	GFR_calc

3.3. Modelling

In this phase, various modelling techniques are selected and applied with consideration of both business and analytical objectives.

The same five algorithms presented earlier are once again chosen for their balance between interpretability and strong performance in health-related tasks (e.g., (Rahman et al., 2023; Sharma et al., 2022a). DT-based algorithms, such as Chi-squared Automatic Interaction Detection (CHAID) (Kass, 1980), C5.0 (Quinlan, 1993), and Random Trees (RT) (Breiman, 2001), offer high interpretability, while models like XGBoost (Chen & Guestrin, 2016) and NN (MLP) (Bishop, 2006).are included for their strong predictive capabilities based on findings from previous studies (Awan et al., 2019a; Awan et al., 2019b; Beecy et al., 2020). The Auto Classifier is also used, as it tests the same algorithms but with different built-in parameterizations.

The DT-based algorithms are crucial in achieving analytical objectives due to their interpretability. They not only help identify the most important variables and patient profiles but also aid in understanding how more complex models like XGBoost, NN, and the Auto Classifier operate. By using DT to explain the results of these less interpretable models, insights can be gained without sacrificing accuracy, allowing a balance between predictive power and model clarity (Ahmad et al., 2018).

First, various models are created using the entire sample with the 25 selected variables, applying different parameterizations for each algorithm. To maximize performance, an additional model is created by splitting the sample into two groups: one with cases that have no missing values and one with cases containing missing values. All models are tested on both groups, and the two strongest models from each group are selected. These models are then combined to create the comprehensive final model. For

example, XGBoost performs well but excludes cases with missing values, which makes splitting the sample necessary. This procedure not only enhances model performance but also provides insights into how the models behave with complete cases and those containing missing values, which reflect a more realistic scenario of the data available at the moment. By analyzing both groups separately, it helps to determine whether the variables with missing values are essential for future improvements, allowing for better decision-making on whether retrieving these variables should be prioritized. The best parametrizations for each algorithm across the three sample groups (complete, with nulls, and without nulls) are detailed in Table 3.16 and Table 3.17.

Table 3.16: Parameterizations of the best models for each algorithm by sample type (complete, with nulls, without nulls) (Part 1)

Sample	C	W/o Nulls	C	W/Nulls	W/o Nulls	C	W/Nulls	W/o Nulls
Algorithm	XGBoost		CHAID			RT		
Ensembles	-	-	-	Bagging	Bagging	-	-	-
Maximum tree depth	12	12	5	4	5	5	4	6
Records in parent branch	-	-	2	2	2	-	-	-
Records in child branch	-	-	1	1	1	10	10	10
Misclassification costs	-	-	Yes	Yes	Yes	No	Yes	No

Notes: C: Complete; W/Nulls: With nulls; W/o Nulls: Without nulls

Table 3.17: Parameterizations of the best models for each algorithm by sample type (complete, with nulls, without nulls) (Part 2)

Sample	C	W/ Nulls	W/o Nulls	C	W/ Nulls	W/o Nulls
Algorithm	C5.0			NN		
Ensembles	-	-	-	-	-	Bagging
Maximum tree depth	-	-	-	-	-	-
Records in parent branch	-	-	-	-	-	-
Records in child branch	12	-	10	-	-	-
Misclassification costs	Yes	Yes	Yes	-	-	-

Notes: C: Complete; W/Nulls: With nulls; W/o Nulls: Without nulls

3.4. Evaluation

In the evaluation phase, the models are assessed to ensure they meet the objectives. This includes comparing evaluation metrics, selecting the best model, and subsequently analyzing prediction errors. Techniques such as sensitivity analysis and logistic regression are applied to ensure the interpretability of the chosen model.

The holdout method with a 70-30 split is used to validate the models, meaning 70% of the data is used for training, while the remaining 30% is reserved as unseen data

for testing. For the Auto Classifier and, sometimes, for C5.0, the holdout is combined with 5-fold cross-validation, where the data is split into 5 subsets, and the model is trained and tested 5 times, each time using a different subset as the test set while the others are used for training. This combination, frequently seen in the SLR (e.g., Özbay Karakuş & Er, 2022; Tong et al., 2023), ensures robust validation. However, not all models allow this level of parameterization.

To evaluate the quality of the models, several standard metrics (Rahman et al., 2023; Rizinde et al., 2023) are selected, based on the confusion matrix (Table 3.18). These include specificity, recall, accuracy, F1 score, and precision, with their formulas and descriptions presented in Table 3.19. In this context, the positive class refers to a patient decompensating in the 30 days following the consultation.

Table 3.18: Confusion matrix

	Predicted Class	
Actual Class	Stable	Decompensation Event
Stable	TN (True Negative)	FP (False positive)
Decompensation Event	FN (False Negative)	TP (True positive)

Table 3.19: Metrics for evaluation

Metric	**Formula**	**Explanation**
Specificity	$\frac{TN}{TN + FP}$	True Negative Rate, measures the proportion of actual negatives that are correctly identified.
Recall	$\frac{TP}{TP + FN}$	True Positive Rate, measures the proportion of actual positives that are correctly identified.
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Overall correctness, measures the proportion of correct predictions (both true positives and true negatives).
F1 score	$\frac{2 * Precision * Recall}{Precision + Recall}$	Harmonic mean of precision and recall, providing a balanced measure of model performance.
Precision	$\frac{TP}{TP + FP}$	Positive Predictive Value, measures the proportion of positive results that are true positives.

In addition, AUC is used, derived from the Receiver Operating Characteristic (ROC) curve. The ROC curve represents the trade-off between the true positive rate (recall) and the false positive rate (1 - specificity) across different classification thresholds. AUC measures the overall ability of the model to distinguish between the positive and negative classes. An AUC value ranges from 0 to 1, where 1 indicates perfect classification, and 0.5 suggests the model performs no better than random guessing (Davis & Goadrich, 2006). This metric was the most reported in the SLR, showing its importance

in model evaluation. The evaluation metrics are applied to both the training and test sets to detect overfitting or underfitting. Overfitting occurs when the model performs well on training data but poorly on test data, while underfitting means the model performs poorly on both (Ying, 2019). Based on these results, parameterizations can be adjusted to improve overall model performance.

After identifying the best model, which meets the success criteria mentioned earlier in the business understanding phase, with a strong emphasis on recall and AUC, the cases where predictions were incorrect are reviewed to understand the causes of misclassification and identify potential areas for model improvement. Following this, two additional objectives are pursued: identifying the most important variables in the model and determining the profile of patients more prone to decompensation, with a confidence level of 80% and a support of at least 50 cases. This analysis provides valuable insights for healthcare professionals, helping them manage patients more effectively and prioritize interventions where needed.

3.4.1. Interpretability

To ensure the model is suitable for real clinical application, interpretability is essential (Ahmad et al., 2018). Three steps are taken to ensure this: creating a decision tree for a global explanation of a less interpretable model, conducting a sensitivity analysis of the model, and using LR to assess the impact of variables on the outcome.

Firstly, a DT is applied to approximate the behavior of the less interpretable model. This tree replicates the decision-making process of the model while maintaining a simpler and more interpretable structure. By ensuring comparable metrics, the decision tree faithfully mirrors the predictions of the original model in this classification task, providing a global explanation that balances accuracy with interpretability (Ahmad et al., 2018).

Additionally, a sensitivity analysis is conducted using heatmaps based on the score (probability of positive) of the model to evaluate how changes in key variables influence the predicted scores. The score ranges from 0 to 1, where a higher score indicates greater confidence in a positive classification. The use of the score, rather than simply observing binary predictions, allows for a more comprehensive understanding of the behavior of the model. This approach offers deeper insight into the robustness of the model by visually illustrating the influence of predictors (Kerexeta et al., 2023).

To gain a deeper understanding of the impact of the predictors on HF decompensation, a LR analysis is conducted (Kerexeta et al., 2023), generating coefficients and odds ratios to quantify the influence of each variable on the target and assess their statistical significance. The LR analysis builds on insights from the feature selection process. Correlated variables are reduced by keeping those with greater statistical and clinical significance (e.g., Hb was kept over Reed Blood Count (RBC) and Hematocrit (Hct)). Significant variables are analyzed through forward stepwise and enter methods (Laureano, 2020), following the same structured approach as the feature selection process by dividing this analysis by dimension. Two final regressions are performed: one including all variables that remain significant after the forward stepwise process, and another using the same significant variables but excluding past-dependent ones (e.g., ES_6M, Var1_0.3_Cr) for applicability to new patients. Control variables, such as Age, DM, and GFR, are included regardless of significance, although they are consistently removed stepwise [9].

3.5. Deployment

The deployment phase of the CRISP-DM methodology is characterized by the application of insights gained from this investigation for academic and professional use. The development of this investigation itself represents a form of deployment, laying the groundwork for future applications.

In this context, deployment is reflected in the elaboration of an article based on the SLR of this investigation, which was presented at a healthcare informatics track at a scientific conference, and in this document, which presents the entire process and interpretations of the study. Both documents, the article and this thesis, are shared with healthcare professionals at the HF department, highlighting the study results. Additionally, the Excel formulas and Power Query scripts used to structure the data are provided to the doctors, enabling them to apply these methods in their practice. However, the implementation of the model in the decision-making process of the HF department was not feasible within the one-year development of this investigation. Additional time is required to ensure effective integration and to assess whether the business objectives outlined in the study are met. The model is available for the department to consider its implementation, with the potential to enhance the management of HF patients.

[9] Full results from all regressions are presented in appendix G.

4. Results and Discussion

This chapter presents the results obtained through the applied methodology, providing interpretations aligned with the study objectives and discussing their broader significance and implications. The sample is composed of 17.13% positive cases (n=344), where patients experience HF decompensation within 30 days of their HF consultation, and 82.87% negative cases (n=1664), representing stable patients who do not decompensate in that timeframe.

4.1. Predictive Models for HF Decompensation

To address the first objective, multiple predictive models are developed following the outlined methodology. Given the nature of the models, different approaches are applied to ensure robust results and enable a thorough evaluation. This section presents the results for both the complete dataset and the combined model, integrating predictions from cases with and without null values. Several experiments are conducted for each of the five algorithms, adjusting parameters and samples to optimize performance and avoid underfitting or overfitting. The best models are identified based on the success criteria defined for the first analytical objective. The parameter settings for each model were detailed previously. Results for the models using divided samples, which are not part of the combined model, are not presented here [10].

4.1.1. Evaluation of the Models and Selection of the Best-Performing One

The results of the best models, by algorithm, for the complete sample, as well as the two best for each divided sample (with nulls and without nulls) and the combined model, are shown in Table 4.1. The best model is selected based on a recall threshold of at least 60% and an AUC of 0.65 or higher in the test sample. For the complete sample, ignoring XGBoost due to its exclusion of null cases, the focus is on CHAID, RT, C5.0, and NN. The overall behavior of the models indicates some overfitting, as the training metrics outperform the test set substantially. Notably, C5.0 exhibits the best recall (85%) and AUC (0.903) on the training set, but RT stands out with the highest recall on the test set (58.82%) and an AUC of 0.666. None of these models fully meet both success criteria, although RT comes closest. In terms of other metrics, the models generally show higher

[10] See appendix H

specificity, suggesting better performance in correctly identifying stable patients, but overall precision remains low, indicating some difficulty in accurately predicting decompensation cases.

Due to these limitations, a combination of the best models from the divided samples is explored. XGBoost (for the sample without nulls) and RT from Auto Classifier (for the sample with nulls) individually provide the best performance metrics in their respective samples. This combined model achieves high metrics, such as 96.94% specificity, 81.40% recall, and an AUC of 0.964, outperforming all other models. While this approach limits interpretability due to the nature of both XGBoost and Auto Classifier, the significantly improved performance justifies the choice of prioritizing accuracy over transparency in this case.

Table 4.1: Results of the Best Predictive Models for HF Decompensation per Sample

Metrics		Complete					With Nulls	Without Nulls	Combined
		XGBoost	CHAID	RT	C5.0	NN	RT*	XGBoost	XGBoost + RT*
Training	S	97.27%	76.68%	73.14%	80.14%	76.25%	97,13%	96.95%	96.47%
	R	100.00%	75.46%	64.29%	85.82%	69.07%	84,07%	100.00%	91.12%
	A	98.60%	76.09%	68.84%	82.89%	72.78%	90,22%	98.49%	93.81%
	F1	98.59%	75.42%	66.73%	82.92%	71.04%	90,09%	98.52%	93.61%
	P	97.21%	75.39%	70.32%	80.21%	73.12%	0,9704	97.08%	96.23%
	AUC	1.000	0.856	0.744	0.903	0.813	0,955	1.000	0.977
Test	S	97.51%	70.75%	69.37%	73.91%	69.96%	97,17%	93.71%	96.94%
	R	100.00%	54.62%	58.82%	57.14%	55.46%	60,00%	100.00%	81.40%
	A	97.88%	67.68%	67.36%	70.72%	67.20%	91,27%	94.63%	94.32%
	F1	93.33%	39.16%	40.70%	42.63%	39.17%	68,57%	84.51%	82.84%
	P	87.50%	30.52%	31.11%	34.00%	30.28%	82,61%	73.17%	84.34%
	AUC	0.998	0.679	0.666	0.691	0.695	0,907	1.000	0.964

Supporting the table of results are the gain charts for the best models of the complete sample and the combined model, shown in Figure 4.1 and Figure 4.2 , respectively. In terms of gain, the combined model achieves a gain of 87.2% for the top 40% of patients, compared to C5.0 with 71.4% and RT with 53.6%. This means that when selecting the top 40% of patients most prone to decompensate, the combined model identifies 87.2% of the positive cases, improving over the other models [11].

[11] . Additionally, the ROC curves, which visualize the previously discussed AUC values, are available in the Appendix I.

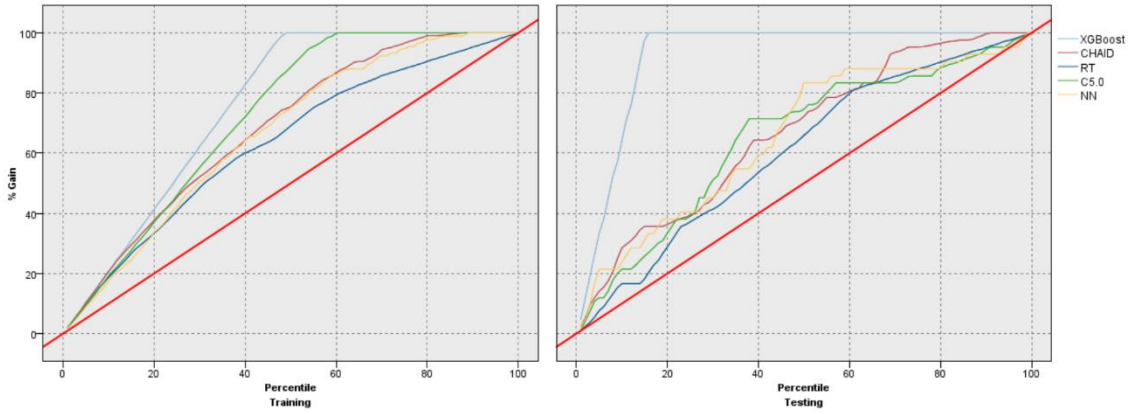


Figure 4.1: Gain charts of the best models by algorithm (Complete Sample)

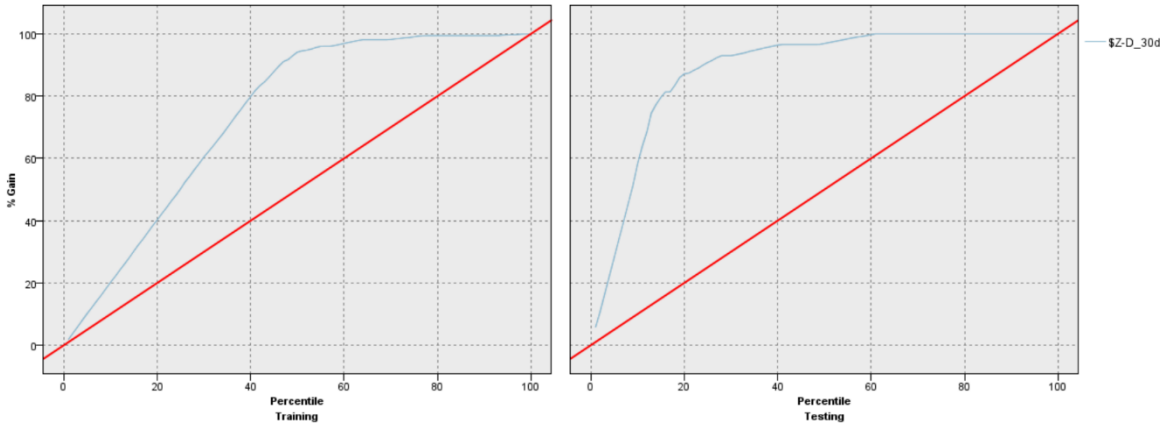


Figure 4.2: Gain charts of the combined model

The error analysis emphasizes how the model performance varies between categories within each predictor. Focusing on these differences helps identify specific groups where misclassifications occur more frequently, highlighting patterns for model refinement.

For the variable Symptoms the model shows an error rate of 6.6% when there are no symptoms (0), 5.5% for 1 symptom, and only 1.2% for 3 symptoms, making no errors when there are 2 or 4 symptoms (Figure 4.3).

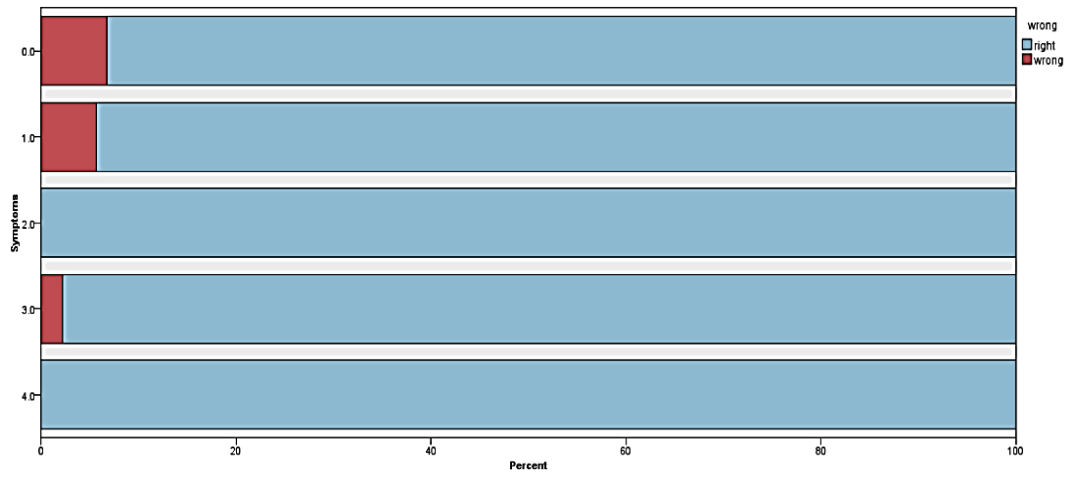


Figure 4.3: Error Distribution by Number of Symptoms

In contrast, when examining a continuous variable, the Cr distribution (Figure 4.4) reveals more outliers for correct predictions, indicating that cases with higher Cr values are predicted more accurately. The model struggles more with lower Cr values, though the median remains similar between correct and incorrect predictions.

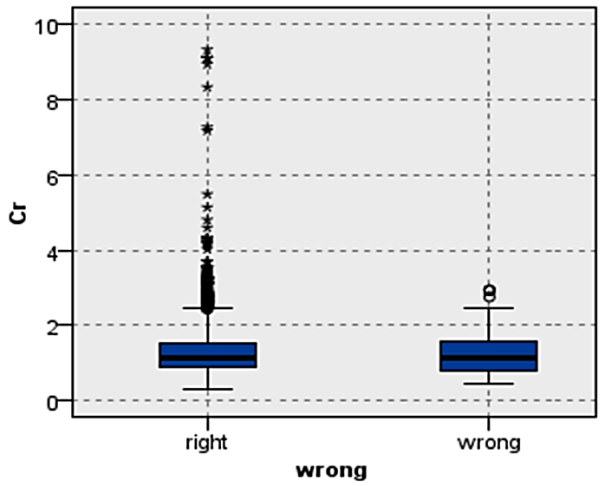


Figure 4.4: Cr-distribution by type of decision

As for the remaining variables, the analysis indicates consistent patterns across different categories.[12] In the clinical presentation dimension, the NYHA classification displays a similar pattern to symptoms, with higher misclassification rates when NYHA is not documented (8% error rate) and for Class 1 (7.2%), suggesting that undocumented or lower classifications are more prone to error. Laboratory variables such as NT-proBNP and WBC follow a similar trend to Cr, where higher values generally improve accuracy, with WBC displaying the fewest outliers among them. However, Na behaves differently, showing better accuracy with lower values. Other lab results exhibit no significant differences between correct and incorrect predictions. All continuous hospital admission variables show a similar pattern, where higher values lead to more accurate predictions, while lower values are associated with increased difficulty. This trend is most evident with DH_2020_Now and DH_6m, where the distribution of errors is centered around lower values of DH visits. Lastly, binary variables, including sex, the presence of an ICD, and previous hospitalizations within the last year, do not show significant performance differences, as their distributions reflect no major impact on prediction accuracy. However, the smoker status variable reveals a trend where the model achieves more accurate predictions in cases where the patient is a current smoker.

[12] Appendix J provides distribution graphs for the majority of these variables, illustrating differences in model accuracy by error distribution across categories.

4.2. Most Important Predictors

To address the third analytical objective of identifying the most important variables that predict HF decompensation, DT-based algorithms are created to mimic the behavior of the combined model, effectively opening the "black box". The C5.0 algorithm, selected for its interpretability, achieves the highest metrics: 91.03% specificity, 93.73% recall, 92.12% accuracy, 90.62% F1 score, 87.71% precision, and an AUC of 0.943. These strong results allow for a confident understanding of the predictions of the combined model. Therefore, the predictor importance of the C5.0 algorithm is analyzed (Figure 4.5), and Figure 4.6 presents the top 10 most important predictors from the best-performing model for the complete sample (RT).

The combined model highlights an even distribution between laboratory results and patient profile factors, with hospital admissions such as ES_6m (0,1359) and Hosp_2020_Now (0,0866) playing a dominant role. Laboratory variables play a key role in predicting HF decompensation, with Cr and Na ranking among the top three predictors, showing high importance values (0.1004 and 0.0951, respectively). Other markers, such as WBC and RBC, also rank within the top 10 predictors, with values of 0.0678 and 0.0577, respectively, further highlighting the support on laboratory indicators for accurate predictions.

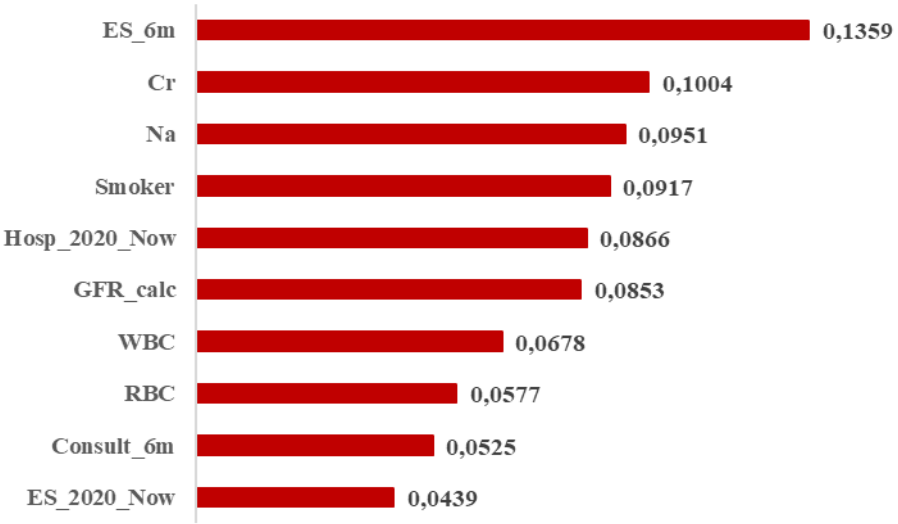


Figure 4.5: Top 10 - Variable Importance for the combined model

In contrast, the complete model includes clinical presentation variables, which are entirely absent in the combined model. NYHA_cod, for example, is the top predictor in the complete model (1), signaling the relative importance of clinical symptoms in

identifying decompensation risk. Laboratory variables like Hb, Hct, and NT-proBNP are present in the complete model but only appear starting from the fourth place in the predictor relative importance list (0.79, 0,71 and 0.48, respectively). Both models emphasize unplanned hospital visits, particularly ES visits. However, the complete model also highlights the importance of DH visits, while the combined model focuses more on hospitalizations. This illustrates how each model prioritizes different factors in managing HF.

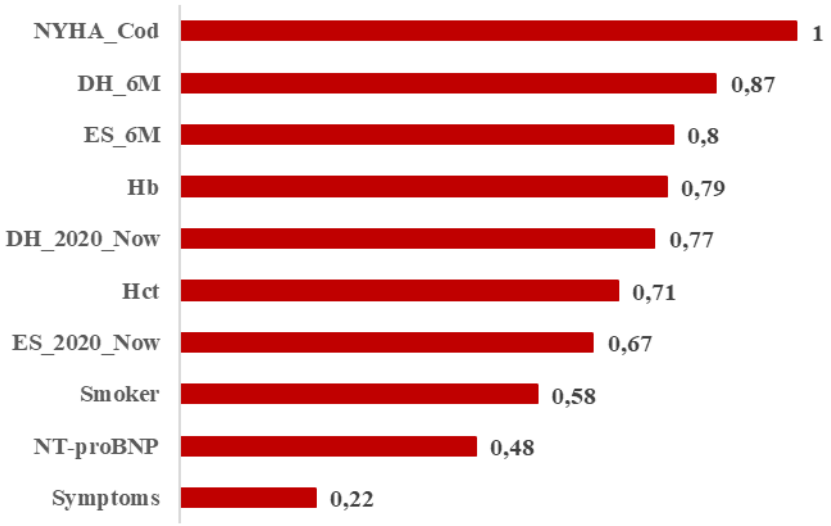


Figure 4.6: Top 10 – Variable by Relative Importance for the complete sample

4.2.1. Sensibility Analysis

To provide more insight and enhance the explainability of the combined model, a sensitivity analysis is conducted, focusing on NT-proBNP, GFR, and Na, due to their strong influence in the models and clinical significance according to the HF experts and literature. NT-proBNP is a key biomarker for HF diagnosis and monitoring, while GFR is preferred over absolute Cr level, because of the inclusion of specific patient variables in its equation, as patient sex, age, and creatinine levels, providing a more accurate assessment of renal function. Na, crucial for maintaining electrolyte and volume balance, offers valuable information about the fluid volume status of the patient. The goal is to explore how these variables relate with ES_6m, the most significant predictor in the combined model.

The continuous variables are transformed into nominal categories (bins). ES_6m is categorized in two ways: one that compares no emergency service visits in the last six months with at least one, and another that divides the visits into three groups for a more in-depth comparison. GFR and Na follow standard clinical reference ranges (Figure 4.7

and Figure 4.8, respectively), while for NT-proBNP, less than 300 pg/mL is used as the clinical reference, and 3646 pg/mL, the average value in the sample, differentiates the middle and upper risk categories (Figure 4.9).

The heatmap bellow shows that the highest risk of decompensation (mean score 0.9) is observed for patients with a low GFR (below 30 mL/min./1.) and five or more ES visits, indicating that renal dysfunction coupled with frequent hospitalizations increases the chances of HF decompensation. In contrast, the lowest risk (mean score 0.3) is seen in patients with GFR above 90 mL/min./1. and fewer than two visits.

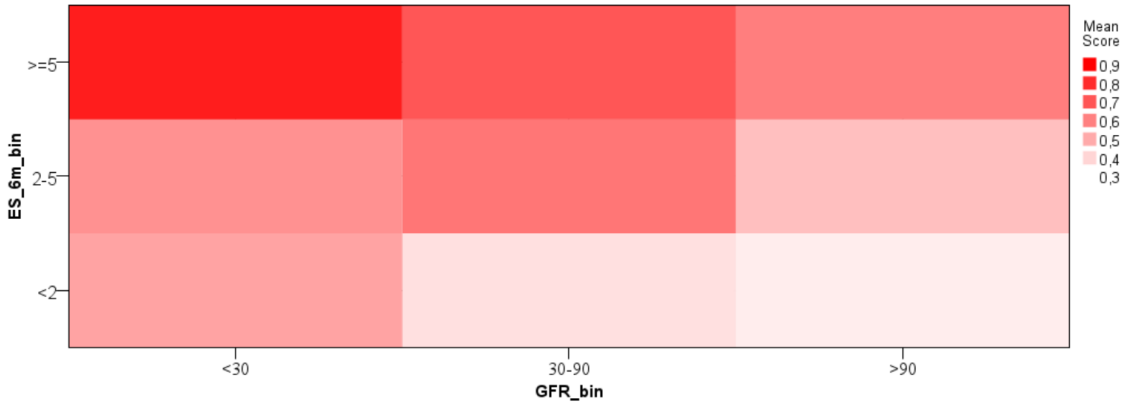


Figure 4.7: HF Decompensation Score by GFR levels and ES visits

When comparing Na (Figure 4.8) and NT-proBNP (Figure 4.9), both markers show increased decompensation risk when their values are elevated, but sodium presents a higher risk. Patients with Na levels above 145 mmol/L and more than one ES visit face a higher mean score of 0.7. In comparison, NT-proBNP levels above 3646 pg/mL, under similar conditions, result in a slightly lower score of 0.6. On the lower end, patients with normal Na levels (135-145 mmol/L) and NT-proBNP levels below 300 pg/mL, combined with fewer than one ES visit, remain in the low-risk category (mean score 0.3).

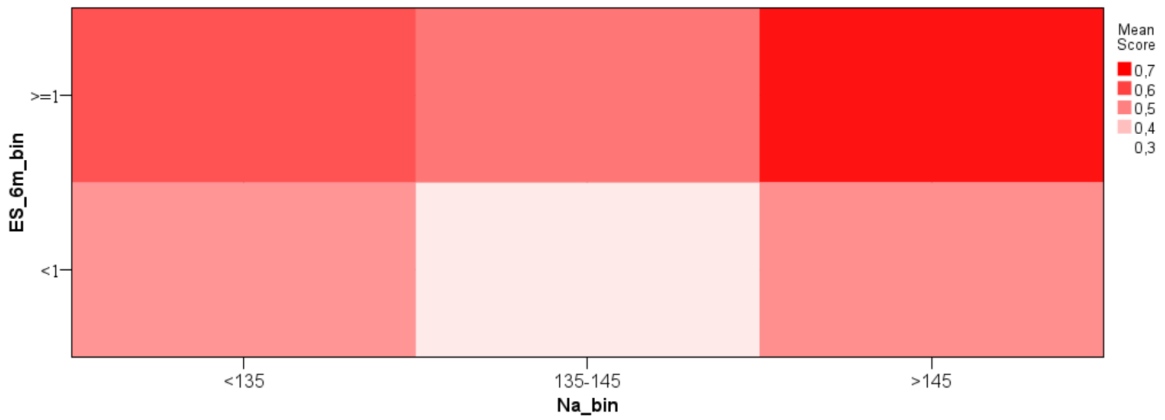


Figure 4.8: HF Decompensation Score by Na and ES visits

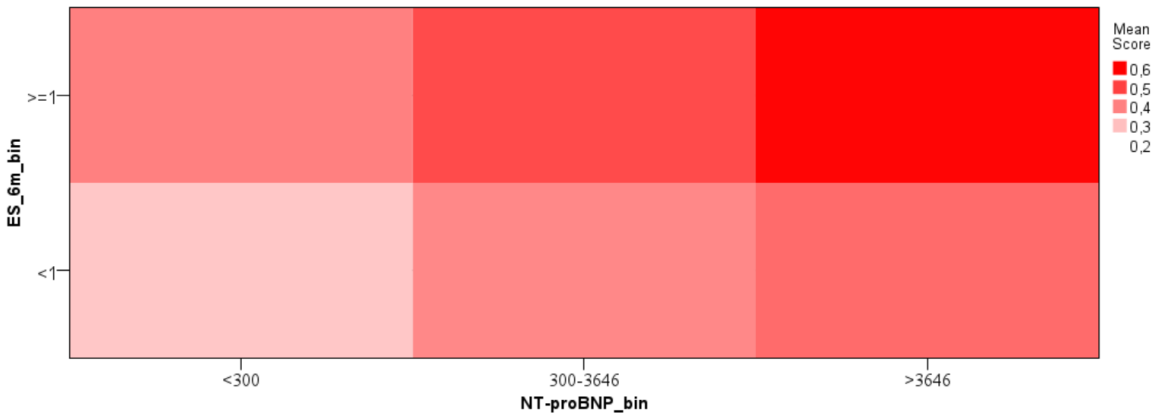


Figure 4.9: HF Decompensation Score by NT-proBNP and ES visits

4.2.2. Variable Impact through Logistic Regression

To assess the impact of key variables identified in the predictive models, two LR analyses are performed to predict HF decompensation probability, as outlined in the evaluation subchapter of the methodology. The first model (equation 1) includes all significant variables, while the second (equation 2) excludes past variables, providing insights for new patients. By analyzing the coefficients and odds ratios, the contribution of each factor to HF decompensation risk becomes clearer [13].

While NT-proBNP and unplanned visit-related variables (ES_6m and DH_2020_Now) are included in the LR because of their significance (p-value of 0.002, <0,001 and <0,001, respectively), their specific impact is not detailed in this analysis. This is because NT-proBNP has an odds ratio close to 1 (1.003), indicating that one-unit increase of NT-proBNP has minimal effect likely due to its high variability, and the impact of unplanned visits is considered more intuitive and thus not shown.

Therefore, the focus is on Hb and clinical status indicators (such as PND presence, NYHA classification above 1, and a Cr increase equal or superior to 0.3 mg/dL) that are identified as significant, both statistically (p-value <0,1) and clinically.

$$1)\ P(Decompensation\ in\ 30\ days) = \frac{1}{1 + e^{-(-0{,}151+0{,}496ICD(1)+0{,}684PND(1)+0{,}262ES_{6m}+0{,}12DH_{2020_{Now}}-0{,}147Hb+0{,}614Var1_{30_{Cr(1)}}-0{,}511NYHA(1)}} \quad (1)$$

$$2)\ P(Decompensation\ in\ 30\ days) = \frac{1}{1 + e^{-(0{,}182+0{,}372Sex(1)+0{,}521ICD(1)-0{,}151Hb+0{,}003NT-proBNP-0{,}339NYHA(1))}} \quad (2)$$

[13] See appendix G

The Figure 4.10 illustrates the impact of Hb variation on HF decompensation probability for the patients listed in Table 4.2, while Table 4.3**Table *4.3*** details how changes in Hb levels and clinical indicators influence these probabilities. The probabilities for patients 1_0 and 1_1 are derived from equation (1), whereas patients 2_0 and 2_1 follow equation (2), where the first number indicates the equation and the second denotes the presence of clinical indicators (0 for none, 1 for presence)

There is a clear negative relationship between Hb levels and the probability of HF decompensation within 30 days. For patient 1_0, without clinical status indicators, a reduction in Hb increases the decompensation probability by 25.3%, while an increase has a similar magnitude in the opposite direction (-21.3%). However, in the presence of clinical status indicators, as seen with patient 1_1, the effect of Hb changes is much smaller, with only a 10.8% increase for a reduction in Hb and an 11.6% decrease when Hb increases. Additionally, maintaining Hb at 13 but adding clinical status indicators nearly triples the probability of decompensation (196%), increasing from 0.208 to 0.616. This change moves patients from a lower-risk category to a moderate risk, where the model now indicates a higher chance of decompensation. This highlights how the combination of multiple factors such as PND, NYHA and specific variations on Cr and Hb significantly increases the overall risk of HF decompensation.

For patients 2_0 and 2_1, the impact of Hb variations follows a similar pattern to that of patient 1_0. However, the difference between 2_0 and 2_1 is much smaller due to the presence of only one clinical status indicator (NYHA). The addition of NYHA above 1 leads to a modest 28.6% increase in decompensation probability, keeping the patient in a low-risk category with a probability of 0.261.

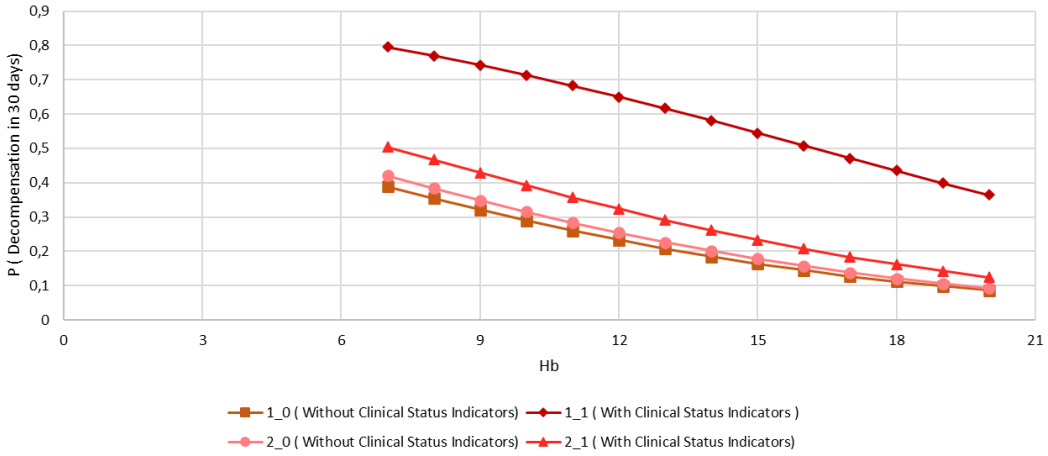


Figure 4.10: Impact of Hb Levels on 30-Day Decompensation Probability by Clinical Status

Table 4.2: Patient profiles and respective 30-day decompensation probability

Patient	ICD*	PND*	ES_6m	DH_2020_Now	Hb	Var1_0.3_Cr*	NYHA_1*	Sex*	NTproBNP_scale	Probability
1_0	1	0	1	4	13	0	1	-	-	0,208
1_1	1	1	1	4	13	1	0	-	-	0,616
2_0	1	-	-	-	13	-	1	1	0,336	0,227
2_1	1	-	-	-	13	-	0	1	0,336	0,261

Note: * are binary and coded as 1 for a positive outcome. This means the impact is measured when conditions such as the presence of PND, having an ICD, being female, or having a NYHA classification of 1 are met.

Table 4.3: Impact of Hb variation and clinical status indicators on 30-day decompensation probability

	Hb		Clinical Status Indicators
Patient	13 to 11	13 to 15	Absent to Present
1_0	25,3%	-21,3%	195,8%
1_1	10,8%	-11,6%	
2_0	25,2%	-21,4%	28,6%
2_1	22,7%	-20,0%	

Focusing on the remaining variables in the equations, the odds ratios and coefficients reveal that: being female, more emergency service visits in the past 6 months, more day hospital visits since 2020, higher levels of NT-proBNP, and having an ICD all positively influence the likelihood of decompensation. However, ICD is a preventive measure and is therefore expected to have a negative impact on decompensation. The positive association observed may reflect that patients more prone to decompensation are often those who receive the device. This underscores the importance of interpreting statistical findings within a clinical context

4.3. Decision Rules to Support Patient Profiles

The fourth and final objective, identifying profiles of patients more and less prone to decompensation after 30 days of the consultation, and those where the model struggles to classify, is now addressed. To offer a thorough comparison, decision rules that can support patient profiles are provided for both the complete model and the combined model, balancing explainability with detailed comparison. For each model, three profiles are presented: stable patients (low score), those who experience decompensation within 30 days (high score), and cases where classification is uncertain (score around 0.5).

The rules for these profiles are derived from the C5.0 model for the combined data. Profiles are determined based on rules with a confidence level of at least 80% and

support of at least 50 consultations for less prone (classification of 0) and more prone (classification of 1) profiles. For the uncertain profiles, confidence ranges from 0.5 to 0.7, with support of more than 20 cases. For the complete model, the RT algorithm does not provide the same rule evaluation metrics as it lacks the support number, so rule accuracy is used instead. This metric is similar to confidence and follows the same success criteria as the combined model.

Rules from the combined model (C5.0):

- NYHA_cod = {0, 1, 2, 3, 4, 5} and DH_2020_NOW < 4 and Hosp_6m < 1 and ES_6m < 1 and Hb > 14.25 and Na ≤ 138.55, then 0 (support: 92; confidence: 99.8%);
- NYHA_cod = {0, 1, 2, 3, 4, 5} and DH_2020_NOW < 4 and Hosp_6m < 1 and ES_6m < 1 and Hb > 14.25 and WBC < 138.75 and Na > 138.75, then 0 (support: 171; confidence: 98.4%);
- DH_2020_NOW < 4 and Hosp_6m < 1 and ES_6m < 1 and Hb < 14.25 and 85.350 < Urea < 172.5 and ICD = 0 and Hosp_2020_Now < 1 and Hct < 45.450 and Symptoms < 1 and NYHA_cod = 1, then 0 (support: 93; confidence: 96.8%);
- NYHA_cod = {0, 1, 2, 3, 4, 5} and DH_2020_Now< 4 and Hosp_6m < 1 and ES_6m > 1 and DH_2020 < 2 and WBC < 6.650 and Cr > 1.035 and GFR_calc < 74.5 and 1 < Consult_6m < 2, then 1 (support: 72; confidence: 81.9%);
- NYHA_cod = {0, 1, 2, 3, 4, 5} and DH_2020_Now < 5 and Hosp_6m > 1 and ES_6m > 4, then 1 (support: 80; confidence: 82.5%);
- NYHA_cod = {0, 1, 2, 3, 4, 5} and DH_2020_Now < 2 and 0< Hosp_6m < 2 and ES_6m > 1 and GFR_calc < 86.5 and ES_2020 < 13 and Urea < 113.5 and Smoker = 0 and Na < 143.450 and K > 3.860 and 6.65 < WBC < 10.5, then 1 (support: 74; confidence: 82.9%);
- NYHA_cod = {0, 1, 2, 3, 4, 5} and Hosp_6m < 1 and ES_6m > 2 and DH_2020_Now ≤ 2 and WBC ≤ 6.65 and Cr > 1.035 and GFR_calc < 74.5 and Consult_6m ≤ 0 and Symptoms ≤ 3, then 0 (support: 21; confidence: 66.3%);
- DH_2020_Now < 4 and Hosp_6m ≤ 1 and ES_6m ≤ 1 and Hb > 14.25 and ICD = 0 and Hosp_2020_Now ≤ 1 and Urea < 85.350 and NYHA_cod = {1, 0} and Hct < 45.45 and Symptoms > 2 and Previously_hospitalized_12m = 0, then 1 (support: 39; confidence: 68.7%);
- NYHA_cod = {0, 1, 2, 3, 4, 5} and DH_2020_NOW ≤ 4 and Hosp_6m > 2 and ES_6m < 4 and WBC > 4.95 and Sex = 0 and ES_6m > 3 and Hb ≤ 13, then 1 (support: 39; confidence: 75.4%).

Profiles from the complete model (RT):

- Hb ≤ 12.5 and DH_6m ≥ 2, then 1 (rule accuracy: 95.5%);
- DH_6m > 2 and Hct ≤ 48 and NYHA_cod = {0, 1, 2, 3, 4, 5}, then 1 (rule accuracy: 91.2%);

- DH_2020_NOW ≤ 5 and NT-proBNP > 5604 and ES_6m > 2 and DH_6m ≤ 0, then 1 (rule accuracy: 87.0%);
- DH_2020_NOW ≤ 5 and NT-proBNP < 5604 and ES_6m ≤ 1 and Hb > 14.2, then 0 (rule accuracy: 78.9%);
- Hb > 13 and DH_6m ≤ 0 and NT-proBNP ≤ 3358 and ES_6m ≤ 1, then 0 (rule accuracy: 76.1%);
- DH ≤ 2 and Hct > 38.6 and ES_6m < 1 and NYHA_cod < 2, then 0 (rule accuracy: 76.7%);
- DH_2020_NOW ≤ 5 and ES_6m > 1 and ES_2020_NOW > 5 and NT-proBNP < 2195, then 1 (rule accuracy: 51.9%);
- Hb > 12.5 and ES_6m > 1 and ES_2020_NOW ≤ 4 and GFR_calc ≤ 98, then 0 (rule accuracy: 52.1%);
- Hb ≤ 13 and DH_6m ≤ 2 and hosp_6m ≤ 1 and LOS_last ≤ 5, then 0 (rule accuracy: 53.8%);

From the presented decision rules, it is evident that demographic variables are largely absent, with Sex being the only one included, appearing just once. NYHA also features frequently but only to distinguish from the most severe cases, as no rules are created for cases where NYHA_cod is 7, corresponding to a NYHA class of 4. Instead, the rules are primarily composed of lab results and hospital admission variables, with symptoms appearing a smaller number of times. It is important to note that the RT model achieved high rule accuracy for profiles more prone to decompensate but marginally missed the success criteria for rules associated with profiles less prone to decompensate, which is consistent with its previously observed evaluation metrics. In contrast, the combined model successfully meets all criteria, thereby fulfilling the final analytical objective. The translation of these rules into patient profiles is provided in the next subchapter, Discussion and Practical Implications.

4.4. Discussion and Practical Implications

In recent years, there has been a growing focus on developing ML models for HF, primarily targeting outcomes like readmission and mortality. Despite this progress, there has been limited attention specifically on HF decompensation, a concept whose definition remains debated, making further research essential. In the health domain, the need for interpretability ML models is crucial to ensure that models can be understood and trusted by clinicians, ultimately leading to better decision-making. Addressing these needs, this study contributes to the understanding of HF decompensation by developing an ML model focused on a 30-day timeframe, exploring key predictive factors and their influence on HF decompensation risk.

The study conducted by Kerexeta et al. (2023) has already made steps in this direction by developing a predictive model for HF decompensation within 7 days, relying on telemonitoring data. In contrast, this study uses a dataset of 584 patients followed in HF outpatient consultations, incorporating unstructured information such as consultation records, hospital admissions, and laboratory results, an approach not seen in the articles included in the SLR. Additionally, the longer 30-day follow-up period used here provides a more extended view of HF decompensation risk.

The methodology applied is CRISP-DM, a framework not used in any of the reviewed studies. During the business understanding phase, the first objective is addressed by defining HF decompensation through expert input and aligning it with definitions from Bozkurt (2023) and Greene et al. (2023). From the data understanding and preparation stages, over 600 variables are initially created, classified into three dimensions: patient profile, laboratory results, and clinical presentation. Additionally, the target variable is created by combining ES and DH visits, hospitalizations, and therapy optimizations, all occurring within 30 days after one consultation. DH visits represent an additional type of unplanned hospital visit, not included in the Kerexeta et al. (2023) study. Therapy optimization is defined as the initiation or increment of furosemide and the initiation of levosimendan intermittent cycles. The feature selection process, following an approach similar to Rahman et al. (2023) and Sarijaloo et al. (2021), refines the variables within the three dimensions, resulting in 26 key factors for the modeling phase. The iterative nature of CRISP-DM proves essential, allowing for constant adjustments during data preparation and understanding, ensuring a robust final model and a clear pathway for future replication (Chapman et al., 1999). Several techniques are applied, including descriptive techniques, exploratory analyses, and predictive analyses through ML with supervised learning, all guided by both the literature and the need to meet the analytical and business objectives.

The results presented in this chapter highlight that the combination of models resulting from the division of samples based on null values, specifically XGBoost and RT (via the Auto Classifier), delivers the best performance. Although this technique of sample division is not seen in the literature, the outstanding performance of XGBoost has already been demonstrated in multiple studies (Polo Friz et al., 2022; Rahman et al., 2023), and its strong results in this context further reinforce its effectiveness. This model achieves all metrics above 80%, with a recall of 81.40% and an AUC of 0.964, surpassing the decompensation model of Kerexeta et al. (2023). and the 30-day readmission models

(Polo Friz et al., 2022; Sabouri et al., 2023; Sharma et al., 2022), which are the most comparable to predict decompensation.

This model functions as a 'black box' due to the combination of a simpler DT (RT) with a powerful ensemble method (XGBoost), which is non-interpretable by itself, making it necessary to enhance the interpretability of the overall model. These efforts include creating a DT algorithm (C5.0) to mimic the combined model, achieving strong metrics, with all above 90% except for precision at 87.7%. An error analysis is also conducted to improve performance and show how the model works, along with a sensitivity analysis of the most important variables and a LR to assess the impact of significant variables. For additional insights, the results of the RT model are presented as references, with metrics comparable to the literature but falling below the success criteria, specifically in terms of recall.

The analysis of the most important variables from the combined model shows a strong emphasis on lab results and hospital interactions. Key predictors include Cr, Na, GFR, RBC, and WBC, aligning with findings in the literature (Okoye et al., 2023). In this analysis, WBC is used in place of NLR, a predictor mentioned in the literature but not collected by the hospital, and a known important indicator of inflammation, making it a reliable substitute. While the number of hospitalizations is often highlighted in the literature (e.g., Sharma et al., 2022; Xu et al., 2023), this study underscores the predictive value of the number of unplanned ES visits in the last six months, regardless of whether they led to hospitalization or not. This aligns with established scores like LACE and LaCE, where this variable is a core component, and supports the findings from Soliman et al. (2023). This indicates that even less severe decompensations, not requiring hospitalization, already reflect a higher risk for future decompensation. As such, this study allows for earlier intervention, at a less advanced stage when in-patient admission for optimization is not yet required. Although age is frequently noted in the literature as a predictor of HF outcomes, it does not emerge as significant in this analysis. Instead, lab results and healthcare interactions stand out as more critical factors in predicting decompensation. In the lifestyle-related variables, smoking is significant in both the combined and complete models, pointing to its potential impact on HF outcomes, but not previously mentioned by the literature. In the complete model, additional variables like Hb and Hct also emerge as significant, aligning with established research on HF management (Beecy et al., 2020).

The sensitivity analysis enhances the interpretability of the combined model by exploring how key clinical variables interact with ES_6m, the most significant predictor. Results show that sodium levels above 145, combined with more than one emergency visit, are classified by the model as prone to decompensation. For NT-proBNP, levels below 300 pg/mL are associated with the lowest score, while higher levels, especially when combined with frequent emergency visits, lead to significantly higher scores. Similarly, GFR below 30 mL/min./1., combined with more than five emergency visits, achieves a score of 0.90, highlighting the impact of renal function and healthcare interactions.

The LR provides a clearer understanding of how the most significant variables directly influence the likelihood of decompensation, a technique also applied in the study by Kerexeta et al. (2023). By constructing two models, one without patient history variables and one with them, the regression results show that patient history, specifically the number of ES visits in the last 6 months and the number DH visits since 2020, is critical for predicting decompensation probabilities above 50%. Additionally, the regressions confirm the expected negative relationship between Hb and decompensation, as well as the significant impact of symptoms like PND, increased NYHA classification, and a Cr increase of 0.3 mg/dL or more on HF decompensation probability.

The model identifies patients as less prone to decompensation with higher levels of hemoglobin (above 14.25 g/L) combined with low ES visits and hospitalizations. Patients with Hb below 14.5g/L remain less prone to decompensate when combined with absence or up to one symptom. In contrast, profiles with more than one ES visit, combined with poor renal function (defined as creatinine levels above 1.035 mg/dL and GFR below 74.5 mL/min./1.), are consistently identified as having a higher risk of decompensation. In cases where renal function is preserved but Na levels are reduced (below143 mmol/L), patients are still flagged as likely to decompensate. The combination of frequent healthcare interactions and deteriorating renal markers significantly drives this HF decompensation risk. For patients where the model is uncertain, conflicting clinical indicators complicate the accuracy of the outcome. These include low ES visits and normal WBC levels (between 6.66 and 10.5 X 10^9 / L) alongside poor renal function (GFR below 74.5 mL/min./1.), more than one symptom, or low Hb levels (below 13 g/L). Such cases reflect a conflict between risk and protective factors within the same profile, requiring closer evaluation and monitoring to clarify the decompensation risk of the patient.

In comparison, the complete model confirms many of these insights while incorporating additional variables. It highlights low Hb and Hct (below 12.5 g/L and below 48%, respectively) with frequent unplanned hospital visits as significant predictors of decompensation, with high levels of NT-proBNP (above 5604 pg/mL) further distinguishing those at risk. Interestingly, the model shows that NT-proBNP reduction, even with lower levels of Hb, indicates stability. For stable patients, the complete model aligns with the combined model by emphasizing higher Hb and Hct (above 13 g/L and above 38.6%, respectively), fewer hospital visits, and normal NT-proBNP, reinforcing the protective role of lack of anemia and minimal unplanned healthcare interactions. For uncertain profiles, both models struggle when hospital visits coexist with borderline lab values, particularly in renal function and hemogram parameters, complicating predictions.

This analysis demonstrates that all four study objectives are achieved, offering a solid foundation for HF decompensation prediction, with Table 4.4 providing targeted recommendations for stakeholders to guide further research.

Table 4.4: Recommendations to key stakeholders

Recommendation	Stakeholder
Promote research on HF decompensation to refine its definition and identify consistent predictive factors over a 30-day period.	Academics, Healthcare Policy Makers
Establish a standard practice to always request key lab tests (e.g., Cr, Na, NT-proBNP) to reduce missing values and improve predictive accuracy.	Health care professionals
Standardize the language and abbreviations in unstructured records, including non-clinical factors, to improve data extraction and insights into patient compliance.	Health care professionals
Apply the CRISP-DM methodology for replicability and iterative improvement in clinical ML model development.	Researchers of Business analytics and health analytics
Conduct cost-effectiveness analyses to support HF model implementation in routine clinical practice.	Hospital Administrators, Healthcare Policy Makers
Utilize the SLR, which summarizes existing knowledge on ML predictive models for HF outcomes, as a foundation for future research.	Researchers of health analytics
Implement online symptom self-reporting, enabling flexible management of mild cases and creating a valuable data source for research.	Healthcare providers, patients, health IT developers, researchers.
Raise public awareness about heart failure and the benefits of early intervention and regular check-ups to reduce HF risks.	Public Health Organizations, Healthcare professionals, Media Channels
Utilize this investigation as a practical example of applying predictive analytics in cardiology, integrating Business Analytics techniques to enhance patient outcomes in the Portuguese healthcare context.	Business Analytics Professors, researchers
Use this research as a practical example for identifying key factors in HF decompensation, supporting early interventions to reduce hospital readmissions and emergency visits.	Professors and Healthcare professionals of cardiology

5. Conclusion

HF affects millions of individuals worldwide, and as this syndrome progresses, decompensation events occur, significantly worsening the overall condition of the patient. These events are critical in the trajectory of HF, often leading to hospitalizations, reduced quality of life, and increased mortality risk. Due to this, HF places a significant burden on healthcare systems, contributing to increased spending. By anticipating these decompensation events, cardiologists can intervene earlier, reducing hospitalizations and ES visits, and ultimately alleviating the strain on healthcare resources.

Given the significant impact of HF on patients and healthcare systems, along with the added strain that decompensation events create, this investigation aims to answer the question: How can a ML model for HF decompensation prediction enhance the management of HF patients? To achieve this, four objectives were established: to define HF decompensation, develop a predictive model for HF decompensation, identify the key factors influencing these events, and determine patient profiles more prone to decompensation as well as those less susceptible. To address this, an SLR was conducted to gather existing knowledge in this domain, focusing on study contexts, methodologies used, results, and the impact of each study, which resulted in the analysis of 24 articles.

Building on the knowledge gathered from the SLR, the CRISP-DM methodology was applied to the data from 584 patients, resulting in 2008 cases, as the data was structured per consultation rather than per patient. The data passed through all phases of the methodology, with iterations between phases as needed. Each stage was guided by the objectives of the research, ensuring that data preparation and analysis are in line with the objectives of the study. Continuous contact with HF experts was crucial throughout all phases of the process, particularly during business understanding, where the first objective of defining decompensation was completed, and during data understanding and preparation, where unstructured consultation records underwent data mining. In addition, expert input was vital in analyzing the results, providing a clinical perspective to ensure that the conclusions were aligned with practical relevance, not based solely on statistical results.

In this study, HF decompensation is defined as the combination of unplanned hospital visits (ES and DH visits and hospitalization) and scheduled HF appointments with

diuretic and/or outpatient inotropic therapy optimization, all occurring within 30 days after an HF appointment.

For the construction of the ML model, several algorithms were tested, including DT models such as CHAID, C5.0, and RT, chosen for their interpretability. Less interpretable models like XGBoost and NN were also included, as they have demonstrated strong performance in previous studies, despite their lack of interpretability. The best-performing model resulted from combining an XGBoost model, created for a sample without null cases, and a RT model for the sample with nulls. This combined approach achieved the best metrics, meeting the success criteria and surpassing existing models for decompensation and 30-day readmission.

To open the "black box" of the best-performing model, a C5.0 DT was created to mimic its behavior. This was complemented by an error analysis, sensitivity analyses, and a LR analyses. Relying on feature importance from the C5.0 algorithm, the most significant variables were identified, successfully achieving the third objective. The key predictors include the number of emergency visits in the last six months, followed by Cr, Na, smoking status, and the number of hospitalizations from 2020 to the present consultation. The remaining top predictors are primarily related to hospital interactions and laboratory results. The sensitivity analysis provided insights into how clinical key variables, such as NT-proBNP, Na, and GFR, along with the number of emergency visits in the last six months, influenced the probability of the model classifying a patient as decompensated. Specifically, a GFR below 30 mL/min./1, combined with more than five ES visits in the last six months significantly increased the likelihood of classification as decompensated.

Moreover, the LR analysis demonstrated that the presence of PND along with a NYHA classification above one and an increase over 0.3mg/dL in creatinine levels amplified the probability of decompensation by nearly 200%, compared with the absence of this indicators, ceteris paribus. It also highlighted the negative relationship between Hb levels and the likelihood of being classified as decompensated. Furthermore, the presence of historical variables, such as the number of ES visits in the last six months, proved essential for any model to achieve a reliable level of certainty in predicting decompensation.

Regarding the identification of profiles, terminal nodes of the C5.0 were employed to establish three profiles for each type: those more prone to decompensation, those less prone, and those classified with uncertainty. The decision rules to support patient profiles

for both the more prone and less prone categories were created with confidence levels exceeding 80% and support greater than 50 consultations. In contrast, the uncertainty profiles were characterized by confidence levels closer to 50% and support above 20. The rules from the best-performing model in the complete sample, the RT model, were also presented for comparison, resulting in a total of 18 profiles. This process successfully fulfilled the last objective.

With all objectives achieved, it is possible to answer to the research question "How can a ML model for HF decompensation prediction enhance the management of HF patients?" This thesis demonstrates that by creating two distinct models, XGBoost for cases without null values and RT for cases with nulls, and combining them, metrics exceeding 80% are consistently achieved. By applying a C5.0 algorithm alongside sensitivity analysis and LR analysis the black box model can be opened, allowing for the identification of the most important variables and HF patient profiles more prone and less prone to decompensation within 30 days after their consultation, contributing to the explainability (a pre-requisite). The implementation of this model in clinical practice can provide cardiologists with valuable insights to better identify high-risk patients, facilitating more targeted interventions, closer monitoring, and adjustments in therapy, ultimately leading to less HF decompensation episodes. This, in turn, leads to more efficient resource allocation and improves outcomes in the management of HF patients.

5.1. Contributions

This research presents advancements in the study of HF decompensation prediction through ML techniques, benefiting both the academic domain and healthcare professionals. By employing the CRISP-DM methodology, an approach not previously applied in existing literature, it demonstrates the effectiveness of this methodology in healthcare analytics.

For future research on predicting HF outcomes, this study serves as a valuable starting point, with a SLR summarizing findings related to readmission, mortality, and combined endpoints predictions for HF patients. It identifies key insights across various dimensions, including scope and context, methodology, results, and study impacts. One work, based on this research, was presented at a conference on healthcare informatics track and is included in the proceedings, which are scheduled for publication, providing another resource for researchers and practitioners in the field.

The results significantly enhance the literature, as there is limited information on decompensation prediction. By extending the follow-up period from 7 days to 30 days, this study provides valuable insights for longer-term predictions. It utilizes clinical notes, healthcare interaction history, and previous lab results, which eliminates the need for costly telemonitoring devices. Conducting this research in the Portuguese context offers a fresh perspective on the application of predictive techniques for anticipating HF decompensation.

The findings suggest that predicting decompensation can be complex, indicating that more interpretable models may not be adequate for such intricate outcomes that involve numerous variable interactions. This underscores the need to explore black box models, which, despite their complexity, can be enhanced with interpretability techniques presented in this study. Additionally, the feature selection process outlined here can be replicated, aiding in the reduction of model complexity by requiring only 25 variables. This approach demonstrates that extensive data collection is not necessary, thereby facilitating future investigations.

Notably, this investigation demonstrates that a simple tool like Excel can be effectively used for data mining. Sharing the developed formulas and queries, along with recommendations for improving clinical note-taking to facilitate data extraction, can help the HF team streamline their data collection process, improving both the efficiency and quality of future data retrieval. As cardiologists are expected to conduct research independently, these insights can significantly reduce data collection time and enhance data quality. Another tool highlighted in this thesis is SPSS Modeler, a low-code platform with a visual programming interface that further enhances efficiency in data understanding, preparation, and modeling. Its user-friendly design supports rapid insights and empowers self-service data science by reducing the time needed for meaningful results

Furthermore, identifying the most important factors associated with HF decompensation, along with profiling patients as more prone or less prone to decompensation, will aid in refining patient management strategies. By identifying patients at higher risk, cardiologists can tailor their consultation plans. They may implement more frequent follow-ups for high-risk patients, while conducting telephone consultations for stable patients to reduce the need for in-person visits, thereby optimizing resources based on patient profiles. Additionally, meetings for case discussion can be

reduced to focus only on patients where the model indicates uncertainty, streamlining the process and improving efficiency.

Ultimately, the insights gained from this research are valuable for all cardiologists and could be applicable to other pathologies characterized by stable phases interrupted by episodes of decompensation, even though the study focused on a specific condition in patients from a particular hospital. By integrating this domain with business analytics, this work seeks to generate greater interest in similar projects while ensuring that concerns about interpretability are addressed, without allowing these issues to hinder future research.

5.2. Limitations and Future Research Recommendations

Despite the strong results achieved by the model, several limitations should be acknowledged. The investigation faced challenges due to the large amount of information in an unstructured format, which significantly extended the data preparation phase. This complexity resulted in the exclusion of important information, such as prognosis-modifying medication, specific medication dosages and the results from complementary exams like transthoracic echocardiograms and electrocardiograms, and it also impacted the quality of data concerning symptoms and signals. This compromised quality could have led to the reduced significance of these variables in relation to the outcome.

Additionally, the presence of missing values led to the exclusion of crucial lab results, including albumin, cystatin C, ferritin and urinary sodium levels, with the feature selection process also discarding variables intended to provide historical context for lab results and symptoms. It is important to note that the formulas for data extraction were specifically created for the HF team in question, using Portuguese terminology that this team tends to use. Therefore, they should only be viewed as a guideline for how other hospitals might structure their data.

Lastly, despite efforts to enhance interpretability, the complexity of decompensation, which is influenced by many factors and different combinations of predictors, makes more interpretable models less effective. Although some authors suggest using decision trees to interpret black box models, other tactics that could aid in this process were not available in the software used.

There is significant potential for improvement in HF decompensation predictions. Future studies should build upon these findings and enhance model performance by

focusing on the following recommendations: 1) Utilization of large language models for data extraction to improve research efficiency; 2) Inclusion of guideline-directed therapy data to improve predictive accuracy, as low doses or absence due to intolerance correlates with higher readmissions; 3) Incorporation of non-clinical factors such as autonomy, socioeconomic status, and patient perception of their own health to gain insights into patient compliance and its impact on HF progression; 4) Improvement of model interpretability by employing techniques such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive Explanations (SHAP) to clarify factors influencing predictions, aiding clinicians in decision-making; 5) Evaluation of the financial impact of implementing this model to understand its cost-effectiveness.

References

Adler, E. D., Voors, A. A., Klein, L., Macheret, F., Braun, O. O., Urey, M. A., Zhu, W., Sama, I., Tadel, M., Campagnari, C., Greenberg, B., & Yagil, A. (2020). Improving risk prediction in heart failure using machine learning. *European Journal of Heart Failure*, *22*(1), 139–147. https://doi.org/10.1002/EJHF.1628

Ahmad, M. A., Eckert, C., Teredesai, A., & Mckelvey, G. (2018). *Interpretable Machine Learning in Healthcare*.

Ahsan, M. M., & Siddique, Z. (2022). Machine learning-based heart disease diagnosis: A systematic literature review. *Artificial Intelligence in Medicine*, 128. https://doi.org/10.1016/j.artmed.2022.102289

Althnian, A., AlSaeed, D., Al-Baity, H., Samha, A., Dris, A. Bin, Alzakari, N., Abou Elwafa, A., & Kurdi, H. (2021). Impact of dataset size on classification performance: An empirical evaluation in the medical domain. *Applied Sciences*, *11*(2), 1–18. https://doi.org/10.3390/app11020796

Ambrosy, A. P., Gheorghiade, M., Chioncel, O., Mentz, R. J., & Butler, J. (2014). Global Perspectives in Hospitalized Heart Failure: Regional and Ethnic Variation in Patient Characteristics, Management, and Outcomes. *Current Heart Failure Reports*, 11(4), 416–427. https://doi.org/10.1007/s11897-014-0221-9

Au, A. G., McAlister, F. A., Bakal, J. A., Ezekowitz, J., Kaul, P., & Van Walraven, C. (2012). Predicting the risk of unplanned readmission or death within 30 days of discharge after a heart failure hospitalization. *American Heart Journal*, *164*(3), 365–372. https://doi.org/10.1016/J.AHJ.2012.06.010

Austin, D. E., Lee, D. S., Wang, C. X., Ma, S., Wang, X., Porter, J., & Wang, B. (2022). Comparison of machine learning and the regression-based EHMRG model for predicting early mortality in acute heart failure. *International Journal of Cardiology*, *365*, 78–84. https://doi.org/10.1016/j.ijcard.2022.07.035

Awan, S. E., Bennamoun, M., Sohel, F., Sanfilippo, F. M., Chow, B. J., & Dwivedi, G. (2019). Feature selection and transformation by machine learning reduce variable numbers and improve prediction for heart failure readmission or death. *PLoS ONE*, *14*(6). https://doi.org/10.1371/journal.pone.0218760

Awan, S. E., Bennamoun, M., Sohel, F., Sanfilippo, F. M., & Dwivedi, G. (2019). Machine learning-based prediction of heart failure readmission or death: implications of choosing the right model and the right metrics. *ESC Heart Failure*, *6*(2), 428–435. https://doi.org/10.1002/ehf2.12419

Baracaldo-Santamaría, D., Feliciano-Alfonso, J. E., Ramirez-Grueso, R., Rojas-Rodríguez, L. C., Dominguez-Dominguez, C. A., & Calderon-Ospina, C. A. (2023). Making Sense of Composite

Endpoints in Clinical Research. *Journal of Clinical Medicine*, *12*(13), 4371. https://doi.org/10.3390/JCM12134371

Bayés-Genís, A., Lopez, L., Zapico, E., Cotes, C., Santaló, M., Ordonez-Llanos, J., & Cinca, J. (2005). NT-ProBNP reduction percentage during admission for acutely decompensated heart failure predicts long-term cardiovascular mortality. *Journal of Cardiac Failure*, *11*(5 Suppl). https://doi.org/10.1016/j.cardfail.2005.04.006

Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA - Journal of the American Medical Association*, 319(13), 1317–1318. https://doi.org/10.1001/jama.2017.18391

Beecy, A. N., Gummalla, M., Sholle, E., Xu, Z., Zhang, Y., Michalak, K., Dolan, K., Hussain, Y., Lee, B. C., Zhang, Y., Goyal, P., Campion, T. R., Shaw, L. J., Baskaran, L., & Al'Aref, S. J. (2020). Utilizing electronic health data and machine learning for the prediction of 30-day unplanned readmission or all-cause mortality in heart failure. *Cardiovascular Digital Health Journal*, *1*(2), 71–79. https://doi.org/10.1016/j.cvdhj.2020.07.004

Benedetto, U., Dimagli, A., Sinha, S., Cocomello, L., Gibbison, B., Caputo, M., Gaunt, T., Lyon, M., Holmes, C., & Angelini, G. D. (2022). Machine learning improves mortality risk prediction after cardiac surgery: Systematic review and meta-analysis. *Journal of Thoracic and Cardiovascular Surgery*, *163*(6), 2075-2087. https://doi.org/10.1016/j.jtcvs.2020.07.105

Bettencourt, P., Azevedo, A., Pimenta, J., Friões, F., Ferreira, S., & Ferreira, A. (2004). N-terminal-pro-brain natriuretic peptide predicts outcome after hospital discharge in heart failure patients. *Circulation*, *110*(15), 2168–2174. https://doi.org/10.1161/01.CIR.0000144310.04433.BE

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Błaziak, M., Urban, S., Wietrzyk, W., Jura, M., Iwanek, G., Stańczykiewicz, B., Kuliczkowski, W., Zymliński, R., Pondel, M., Berka, P., Danel, D., Biegus, J., & Siennicka, A. (2022). An Artificial Intelligence Approach to Guiding the Management of Heart Failure Patients Using Predictive Models: A Systematic Review. *Biomedicines*, 10(9). https://doi.org/10.3390/biomedicines10092188

Bozkurt, B. (2023). Proposed New Conceptualization for Definition of Decompensated HF: Taking the Acute Out of Decompensation. *JACC: Heart Failure* 11 (3), 368–371. https://doi.org/10.1016/j.jchf.2023.02.001

Bozkurt, B., Ahmad, T., Alexander, K. M., Baker, W. L., Bosak, K., Breathett, K., Fonarow, G. C., Heidenreich, P., Ho, J. E., Hsich, E., Ibrahim, N. E., Jones, L. M., Khan, S. S., Khazanie, P., Koelling, T., Krumholz, H. M., Khush, K. K., Lee, C., Morris, A. A., … Ziaeian, B. (2023). Heart Failure Epidemiology and Outcomes Statistics: A Report of the Heart Failure Society of America. *Journal of Cardiac Failure*, *29*(10), 1412. https://doi.org/10.1016/J.CARDFAIL.2023.07.006

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, *24*, 123–140.

Butler, J., Braunwald, E., & Gheorghiade, M. (2014). Recognizing Worsening Chronic Heart Failure as an Entity and an End Point in Clinical Trials. *JAMA*, *312*(8). http://jama.jamanetwork.com/

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (1999). *CRISP-DM 1.0 Step-by-step data mining guide*. *The CRISP-DM Consortium*.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, Z., Li, T., Guo, S., Zeng, D., & Wang, K. (2023). Machine learning-based in-hospital mortality risk prediction tool for intensive care unit patients with heart failure. *Frontiers in Cardiovascular Medicine*, *10*. https://doi.org/10.3389/fcvm.2023.1119699

Chowdhury, S., Lin, Y., Liaw, B., & Kerby, L. (2021). *Evaluation of Tree Based Regression over Multiple Linear Regression for Non-normally Distributed Data in Battery Performance*.

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, 110, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Cox, D.R. (1958) The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society: Series B*, 20, 215-242. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x

Dai, Q., Sherif, A. A., Jin, C., Chen, Y., Cai, P., & Li, P. (2022). Machine learning predicting mortality in sarcoidosis patients admitted for acute heart failure. *Cardiovascular Digital Health Journal*, *3*(6), 297–304. https://doi.org/10.1016/j.cvdhj.2022.08.001

Davis, J., & Goadrich, M. (2006). The Relationship Between Precision-Recall and ROC Curves. *Proceedings of the 23 Rd International Conference on Machine Learning*.

De Veaux, R. D., & Ungar, L. H. (1994). Multicollinearity: A tale of two nonparametric regressions. *Selecting Models from Data*, 89, 393–402. Springer.

Donze, J., Aujesky, D., Williams, D., & Schnipper, J. L. (2013). Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. *JAMA Internal Medicine*, *173*(8), 632–638. https://doi.org/10.1001/JAMAINTERNMED.2013.3023

Dörr, M., Riemer, U., Christ, M., Bauersachs, J., Bosch, R., Laufs, U., Neumann, A., Scherer, M., Störk, S., & Wachter, R. (2021). Hospitalizations for heart failure: still major differences between East and West Germany 30 years after reunification. *ESC Heart Failure*, *8*(4), 2546–2555. https://doi.org/10.1002/ehf2.13407

ElShawi, R., Sherif, Y., Al-Mallah, M., & Sakr, S. (2021). Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Computational Intelligence*, *37*(4), 1633–1650. https://doi.org/10.1111/coin.12410

Felix, S. E. A., Bagheri, A., Ramjankhan, F. R., Spruit, M. R., Oberski, D., De Jonge, N., Van Laake, L. W., Suyker, W. J. L., & Asselbergs, F. W. (2021). A data mining-based cross-industry process

for predicting major bleeding in mechanical circulatory support. *European Heart Journal - Digital Health*, *2*(4), 635–642. https://doi.org/10.1093/ehjdh/ztab082

Fonseca, C., Brito, D., Cernadas, R., Ferreira, J., Franco, F., Rodrigues, T., Morais, J., & Silva Cardoso, J. (2017). For the improvement of Heart Failure treatment in Portugal - Consensus statement: Position Paper endorsed by the Heart Failure Working Group of the Portuguese Society of Cardiology, the Heart Failure Nucleus of the Portuguese Society of Internal Medicine, the College of General Practitioners of the Portuguese Medical Council, the Cardiovascular Disease Nucleus of the Portuguese Association of General Practitioners, and the Nursing Nucleus of the Portuguese Society of Cardiology. *Revista Portuguesa de Cardiologia (English Edition)*, *36*(1), 1–8. https://doi.org/10.1016/J.REPCE.2016.10.016

Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M., & Suganthan, P. N. (2022). Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, *115*. https://doi.org/10.1016/j.engappai.2022.105151

Goodman, B., & Flaxman, S. (2017). European union regulations on algorithmic decision making and a "right to explanation." *AI Magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Gouveia, M., Ascenção, R., Fiorentino, F., Costa, J., Caldeira, D., Broeiro-Gonçalves, P., Fonseca, C., & Borges, M. (2019). The current and future burden of heart failure in Portugal. *ESC Heart Failure*, *6*(2), 254–261. https://doi.org/10.1002/ehf2.12399

Gouveia, M. R. de A., Ascenção, R. M. S. e. S., Fiorentino, F., Costa, J. N. M. P. G. da, Broeiro-Gonçalves, P. M., Fonseca, M. C. F. G. da, & Borges, M. de F. P. F. (2020). Current costs of heart failure in Portugal and expected increases due to population aging. *Revista Portuguesa de Cardiologia*, *39*(1), 3–11. https://doi.org/10.1016/j.repc.2019.09.006

Greene, S. J., Bauersachs, J., Brugts, J. J., Ezekowitz, J. A., Lam, C. S. P., Lund, L. H., Ponikowski, P., Voors, A. A., Zannad, F., Zieroth, S., & Butler, J. (2023). Worsening Heart Failure: Nomenclature, Epidemiology, and Future Directions: JACC Review Topic of the Week. *Journal of the American College of Cardiology*, 81(4), 413–424. https://doi.org/10.1016/j.jacc.2022.11.023

Guo, C. Y., Wu, M. Y., & Cheng, H. M. (2021). The comprehensive machine learning analytics for heart failure. *International Journal of Environmental Research and Public Health*, *18*(9). https://doi.org/10.3390/ijerph18094943

Hessel, F. P. (2021). Overview of the socio-economic consequences of heart failure. *Cardiovascular Diagnosis and Therapy*, 11(1). https://doi.org/10.21037/CDT-20-291

IBM. (n.d.). *Auto Classifier node*. IBM Documentation. Retrieved August 25, 2024, from https://dataplatform.cloud.ibm.com/docs/content/wsd/nodes/autoclassifier.html?context=cpdaas.

IBM. (n.d.). *Balance node*. IBM Documentation. Retrieved August 9, 2024, from https://www.ibm.com/docs/sl/spss-modeler/saas?topic=nodes-balance-node

IBM. (n.d.). *C5.0 node*. IBM Documentation. Retrieved August 7, 2024, from https://www.ibm.com/docs/en/cloud-paks/cp-data/5.0.x?topic=modeling-c50-node

IBM. (n.d.). *XGBoost as build options*. IBM Documentation. Retrieved August 7, 2024, from https://www.ibm.com/docs/sl/spss-modeler/saas?topic=node-xgboost-as-build-options

Inker, L. A., Eneanya, N. D., Coresh, J., Tighiouart, H., Wang, D., Sang, Y., Crews, D. C., Doria, A., Estrella, M. M., Froissart, M., Grams, M. E., Greene, T., Grubb, A., Gudnason, V., Gutiérrez, O. M., Kalil, R., Karger, A. B., Mauer, M., Navis, G., … Levey, A. S. (2021). New Creatinine- and Cystatin C–Based Equations to Estimate GFR without Race. *New England Journal of Medicine*, *385*(19), 1737–1749. https://doi.org/10.1056/nejmoa2102953

Jang, S. Y., Park, J. J., Adler, E., Eshraghian, E., Ahmad, F. S., Campagnari, C., Yagil, A., & Greenberg, B. (2023). Mortality Prediction in Patients With or Without Heart Failure Using a Machine Learning Model. *JACC: Advances*, *2*(7). https://doi.org/10.1016/j.jacadv.2023.100554

Jiang, X., Menon, A., Wang, S., Kim, J., & Ohno-Machado, L. (2012). Doubly Optimized Calibrated Support Vector Machine (DOC-SVM): an algorithm for joint optimization of discrimination and calibration. *PloS One*, *7*(11). https://doi.org/10.1371/journal.pone.0048823

Johnson, K. W., Torres Soto, J., Glicksberg, B. S., Shameer, K., Miotto, R., Ali, M., Ashley, E., & Dudley, J. T. (2018). Artificial Intelligence in Cardiology. *Journal of the American College of Cardiology*, 71(23), 2668–2679. https://doi.org/10.1016/j.jacc.2018.03.521

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a Definition of Mixed Methods Research. *Journal of Mixed Methods Research*, *1*(2), 112–133. https://doi.org/10.1177/1558689806298224

Ju, C., Zhou, J., Lee, S., Tan, M. S., Liu, T., Bazoukis, G., Jeevaratnam, K., Chan, E. W. Y., Wong, I. C. K., Wei, L., Zhang, Q., & Tse, G. (2021). Derivation of an electronic frailty index for predicting short-term mortality in heart failure: a machine learning approach. *ESC Heart Failure*, *8*(4), 2837–2845. https://doi.org/10.1002/ehf2.13358

Karakuş, M. Ö., & Er, O. (2022). A comparative study on prediction of survival event of heart failure patients using machine learning algorithms. *Neural Computing and Applications*, *34*(16), 13895–13908. https://doi.org/10.1007/s00521-022-07201-9

Kass, G. V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, *29*(2), 119. https://doi.org/10.2307/2986296

KDIGO 2024 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. (2024). *Kidney International*, *105*(4). https://doi.org/10.1016/s0085-2538(24)00110-8

Kerexeta, J., Larburu, N., Escolar, V., Lozano-Bahamonde, A., Macía, I., Beristain Iraola, A., & Graña, M. (2023). Prediction and Analysis of Heart Failure Decompensation Events Based on Telemonitored Data and Artificial Intelligence Methods. *Journal of Cardiovascular Development and Disease*, *10*(2). https://doi.org/10.3390/jcdd10020048

Kitchenham, B. (2004). *Procedures for Performing Systematic Reviews*.

Laureano, R. (2020). Testes de Hipóteses e Regressão: O Meu Manual de Consulta Rápida (2ª Ed.) Edições Sílabo.

Lee, D. S., Stitt, A., Austin, P. C., Stukel, T. A., Schull, M. J., Chong, A., Newton, G. E., Lee, J. S., & Tu, J. V. (2012). Prediction of heart failure mortality in emergent care: a cohort study. *Annals of Internal Medicine*, *156*(11), 767–775. https://doi.org/10.7326/0003-4819-156-11-201206050-00003

Levy, W. C., Mozaffarian, D., Linker, D. T., Sutradhar, S. C., Anker, S. D., Cropp, A. B., Anand, I., Maggioni, A., Burton, P., Sullivan, M. D., Pitt, B., Poole-Wilson, P. A., Mann, D. L., & Packer, M. (2006). The Seattle Heart Failure Model: prediction of survival in heart failure. *Circulation*, *113*(11), 1424–1433. https://doi.org/10.1161/CIRCULATIONAHA.105.584102

Lewsey, S. C., & Breathett, K. (2021). Racial and Ethnic Disparities in Heart Failure: Current State and Future Directions. *Current Opinion in Cardiology*, *36*(3), 320–328. https://doi.org/10.1097/HCO.0000000000000855

Li, K., Rollins, J., & Yan, E. (2018). Web of Science use in published research and review papers 1997–2017: a selective, dynamic, cross-domain, content-based analysis. *Scientometrics*, *115*(1), 1–20. https://doi.org/10.1007/s11192-017-2622-5

Lipton, Z. C. (2016). *The Mythos of Model Interpretability*.

Martins, B., Ferreira, D., Neto, C., Abelha, A., & Machado, J. (2021). Data Mining for Cardiovascular Disease Prediction. *Journal of Medical Systems*, *45*(1). https://doi.org/10.1007/s10916-020-01682-8

McDonagh, T. A., Metra, M., Adamo, M., Baumbach, A., Böhm, M., Burri, H., Čelutkiene, J., Chioncel, O., Cleland, J. G. F., Coats, A. J. S., Crespo-Leiro, M. G., Farmakis, D., Gardner, R. S., Gilard, M., Heymans, S., Hoes, A. W., Jaarsma, T., Jankowska, E. A., Lainscak, M., … Koskinas, K. C. (2021). 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. *European Heart Journal* 42(36), 3599–3726. https://doi.org/10.1093/eurheartj/ehab368

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., Antes, G., Atkins, D., Barbour, V., Barrowman, N., Berlin, J. A., Clark, J., Clarke, M., Cook, D., D'Amico, R., Deeks, J. J., Devereaux, P. J., Dickersin, K., Egger, M., Ernst, E., Gøtzsche, P. C., … Tugwell, P. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLOS Medicine*, *6*(7), e1000097. https://doi.org/10.1371/JOURNAL.PMED.1000097

Moreno-Sánchez, P. A. (2023). Improvement of a prediction model for heart failure survival through explainable artificial intelligence. *Frontiers in Cardiovascular Medicine*, *10*. https://doi.org/10.3389/fcvm.2023.1219586

Najafi-Vosough, R., Faradmal, J., Hosseini, S. K., Moghimbeigi, A., & Mahjub, H. (2021). Predicting hospital readmission in heart failure patients in Iran: A comparison of various machine learning

methods. *Healthcare Informatics Research*, *27*(4), 307–314. https://doi.org/10.4258/HIR.2021.27.4.307

Negassa, A., Ahmed, S., Zolty, R., & Patel, S. R. (2021). Prediction Model Using Machine Learning for Mortality in Patients with Heart Failure. *American Journal of Cardiology*, *153*, 86–93. https://doi.org/10.1016/j.amjcard.2021.05.044

Norhammar, A., Bodegard, J., Vanderheyden, M., Tangri, N., Karasik, A., Maggioni, A. Pietro, Sveen, K. A., Taveira-Gomes, T., Botana, M., Hunziker, L., Thuresson, M., Banerjee, A., Sundström, J., & Bollmann, A. (2023). Prevalence, outcomes and costs of a contemporary, multinational population with heart failure. *Heart*, *109*(7), 548–556. https://doi.org/10.1136/heartjnl-2022-321702

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *New England Journal of Medicine*, *375*(13), 1216–1219. https://doi.org/10.1056/nejmp1606181

Okoli, C., & Schabram, K. (2015). A Guide to Conducting a Systematic Literature Review of Information Systems Research. *Communications of the Association for Information Systems*, 37.

Okoye, C., Mazzarone, T., Niccolai, F., Bencivenga, L., Pescatore, G., Bianco, M. G., Guerrini, C., Giusti, A., Guarino, D., & Virdis, A. (2023). Predicting mortality and re-hospitalization for heart failure: a machine-learning and cluster analysis on frailty and comorbidity. *Aging Clinical and Experimental Research*, *35*(12), 2919–2928. https://doi.org/10.1007/s40520-023-02566-w

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., Mcdonald, S., … Mckenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*, *372*. https://doi.org/10.1136/BMJ.N160

Park, J., Hwang, I.-C., Yoon, Y. E., Park, J.-B., Park, J.-H., & Cho, G.-Y. (2022). Predicting Long-Term Mortality in Patients With Acute Heart Failure by Using Machine Learning. *Journal of Cardiac Failure*, *28*(7), 1078–1087. https://doi.org/10.1016/j.cardfail.2022.02.012

Paul, J., & Criado, A. R. (2020). The art of writing literature review: What do we know and what do we need to know? *International Business Review*, *29*(4). https://doi.org/10.1016/j.ibusrev.2020.101717

Paul, J., Khatri, P., & Kaur Duggal, H. (2023). Frameworks for developing impactful systematic literature reviews and theory building: What, Why and How? *Journal of Decision Systems*. https://doi.org/10.1080/12460125.2023.2197700

Paul, J., Lim, W. M., O'Cass, A., Hao, A. W., & Bresciani, S. (2021). Scientific procedures and rationales for systematic literature reviews (SPAR-4-SLR). *International Journal of Consumer Studies*. https://doi.org/10.1111/ijcs.12695

Paul, J., & Menzies, J. (2023). Developing classic systematic literature reviews to advance knowledge: Dos and don'ts. *European Management Journal*, 41(6), 815–820. https://doi.org/10.1016/j.emj.2023.11.006

Pocock, S. J., Ariti, C. A., McMurray, J. J. V., Maggioni, A., Køber, L., Squire, I. B., Swedberg, K., Dobson, J., Poppe, K. K., Whalley, G. A., & Doughty, R. N. (2013). Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. *European Heart Journal*, *34*(19), 1404–1413. https://doi.org/10.1093/EURHEARTJ/EHS337

Polo Friz, H., Esposito, V., Marano, G., Primitz, L., Bovio, A., Delgrossi, G., Bombelli, M., Grignaffini, G., Monza, G., & Boracchi, P. (2022). Machine learning and LACE index for predicting 30-day readmissions after heart failure hospitalization in elderly patients. *Internal and Emergency Medicine*. https://doi.org/10.1007/s11739-022-02996-w

Pranckutė, R. (2021). Web of Science (WoS) and Scopus: the titans of bibliographic information in today's academic world. *Publications*, 9(1). https://doi.org/10.3390/publications9010012

Quinlan, J. R. (1993). C4.5: Programs for machine learning. *Machine Learning*, *16*, 235-240.

Rahman, M. S., Rahman, H. R., Prithula, J., Chowdhury, M. E. H., Ahmed, M. U., Kumar, J., Murugappan, M., & Khan, M. S. (2023). Heart Failure Emergency Readmission Prediction Using Stacking Machine Learning Model. *Diagnostics*, *13*(11). https://doi.org/10.3390/diagnostics13111948

Rizinde, T., Ngaruye, I., & Cahill, N. D. (2023). Comparing Machine Learning Classifiers for Predicting Hospital Readmission of Heart Failure Patients in Rwanda. *Journal of Personalized Medicine*, *13*(9). https://doi.org/10.3390/jpm13091393

Sabouri, M., Rajabi, A. B., Hajianfar, G., Gharibi, O., Mohebi, M., Avval, A. H., Naderi, N., & Shiri, I. (2023). Machine learning based readmission and mortality prediction in heart failure patients. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-45925-3

Sarijaloo, F. B., Park, J., Zhong, X., & Wokhlu, A. (2021). Predicting 90 day acute heart failure readmission and death using machine learning-supported decision analysis. *Clinical Cardiology*, *44*(2), 230–237. https://doi.org/10.1002/clc.23532

Savarese, G., Becher, P. M., Lund, L. H., Seferovic, P., Rosano, G. M. C., & Coats, A. J. S. (2023). Global burden of heart failure: a comprehensive and updated review of epidemiology. *Cardiovascular research*, 118(17), 3272–3287. https://doi.org/10.1093/cvr/cvac013

Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law*, *7*(4).

Shameer, K., Johnson, K. W., Glicksberg, B. S., Dudley, J. T., & Sengupta, P. P. (2018). Machine learning in cardiovascular medicine: Are we there yet? *Heart*, 104(14), 1156–1164. https://doi.org/10.1136/heartjnl-2017-311198

Sharma, V., Kulkarni, V., Mcalister, F., Eurich, D. E. A. N., Keshwani, S., Simpson, S. H., Voaklander, D. O. N., & Samanani, S. (2022). Predicting 30-Day Readmissions in Patients With Heart Failure

Using Administrative Data: A Machine Learning Approach. *Journal of Cardiac Failure*, *28*(5), 710–722. https://doi.org/10.1016/j.cardfail.2021.12.004

Sohrabi, B., Vanani, I. R., Gooyavar, A., & Naderi, N. (2019). Predicting the Readmission of Heart Failure Patients through Data Analytics. *Journal of Information and Knowledge Management*, *18*(1). https://doi.org/10.1142/S0219649219500126

Soliman, A., Agvall, B., Etminani, K., Hamed, O., & Lingman, M. (2023). The Price of Explainability in Machine Learning Models for 100-Day Readmission Prediction in Heart Failure: Retrospective, Comparative, Machine Learning Study. *Journal of Medical Internet Research*, *25*. https://doi.org/10.2196/46934

Spinar, J., Jarkovsky, J., Spinarova, L., Mebazaa, A., Gayat, E., Vitovec, J., Linhart, A., Widimsky, P., Miklik, R., Zeman, K., Belohlavek, J., Malek, F., Felsoci, M., Kettner, J., Ostadal, P., Cihalik, C., Vaclavik, J., Taborsky, M., Dusek, L., … Parenica, J. (2016). AHEAD score--Long-term risk classification in acute heart failure. *International Journal of Cardiology*, *202*, 21–26. https://doi.org/10.1016/J.IJCARD.2015.08.187

Tasci, E., Zhuge, Y., Camphausen, K., & Krauze, A. V. (2022). Bias and Class Imbalance in Oncologic Data—Towards Inclusive and Transferrable AI in Large Scale Oncology Data Sets. *Cancers*, *14*(12). https://doi.org/10.3390/CANCERS14122897

Tohyama, T., Ide, T., Ikeda, M., Kaku, H., Enzan, N., Matsushima, S., Funakoshi, K., Kishimoto, J., Todaka, K., & Tsutsui, H. (2021). Machine learning-based model for predicting 1 year mortality of hospitalized patients with heart failure. *ESC Heart Failure*, *8*(5), 4077–4085. https://doi.org/10.1002/ehf2.13556

Tong, R., Zhu, Z., & Ling, J. (2023). Comparison of linear and non-linear machine learning models for time-dependent readmission or mortality prediction among hospitalized heart failure patients. *Heliyon*, *9*(5). https://doi.org/10.1016/j.heliyon.2023.e16068

van der Galiën, O. P., Hoekstra, R. C., Gürgöze, M. T., Manintveld, O. C., van den Bunt, M. R., Veenman, C. J., & Boersma, E. (2021). Prediction of long-term hospitalisation and all-cause mortality in patients with chronic heart failure on Dutch claims data: a machine learning approach. *BMC Medical Informatics and Decision Making*, *21*(1). https://doi.org/10.1186/s12911-021-01657-w

Van Walraven, C., Dhalla, I. A., Bell, C., Etchells, E., Stiell, I. G., Zarnke, K., Austin, P. C., & Forster, A. J. (2010). Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *CMAJ : Canadian Medical Association Journal*, *182*(6), 551–557. https://doi.org/10.1503/CMAJ.091117

Voors, A. A., Ouwerkerk, W., Zannad, F., van Veldhuisen, D. J., Samani, N. J., Ponikowski, P., Ng, L. L., Metra, M., ter Maaten, J. M., Lang, C. C., Hillege, H. L., van der Harst, P., Filippatos, G., Dickstein, K., Cleland, J. G., Anker, S. D., & Zwinderman, A. H. (2017). Development and

validation of multivariable models to predict mortality and hospitalization in patients with heart failure. *European Journal of Heart Failure*, *19*(5), 627–634. https://doi.org/10.1002/EJHF.785

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1–3), 37–52.

Xu, C., Li, H., Yang, J., Peng, Y., Cai, H., Zhou, J., Gu, W., & Chen, L. (2023). Interpretable prediction of 3-year all-cause mortality in patients with chronic heart failure based on machine learning. *BMC Medical Informatics and Decision Making*, *23*(1). https://doi.org/10.1186/s12911-023-02371-5

Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*, *1168*(2). https://doi.org/10.1088/1742-6596/1168/2/022022

Zhang, D., Yin, C., Zeng, J., Yuan, X., & Zhang, P. (2020). Combining structured and unstructured data for predictive models: a deep learning approach. *BMC Medical Informatics and Decision Making*, *20*(1). https://doi.org/10.1186/s12911-020-01297-6

Appendix

A: Most important variables from the SLR

Most important variables: Demographics

Demographics	Study ID																			Total
	1	2	3	4	5	6	7	8	9	12	13	14	15	16	18	19	20	21	22	23	
Age	✓		✓		✓	✓	✓					✓			✓	✓	✓	✓		✓	11
Sex			✓													✓		✓			3
Weight								✓													1
Substance abuse															✓						1
Alcohol intake			✓																		1
Occupaction		✓																			1
Distric of Residence			✓																		1
Total population																			✓		1
Distance from home to a park																			✓		1
Admission Type		✓							✓											✓	3
LOS			✓		✓				✓							✓					4
LOB																		✓			1
HF duration					✓																1
Discharged to home																			✓		1
Nº Hospitalizations					✓	✓				✓	✓										4
Nº of allied health visits (last 6 months)																				✓	1
Nº VGP																		✓			1
History of HF																				✓	1
Last Hospitalization																✓					1
Nº Hospitalizations for CHF/AHF																		✓			1
Nº ES visits (last 6 months)					✓																1
Time since last non-cardiovascular admission													✓								1
Long-term care residence status													✓								1
AHF														✓							1
BMI																✓					1
Type of HF		✓							✓												2
CFS							✓														1
CCI		✓			✓		✓														3
PNI																	✓				1
NYHA						✓			✓												2
LACE score		✓																			1
BI																✓					1

Note: AHF: Acute Heart Failure; BI: Barthel Index; BMI: Body Mass Index; CCI: Charlson Comorbidity Index; CFS: Clinical Frailty Scale; CHF: Chronic Heart Failure; ED: Emergency Department; HF: Heart Failure; LACE: Length of Stay, Acuity of Admission, Comorbidity, Emergency Department Use; LOB: Length of Stay Beyond; LOS: Length of Stay; NYHA: New York Heart Association Functional Classification; PNI: Prognostic Nutritional Index; VGP: Visits to the General Practitioner.

Most important variables: Lab Results

Lab Results	Study ID																			Total
	1	2	3	4	5	6	7	8	9	12	13	14	15	16	18	19	20	21	22	23	
Albumin														✓							1
Basophil count									✓												1
BNP						✓	✓														2
BUN															✓				✓		2
CK-MB															✓						1
Cholesterol				✓																	1
Cr	✓						✓								✓				✓		4
DD									✓												1
GFR						✓						✓									2
Hb	✓	✓								✓									✓		4
HCO3														✓							1
Htc									✓										✓		2
LFT	✓																				1
MCV									✓												1
NLR		✓															✓				2
NT-proBNP					✓							✓		✓							3
Platelet									✓												1
PTT															✓						1
RBC				✓															✓		2
Na												✓	✓	✓	✓						4
Uric acid	✓			✓					✓												3

Notes: BNP: B-type Natriuretic Peptide; BUN: Blood Urea Nitrogen; CK-MB: Creatine Kinase-MB; Cr: Creatinine; DD: D-Dimer; GFR: Glomerular Filtration Rate; Hb: Hemoglobin; HCO3: Bicarbonate; Htc: Hematocrit; LFT: Liver Function Test; MCV: Mean Corpuscular Volume; Na: Sodium; NLR: Neutrophil to Lymphocyte Ratio; NTproBNP: N-terminal pro–B-type Natriuretic Peptide; PTT: Partial Thromboplastin Time; RBC: Red Blood Cell Count.

Most important variables: Vital Signs; Symptoms and Comorbidities

	Study ID																				Total
	1	2	3	4	5	6	7	8	9	12	13	14	15	16	18	19	20	21	22	23	
Vital signs																					-
HR			✓					✓			✓			✓							4
DBP			✓			✓						✓		✓							4
SBP		✓	✓			✓								✓							4
Symptoms																					
Well-being								✓													
Edema		✓						✓													2
Orthopnea								✓													
Shortness of breath			✓																		1
Comorbidities																					-
Arrhythmia			✓																		1
COPD																		✓			1
DM										✓								✓			2
CKD		✓			✓											✓				✓	4
Depression																				✓	1
Lung Disease														✓							1
Other CVD														✓							1
WRF		✓																			1

Notes: CKD: Chronic Kidney Disease; COPD: Chronic Obstructive Pulmonary Disease; CVD: Cardiovascular Disease; DBP: Diastolic Blood Pressure; DM: Diabetes Mellitus; HR: Heart Rate; SBP: Systolic Blood Pressure; WRF: Worsening Renal Function

Most important variables: Treatments; TTE/ECG

	Study ID																			Total
	1	2	3	4	5	6	7	8	9	12	13	14	15	16	18	19	20	21	22	23	
Treatments																					-
Inotropic Support	✓													✓							2
Dialysis	✓																				1
Diuretic use														✓		✓			✓		3
at least 2 supplies of antineoplastic and immunomodulating agents in the last 6 months																				✓	1
TTE/ECG																					-
R-wave axis																			✓		1
QRS duration																			✓		1
QTc interval																			✓		1
Atrial rate																			✓		1
LVEDD				✓																	1
LVEF														✓	✓						2
RV Dysfunction	✓																				1
TR	✓																				1
LV GLS												✓									1
E/e' Ratio												✓									1
IVC diameter														✓							1
Mitral Valve SEM		✓																			1

Notes: E/e' Ratio: Ratio of early mitral inflow velocity to early diastolic mitral annular velocity; IVC: Inferior Vena Cava; LV GLS: Left Ventricular Global Longitudinal Strain; LVEDD: Left Ventricular End-Diastolic Diameter; LVEF: Left Ventricular Ejection Fraction; QTc interval: Corrected QT Interval; R-wave axis: Axis of the R-wave; RV: Right Ventricle; SEM: Systolic Ejection Murmur; TTE/ECG: Electrocardiogram/Echocardiogram; TR: Tricuspid Regurgitation.

B: Annual Distribution of HF Diagnoses by ICD-10 Code

ICD-10 HF Codes	2020	2021	2022	2023	2024	Total
I50 - HF	159	154	127	95	28	563
I50.1 - Left HF	149	470	751	901	224	2 495
I50.20 - Systolic (congestive) HF, unspecified	4	27	47	43	10	131
I50.22 - Chronic congestive (systolic) HF	5	7	25	26	17	80
I50.23 - Chronic systolic (congestive) HF, exacerbated				1	3	4
I50.30 - Diastolic (congestive) HF, unspecified		1	7	16	2	26
I50.32 - Chronic congestive diastolic HF	2	2	6	7	1	18
I50.33 - Chronic diastolic (congestive) HF, exacerbated				1		1
I50.40 - Combined systolic (congestive) and diastolic (congestive) HF, unspecified			2	1	1	4
I50.810 - Right HF, unspecified					1	1
I50.812 - Chronic right HF				6	1	7
I50.82 - Biventricular HF			1			1
I50.89 - Other HF			10	42	25	77
I50.9 - HF, unspecified	6	8	22	36	13	85
Total	**325**	**669**	**998**	**1 175**	**326**	**3 493**

C: Past-Related Variables Created from Consultation Records

- Variables indicating the presence of symptoms, signs, or treatment prescriptions at the last consultation, such as Last_JVD, Last_HJR, and Last_SOB, are coded as 1 if present and 0 otherwise.
- Variables showing variation from the last consultation in symptoms and signs, such as Var1_SOB and Var1_PND, represent changes: 0 indicates no change, -1 indicates worsening, and 1 indicates improvement. For Var_Symptoms, values >0 indicate worsening, <0 indicate improvement, and 0 indicates no change.

D: Derived Laboratory Variables

From each of the 52 original laboratory measurements, 11 additional variables were created, yielding a total of 572 new variables. An additional 7 binary indicators were generated specifically for NT-proBNP and creatinine.

- Sequential Lab Values: Variables like Last1_Hb, Last2_Na, and Last3_Cr capture lab results from the last three consultations, allowing for trend analysis across recent measurements.
- Statistical Summaries: For each original lab variable, five summary statistics were calculated. For example, Mx_Na, Mn_Hb, and AVG_Cr represent maximum, minimum, and average values, respectively, over the recorded period; SD_WBC and CV_K capture the standard deviation and coefficient of variation, reflecting variability in results.
- Binary Indicators: Seven binary variables act as clinical flags for threshold levels in NT-proBNP and creatinine. For instance, Var1_30_NTproBNP is coded as 1 if NT-proBNP reduction from the previous lab result is below 30%, and 0 otherwise. Similarly, Var2_0.3_Cr is coded as 1 if creatinine increases by 0.3 mg/dL compared to the second-to-last measurement.

E: Exploratory analysis

Due to the number of variables, the presented correlation matrix includes only those variables identified as relevant in the bivariate analysis or considered clinically significant by experts.

Correlation matrix of variables from the dimension Patient profile (Part 1)

Correlations

		Sex	Autonomy	Smoker Status	CDI	CRT-P
Sex	Pearson Correlation	1	-,151**	-,294**	-,229**	-,075**
	Sig. (2-tailed)		<,001	<,001	<,001	<,001
	N	2008	1432	2008	2008	2008
Autonomy	Pearson Correlation	-,151**	1	,138**	,085**	-,016
	Sig. (2-tailed)	<,001		<,001	,001	,543
	N	1432	1432	1432	1432	1432
Smoker Status	Pearson Correlation	-,294**	,138**	1	,139**	,014
	Sig. (2-tailed)	<,001	<,001		<,001	,522
	N	2008	1432	2008	2008	2008
CDI	Pearson Correlation	-,229**	,085**	,139**	1	-,058**
	Sig. (2-tailed)	<,001	,001	<,001		,009
	N	2008	1432	2008	2008	2008
CRT-P	Pearson Correlation	-,075**	-,016	,014	-,058**	1
	Sig. (2-tailed)	<,001	,543	,522	,009	
	N	2008	1432	2008	2008	2008

**. Correlation is significant at the 0.01 level (2-tailed).

Correlation matrix of variables from the dimension Patient profile (Part 2)

Correlations

		ES_2020_Now	ES_6m	DH_6m	DH_2020_Now	Hosp_2020_Now	Hosp_6m	Previsously_Hospitalized_12m	Avg_Time_Hosp	LOS_Last
ES_2020_Now	Pearson Correlation	1	,510**	,098**	,087**	,627**	,256**	,230**	,084**	,108**
	Sig. (2-tailed)		<,001	<,001	<,001	<,001	<,001	<,001	<,001	<,001
	N	2008	2008	2008	2008	2008	2008	2008	2008	2008
ES_6m	Pearson Correlation	,510**	1	,048*	-,035	,326**	,647**	,399**	,121**	,109**
	Sig. (2-tailed)	<,001		,032	,117	<,001	<,001	<,001	<,001	<,001
	N	2008	2008	2008	2008	2008	2008	2008	2008	2008
DH_6m	Pearson Correlation	,098**	,048*	1	,667**	,034	-,005	-,001	-,002	,005
	Sig. (2-tailed)	<,001	,032		<,001	,127	,840	,978	,941	,825
	N	2008	2008	2008	2008	2008	2008	2008	2008	2008
DH_2020_Now	Pearson Correlation	,087**	-,035	,667**	1	-,006	-,063**	-,071**	-,059**	-,051*
	Sig. (2-tailed)	<,001	,117	<,001		,781	,004	,001	,008	,022
	N	2008	2008	2008	2008	2008	2008	2008	2008	2008
Hosp_2020_Now	Pearson Correlation	,627**	,326**	,034	-,006	1	,483**	,480**	,222**	,217**
	Sig. (2-tailed)	<,001	<,001	,127	,781		<,001	<,001	<,001	<,001
	N	2008	2008	2008	2008	2008	2008	2008	2008	2008
Hosp_6m	Pearson Correlation	,256**	,647**	-,005	-,063**	,483**	1	,632**	,223**	,209**
	Sig. (2-tailed)	<,001	<,001	,840	,004	<,001		<,001	<,001	<,001
	N	2008	2008	2008	2008	2008	2008	2008	2008	2008
Previsously_Hospitalized_12m	Pearson Correlation	,230**	,399**	-,001	-,071**	,480**	,632**	1	,290**	,268**
	Sig. (2-tailed)	<,001	<,001	,978	,001	<,001	<,001		<,001	<,001
	N	2008	2008	2008	2008	2008	2008	2008	2008	2008
Avg_Time_Hosp	Pearson Correlation	,084**	,121**	-,002	-,059**	,222**	,223**	,290**	1	,951**
	Sig. (2-tailed)	<,001	<,001	,941	,008	<,001	<,001	<,001		<,001
	N	2008	2008	2008	2008	2008	2008	2008	2008	2008
LOS_Last	Pearson Correlation	,108**	,109**	,005	-,051*	,217**	,209**	,268**	,951**	1
	Sig. (2-tailed)	<,001	<,001	,825	,022	<,001	<,001	<,001	<,001	
	N	2008	2008	2008	2008	2008	2008	2008	2008	2008

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Correlation matrix of variables from the dimension Lab Results

Correlations

		Hb	RBC	Hct	WBC	NT-proBNP	Na	K	Urea	Cr	GFR_calc
Hb	Pearson Correlation	1	,834**	,967**	,157**	-,241**	,137**	,042	-,255**	-,209**	,307**
	Sig. (2-tailed)		<,001	<,001	<,001	<,001	<,001	,073	<,001	<,001	<,001
	N	1902	1902	1902	1902	1232	1829	1785	1777	1853	1853
RBC	Pearson Correlation	,834**	1	,886**	,191**	-,220**	,127**	,020	-,246**	-,208**	,299**
	Sig. (2-tailed)	<,001		<,001	<,001	<,001	<,001	,397	<,001	<,001	<,001
	N	1902	1902	1902	1902	1232	1829	1785	1777	1853	1853
Hct	Pearson Correlation	,967**	,886**	1	,177**	-,220**	,181**	,076**	-,252**	-,200**	,285**
	Sig. (2-tailed)	<,001	<,001		<,001	<,001	<,001	,001	<,001	<,001	<,001
	N	1902	1902	1902	1902	1232	1829	1785	1777	1853	1853
WBC	Pearson Correlation	,157**	,191**	,177**	1	-,038	-,054*	,082**	,026	,055*	-,013
	Sig. (2-tailed)	<,001	<,001	<,001		,182	,020	<,001	,282	,019	,587
	N	1902	1902	1902	1902	1232	1829	1785	1777	1853	1853
NT-proBNP	Pearson Correlation	-,241**	-,220**	-,220**	-,038	1	-,087**	-,105**	,311**	,403**	-,305**
	Sig. (2-tailed)	<,001	<,001	<,001	,182		,002	<,001	<,001	<,001	<,001
	N	1232	1232	1232	1232	1266	1238	1224	1199	1246	1246
Na	Pearson Correlation	,137**	,127**	,181**	-,054*	-,087**	1	-,056*	-,164**	-,076**	,047*
	Sig. (2-tailed)	<,001	<,001	<,001	,020	,002		,016	<,001	,001	,040
	N	1829	1829	1829	1829	1238	1869	1823	1785	1862	1862
K	Pearson Correlation	,042	,020	,076**	,082**	-,105**	-,056*	1	,041	,076**	-,105**
	Sig. (2-tailed)	,073	,397	,001	<,001	<,001	,016		,085	,001	<,001
	N	1785	1785	1785	1785	1224	1823	1834	1741	1827	1827
Urea	Pearson Correlation	-,255**	-,246**	-,252**	,026	,311**	-,164**	,041	1	,655**	-,700**
	Sig. (2-tailed)	<,001	<,001	<,001	,282	<,001	<,001	,085		<,001	<,001
	N	1777	1777	1777	1777	1199	1785	1741	1815	1812	1812
Cr	Pearson Correlation	-,209**	-,208**	-,200**	,055*	,403**	-,076**	,076**	,655**	1	-,754**
	Sig. (2-tailed)	<,001	<,001	<,001	,019	<,001	,001	,001	<,001		<,001
	N	1853	1853	1853	1853	1246	1862	1827	1812	1906	1906
GFR_calc	Pearson Correlation	,307**	,299**	,285**	-,013	-,305**	,047*	-,105**	-,700**	-,754**	1
	Sig. (2-tailed)	<,001	<,001	<,001	,587	<,001	,040	<,001	<,001	<,001	
	N	1853	1853	1853	1853	1246	1862	1827	1812	1906	1906

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Correlation matrix of variables from the dimension Clinical Presentation

Correlations

		JVD	PND	Orthopnea	NYHA_Cod	Symptoms
JVD	Pearson Correlation	1	,060**	,089**	,163**	,345**
	Sig. (2-tailed)		,007	<,001	<,001	<,001
	N	2008	2008	2008	2008	2008
PND	Pearson Correlation	,060**	1	,472**	,116**	,593**
	Sig. (2-tailed)	,007		<,001	<,001	<,001
	N	2008	2008	2008	2008	2008
Orthopnea	Pearson Correlation	,089**	,472**	1	,212**	,718**
	Sig. (2-tailed)	<,001	<,001		<,001	<,001
	N	2008	2008	2008	2008	2008
NYHA_Cod	Pearson Correlation	,163**	,116**	,212**	1	,309**
	Sig. (2-tailed)	<,001	<,001	<,001		<,001
	N	2008	2008	2008	2008	2008
Symptoms	Pearson Correlation	,345**	,593**	,718**	,309**	1
	Sig. (2-tailed)	<,001	<,001	<,001	<,001	
	N	2008	2008	2008	2008	2008

**. Correlation is significant at the 0.01 level (2-tailed).

Variance Inflation Factor of significant variables for the LR

Coefficients

Model		Collinearity Statistics	
		Tolerance	VIF
1	Consult_6m	,569	1,757
	Sex	,675	1,480
	Autonomy	,827	1,209
	Smoker	,763	1,310
	JVD	,635	1,574
	PND	,663	1,509
	Orthopnea	,499	2,003
	Symptoms	,337	2,971
	NYHA_Cod	,670	1,493
	ES_2020_Now	,337	2,971
	ES_6m	,536	1,864
	DH_6m	,321	3,111
	DH_2020_Now	,297	3,372
	Hosp_2020_Now	,418	2,394
	Hb	,704	1,421
	WBC	,854	1,171
	NTproBNP	,696	1,436
	Na	,852	1,174
	Var3_30_NTproBNP	,898	1,113

F: Outputs of the PCA with the best results

Correlation Matrix

		Hb	RBC	Hct	Urea	Cr	GFR_calc
Correlation	Hb	1,000	,834	,967	-,255	-,209	,307
	RBC	,834	1,000	,886	-,246	-,208	,299
	Hct	,967	,886	1,000	-,252	-,200	,285
	Urea	-,255	-,246	-,252	1,000	,655	-,700
	Cr	-,209	-,208	-,200	,655	1,000	-,754
	GFR_calc	,307	,299	,285	-,700	-,754	1,000

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,717
Bartlett's Test of Sphericity	Approx. Chi-Square	10718,619
	df	15
	Sig.	<,001

Communalities

	Initial	Extraction
Hb	1,000	,939
RBC	1,000	,880
Hct	1,000	,973
Urea	1,000	,762
Cr	1,000	,812
GFR_calc	1,000	,839

Extraction Method: Principal Component Analysis.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,380	56,332	56,332	3,380	56,332	56,332	2,791	46,519	46,519
2	1,825	30,418	86,749	1,825	30,418	86,749	2,414	40,231	86,749
3	,354	5,892	92,641						
4	,235	3,916	96,558						
5	,181	3,016	99,574						
6	,026	,426	100,000						

Extraction Method: Principal Component Analysis.

Rotated Component Matrix[a]

	Component	
	1	2
Hct	,978	-,130
Hb	,958	-,146
RBC	,927	-,145
Cr	-,077	,898
GFR_calc	,183	-,897
Urea	-,138	,862

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Component Matrix[a]

	Component	
	1	2
Hct	,851	,499
Hb	,845	,474
RBC	,820	,456
GFR_calc	,696	-,595
Urea	-,639	,595
Cr	-,613	,660

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

G: Logistic Regression Significant Variable Selection by Dimension and Overall Model

Dimension	Lab Results				Patient Profile				Clinical Profile			
Selection Method	Enter		Forward Stepwise		Enter		Forward Stepwise		Enter		Forward Stepwise	
Independent Variables	Exp(B)	P-value	Exp(B)	P-value	Exp(B)	P-value	Exp(B)	P-value	Exp(B)	P-value	Exp(B)	P-value
Hb	0,849	<0,001	0,84	<0,001	-	-	-	-	-	-	-	-
NTproBNP_scale	1,032	0,004	1,032	0,003	-	-	-	-	-	-	-	-
Var3_30_NTproBNP	1,403	0,081	1,446	0,054	-	-	-	-	-	-	-	-
Var1_0.3_Cr (Increase>= 0.3)	1,66	0,065	1,737	0,041	-	-	-	-	-	-	-	-
WBC	0,78	0,99	-	-	-	-	-	-	-	-	-	-
Na	0,109	0,963	-	-	-	-	-	-	-	-	-	-
Sex (Female)	-	-	-	-	1,273	0,16	1,422	0,022	-	-	-	-
ICD (Yes)	-	-	-	-	1,807	0,01	1,672	0,022	-	-	-	-
ES_6M	-	-	-	-	1,257	<0,001	1,357	<0,001	-	-	-	-
DH_6M	-	-	-	-	1,168	0,006	1,1149	0,011	-	-	-	-
DH_2020_Now	-	-	-	-	1,045	0,007	1,049	0,004	-	-	-	-
Previsouly_Hosp_12m (Yes)	-	-	-	-	0,66	0,053	0,714	0,047	-	-	-	-
LOS_Last	-	-	-	-	1,02	0,019	1,02	0,016	-	-	-	-
CRT-P (Yes)	-	-	-	-	0,701	0,245	-	-	-	-	-	-
Consult_6m	-	-	-	-	0,932	0,345	-	-	-	-	-	-
Autonomy (Independent)	-	-	-	-	0,7	0,151	-	-	-	-	-	-
Smoker status (Smoker)	-	-	-	-	0,803	0,366	-	-	-	-	-	-
Smoker status (Ex-smoker)	-	-	-	-	0,778	0,245	-	-	-	-	-	-
ES_2020_Now	-	-	-	-	1,025	0,339	-	-	-	-	-	-
Hosp_2020_Now	-	-	-	-	0,874	0,123	-	-	-	-	-	-
Hosp_6m	-	-	-	-	1,37	0,066	-	-	-	-	-	-
JVD (Yes)	-	-	-	-	-	-	-	-	2,142	0,111	2,376	0,032
NYHA_1	-	-	-	-	-	-	-	-	0,617	0,002	0,666	0,004
NYHA_2	-	-	-	-	-	-	-	-	0,818	0,174	-	-
NYHA_3	-	-	-	-	-	-	-	-	0,953	0,838	-	-
PND (Yes)	-	-	-	-	-	-	-	-	1,411	0,409	1,763	0,071
Orthopnea (Yes)	-	-	-	-	-	-	-	-	1,241	0,519	-	-
Symptpms	-	-	-	-	-	-	-	-	1,036	0,838	-	-
Nagelkerke R^2	0,066		0,062		0,156		0,146		0,017		0,015	

LR Model Specifications: Including and Excluding Historical Data

LR	LR with Historical Data				LR without Historical Data			
Selection Method	Enter		Forward Stepwise		Enter		Forward Stepwise	
Independent Variables	Exp(B)	P-value	Exp(B)	P-value	Exp(B)	P-value	Exp(B)	P-value
Hb	0,87	0,004	0,863	<0,001	0,84	<0,001	0,852	<0,001
NTproBNP_scale	1,01	0,443	-	-	1,037	0,041	1,035	0,002
Var3_30_NTproBNP	1,282	0,216	-	-	-	-	-	-
Var1_0.3_Cr (Increase >= 0.3)	1,993	0,015	1,847	0,026	-	-	-	-
Sex (Female)	1,304	0,142	-	-	1,446	0,03	1,415	0,038
ICD (Yes)	1,731	0,028	1,642	0,039	1,766	0,018	1,769	0,017
ES_6M	1,298	<0,001	1,3	<0,001	-	-	-	-
DH_6M	1,067	0,418	-	-	-	-	-	-
DH_2020_Now	1,101	0,002	1,128	<0,001	-	-	-	-
Previsouly_Hosp_12m (Yes)	0,875	0,476	-	-	-	-	-	-
LOS_Last	1,004	0,692	-	-	-	-	-	-
JVD (Yes)	1,498	0,444	-	-	1,586	0,364	-	-
NYHA1	0,647	0,039	0,6	0,012	0,671	0,041	0,683	0,047
PND (Yes)	1,889	0,135	1,982	0,098	1,605	0,248	-	-
Age	1,001	0,913	-	-	0,997	0,627	-	-
DM (Yes)	1,029	0,868	-	-	1,066	0,694	-	-
GRF	1,004	0,346	-	-	1,003	0,394	-	-
Nagelkerke R^2	0,176		0,168		0,074		0,069	

H: Results of the Best Models for de Divided Samples

Results of the Best Predictive Models For HF Decompensation For The Sample With Null

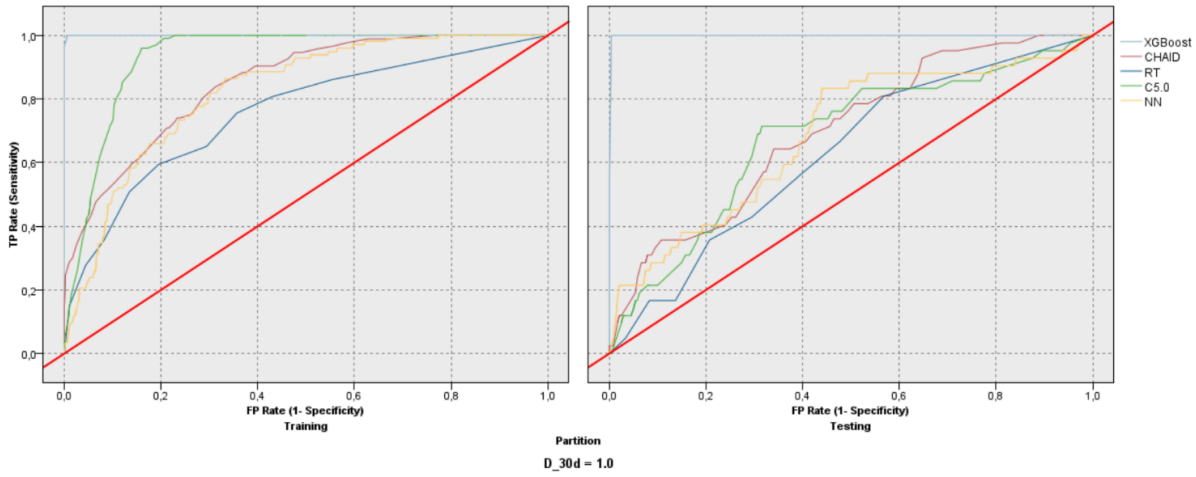
Metrics		Sample with Nulls					
		XGBoost	CHAID	RT	C5.0	NN	RT (Auto)
Training	S	-	87.11%	77.78%	80.00%	84.30%	97,13%
	R	-	91.43%	68.60%	94.60%	58.04%	84,07%
	A	-	89.20%	73.38%	87.05%	71.58%	90,22%
	F1	-	89.10%	71.18%	87.58%	66.43%	90,09%
	P	-	86.88%	73.96%	81.53%	77.64%	97,04%
	AUC	-	0.953	0.811	0.941	0.759	0,955
Test	S	-	80.36%	76.79%	75.71%	80.36%	97,17%
	R	-	57.35%	54.41%	51.47%	54.41%	60,00%
	A	-	75.86%	72.41%	70.98%	75.29%	91,27%
	F1	-	48.15%	43.53%	40.94%	46.25%	68,57%
	P	-	41.49%	36.27%	33.98%	40.22%	82,61%
	AUC	-	0.714	0.675	0.643	0.729	0,907

Results of the Best Predictive Models For HF Decompensation (Sample Without Null)

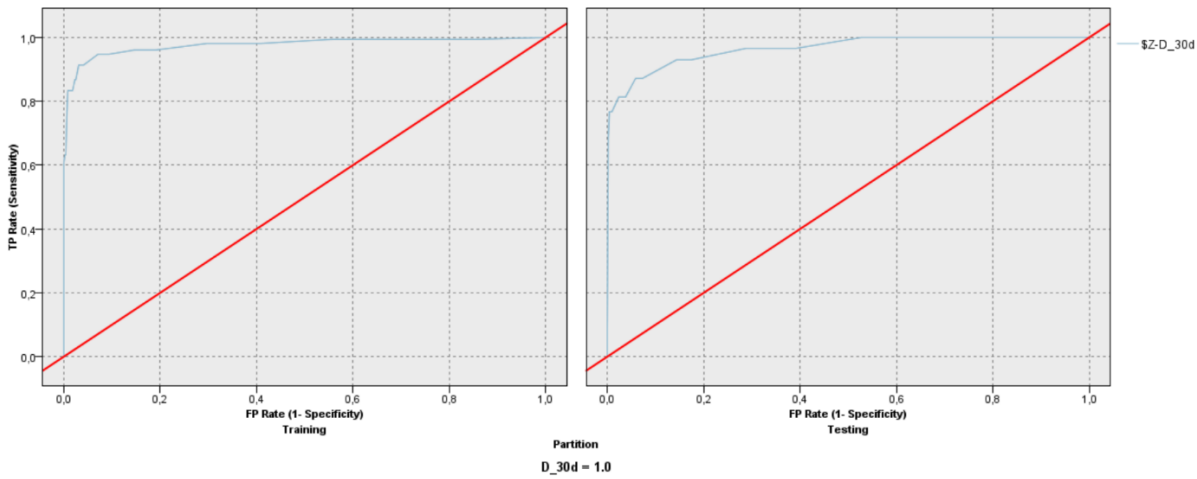
Metrics		Sample without Nulls					
		XGBoost	CHAID	RT	C5.0	NN	RT (Auto)
Training	S	96.95%	83.16%	74.04%	71.60%	81.95%	94.12%
	R	100.00%	99.07%	92.90%	100.00%	91.30%	100.00%
	A	98.49%	91.45%	83.85%	86.45%	86.79%	97.17%
	F1	98.52%	92.35%	85.69%	88.52%	87.74%	97.35%
	P	97.08%	86.48%	79.52%	79.41%	84.44%	94.83%
	AUC	1.000	0.976	0.924	0.912	0.910	0.996
Test	S	93.71%	68.98%	66.67%	63.43%	69.44%	96.30%
	R	100.00%	55.56%	55.56%	52.78%	58.33%	75.00%
	A	94.63%	67.06%	65.08%	61.90%	67.86%	93.25%
	F1	84.51%	32.52%	31.25%	28.36%	34.15%	76.06%
	P	73.17%	22.99%	21.74%	19.39%	24.14%	77.14%
	AUC	1.000	0.622	0.636	0.572	0.669	0.954

I: ROC Curves

ROC Curve of the best models by algorithm (Complete Sample)

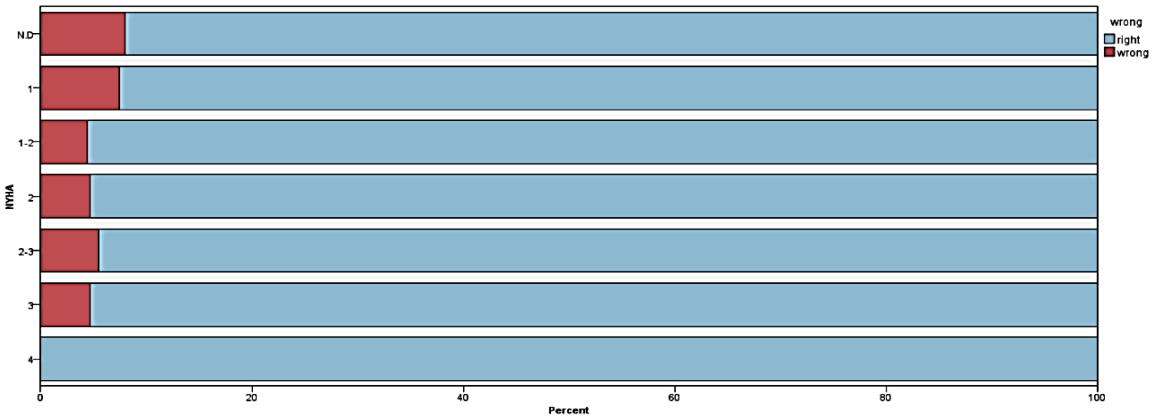


ROC Curve of the combined model

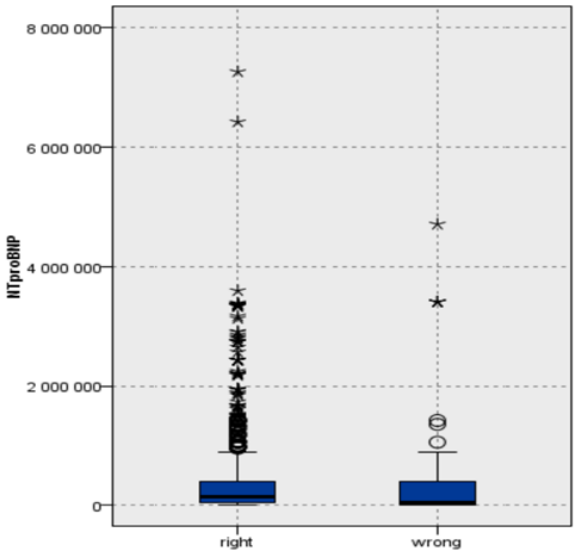


J: Error Analysis for Model Improvement

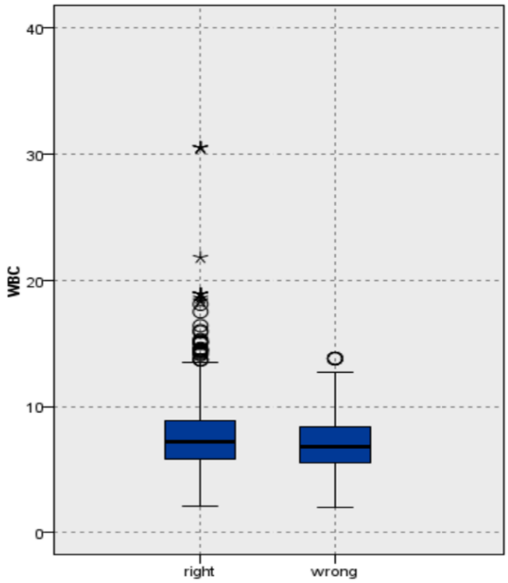
Error Distribution by NYHA



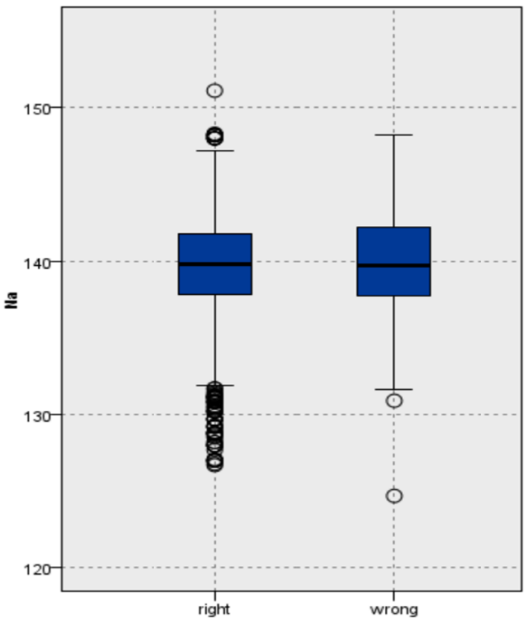
NT-proBNP-distribution by type of decision



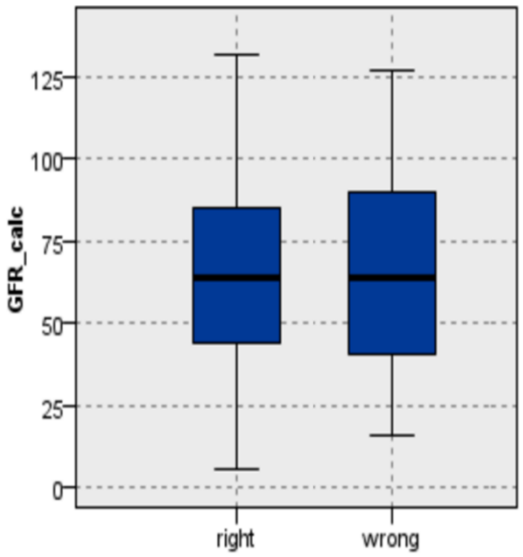
WBC-distribution by type of decision



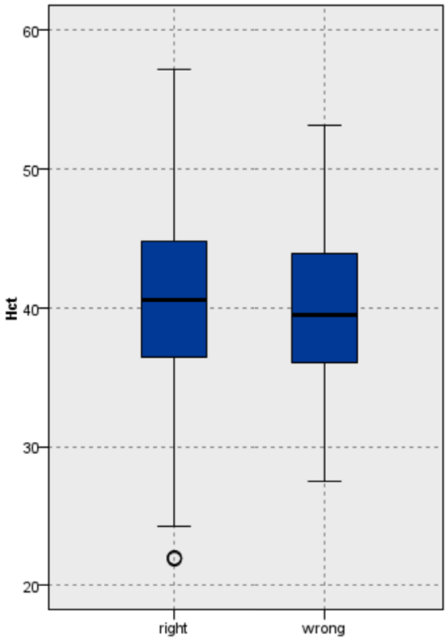
Na-distribution by type of decision



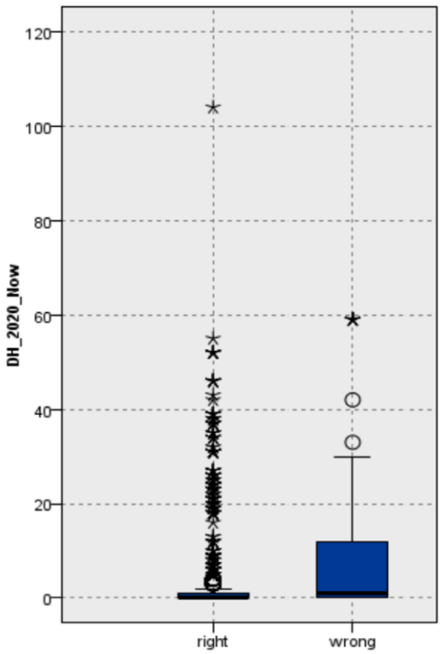
GFR_calc-distribution by type of decision



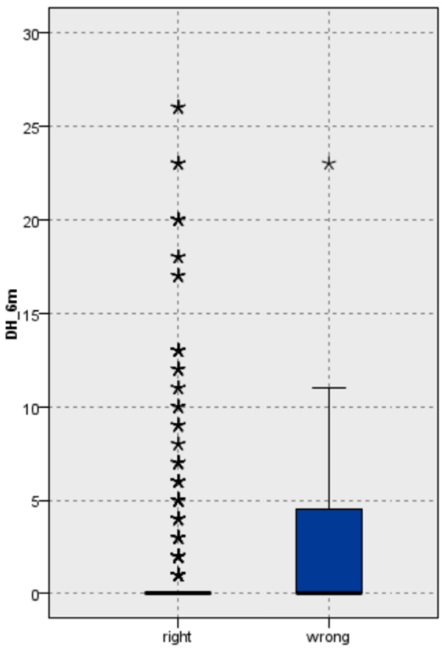
Hct-distribution by type of decision



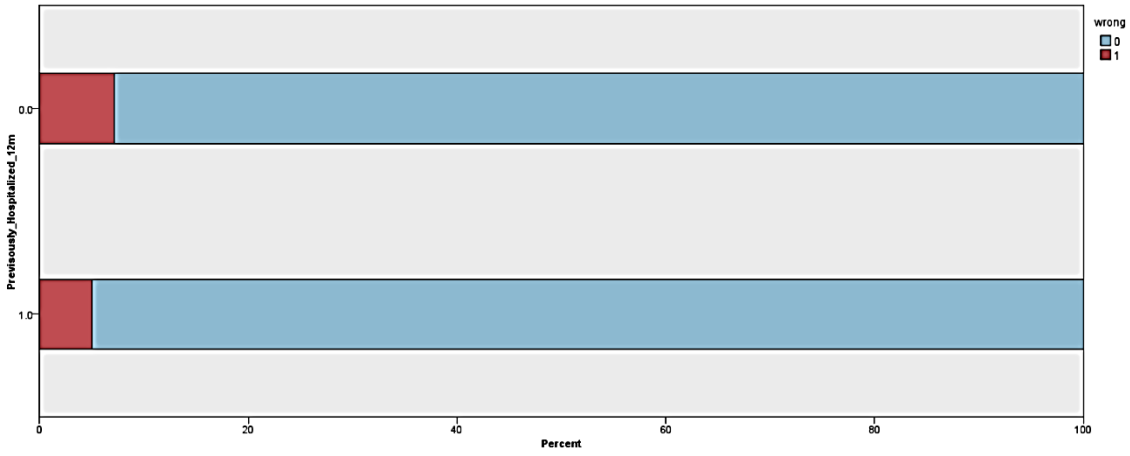
DH_2020_Now-distribution by type of decision



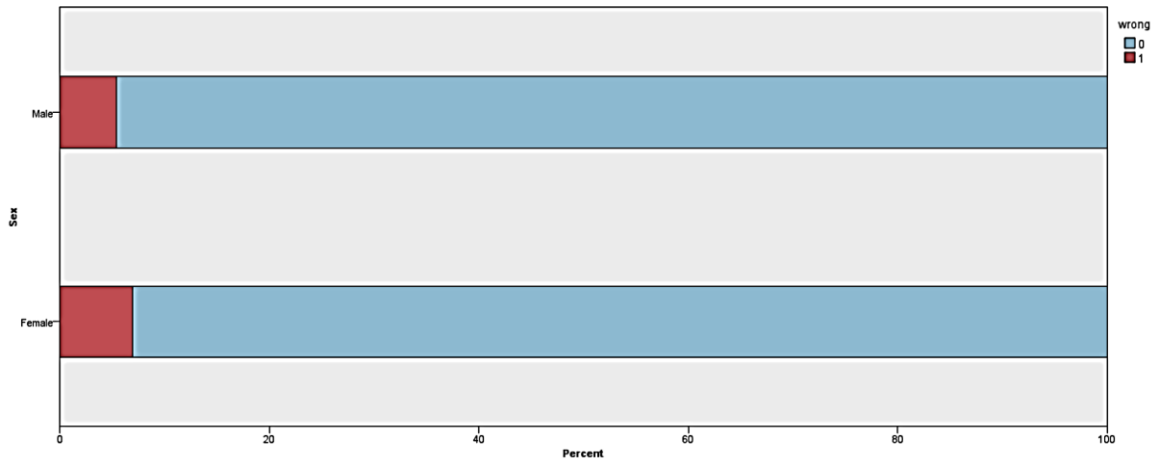
DH_6m-distribution by type of decision



Error Distribution by Previously_Hospitalized_12m



Error Distribution by Sex



Error Distribution by ICD