ÓBUDAI EGYETEM
ÓBUDA UNIVERSITY

DOCTORAL (PHD) THESIS

PETER PIROS

Improving Mortality Prediction with Machine Learning Models

Supervisors:

Prof. Dr. Levente Kovács

Dr. Rita Fleiner

DOCTORAL SCHOOL OF
APPLIED INFORMATICS AND
APPLIED MATHEMATICS

Budapest, November 26, 2023

Contents

Acknowledgments

I would like to express my gratitude to *Levente Kovács*, who, in addition to all his other duties, always assisted me in professional and administrative issues both under my master's and doctoral studies. His problem-solving ability always made possible to define the next strategic plans of the research. I am thankful for *Rita Fleiner* who helped me on a day-to-day basis. Her well-structured thinking and hardworking behaviour taught me a lot and made possible to publish the dissertation.

I am a great admirer of *András Jánosi*, who made a real, unfiltered dataset from the Hungarian Myocardial Infarction Registry (HUMIR) available for the current researches. His career with the heart attack registry is fascinating and I consider it an honor to be involved in his momentous work. I thank to *Tamás Ferenci* who was always ready to give feedbacks and advices on several questions regarding statistics and the dataset itself.

Thank you very much for the support of *Epam Hungary*, who supported the first part of my doctoral research, as an industrial partner.

And finally, I am grateful to my family I am blessed with, for my wife *Julia*, for my wonderful children, and for my Father and Mother. Without their support, I wouldn't have had a chance to publish any scientific results.

Abstract

Nowadays, we encounter new fields of application of artificial intelligence every day, ranging from chatbots that give sophisticated, more human-like responses than ever before, to predicting the future value of corporate stocks, and even machine-written software source codes.

Still, the greatest and most basic need is to put the amazing capabilities of this new – and truly promising – area of computer technology at the service of human life and health.

This dissertation provides an example of this by predicting the short-term and long-term chances of death of patients who have suffered a heart attack. During the research, the motivation was given by the fact that, knowing the result of the forecast, the attending physician can set up a more personalized treatment for the patients.

The first group of theses deals with the international and Hungarian heart attack registers, followed by a solution for using the Hungarian register as an input dataset for artificial intelligence algorithms.

The second thesis specifically deals with the exact results of machine learning models, which infer regularities from three years of complete, unfiltered heart attack cases in Hungary. The published models achieved, and in some cases even exceeded, the predictive capabilities of regression generally (and rightfully) accepted in the field – thus demonstrating the raison d'être of machine learning solutions in the current scientific field and in the Hungarian register.

The third thesis presents a more specific result: I investigated whether there is a difference in the predictive power of the decision tree models tuned with different resampling methods on the data of the Hungarian register. The results of the thesis can serve as a basis for further researches.

Absztrakt

Napjainkban szinte minden nap találkozunk a mesterséges intelligencia egy-egy új alkalmazási területével, kezdve a minden korábbinál kifinomultabb és ”emberibb” válaszokat adó chatrobotoktól, a részvények értékének előrejelzésén át akár a szoftvereket alkotó programkódok gép által történő megírásáig.

Mégis, a legnagyobb és legelemibb szükség abba az irányba mutatkozik, hogy a számítástechnika ezen új – és valóban sokat ígérő – területének bámulatos képességeit az emberi élet és egészség szolgálatába állítsuk.

Jelen disszertáció erre ad példát szívinfarktuson átesett betegek rövid és hosszú távú halálozási esélyeinek előrejelzésével. A kutatás során a motivációt az a tény adta, hogy az előrejelzés eredményének ismeretében a kezelőorvos személyre szabottabb kezelést tud felállítani a pácienseknek.

Az első téziscsoport a nemzetközi és magyarországi szívinfarktus regisztereket veszi sorra, majd megoldást ad a magyar regiszter adatainak mesterséges intelligencia algoritmusok bemeneti adathalmazaként történő alkalmazására.

A második tézis már kifejezetten a gépi tanulási modellek eredményeivel foglalkozik, amelyek három évnyi, teljes, szűretlen, Magyarország területén történt szívinfarktusos esetből következtetnek szabályszerűségekre. A publikált modellek elérték, sőt, bizonyos esetben meg is haladták a területen általánosan (és jogosan) elfogadottnak tekinthető regresszió előrejelző képességeit – ezzel demonstrálva a tudományterületen és a magyar regiszteren a gépi tanulási megoldások létjogosultságát.

A harmadik tézis pedig egy specifikusabb eredményt közöl: azt vizsgáltam a magyar regiszter adatain, hogy van-e különbség a különböző újramintavételezési módszerekkel hangolt döntési fa modellek prediktív erejében. A tézis eredménye további kutatások alapjául szolgálhat.

1 Introduction

1.1 Importance of myocardial registries and mortality prediction

In *Heart Disease and Stroke Statistics*, American Heart Association annually reports, that approximately every 40 seconds, an American will have an myocardial infarction (MI) - they did the same in the recent statistics titled *2022 Update* [1].

In addition, *Heart disease* (which can lead to myocardial infarction) is still at the first position of the ten leading cause of death, followed by cancer, unintentional injuries, chronic lower respiratory diseases, stroke and Alzheimer disease, respectively.

In the recent decades a decline is seen in coronary heart disease mortality [2]. Researches found that there is a difference in the decline based on the socioeconomic background of the patients [3][4]. For instance, the group of less educated people shows a smaller decrement. On one hand, the decline in coronary heart disease mortality can be one of the major significance of public health and a result of the new methods of treatment. On the other hand, more accurate and more complete information is needed to confirm such statements – because cardiovascular disease continues to be one of the most common cause of death in both men and women.

In the area with numbers like these, mortality prediction can and should play a very important role in the hand of physicians: with validated models, it becomes possible to select patients with high-risk of death and use this information in the process of treatment. Using new, real-life datasets to extract hidden information can lead to more effective treatment and prevention. As US surgeon Dr. Ernest Amory Codman suggested: ”Every hospital should follow every patient it treats long enough to determine whether or not the treatment was successful and to inquire ‘if not, why not?’ with a view to preventing similar failures in future.” [5]

Reliable, high-quality datasets are mandatory to build and train any type of predictive model. Hungarian Myocardial Infarction Register (HUMIR) project was introduced in 2010, initially collected AMI information only from five districts of Budapest and the county of Szabolcs-Szatmár-Bereg. In 2014, the Hungarian government selected it as the official myocardial database and obligated all hospitals in the territory of Hungary to report all MI-cases to HUMIR. In the recent years, around 15,000 new patients got registered per year and until December 2022, the 94 participating hospitals reported 157,724 cases in 142,439 patients. In all my related publications and theses I used a dataset from HUMIR to predict the mortality of patients hospitalized with acute myocardial infarction.

1.2 Aims of the research

Nowadays, most of the countries have their own mortality and disease statistics based on International Classification of Diseases – however, these statistics never contain clinical informations, for example results of former examinations, comorbidity or smoking behavior of the patients. Several databases store information about patients and diseases, but only a few system exists that focuses directly on myocardial events and treatments. My first aim was to collect these systems and to take advantage of the opportunities offered by the Hungarian register with developing the solution which turns the registers' raw data into an input data of machine learning algorythms.

Then, in my researches I developed several machine-learning models based on Decision Tree (DT), Neural Networks (NN), Logistic Regression (LR), Random Forest (RF), Generalized Boosted Model (GBM) and Ensembled algorithms to predict 30-day and 1-year mortality on the same, real-world, unfiltered dataset originated from HUMIR. The results achieved with these methods were published in several conferences and papers. The main question I was facing was if there is a competitive opponent for the mostly used and trusted regression in the world of machine learning algorithms.

The idea behind the approach that I was working on the same dataset in all the connected researches is the following: I was trying to establish an order in the list of different modelling techniques by keeping the dataset fixed and trying to maximize the prediction capability of each of our models.

1.3 Clinical introduction

My theses are built around *acute myocardial infarction*, so in this section I will highlight some general clinical information about the term.

As Cleveland Clinic summarizes *MI* itself, [6], it is ”a condition that happens because of a lack of blood flow to one’s heart muscle. The lack of blood flow can occur because of many different factors but is usually related to a blockage in one or more heart’s arteries. Without blood flow, the affected heart muscle will begin to die. If blood flow isn’t restored quickly, a heart attack can cause permanent heart damage and death.” The most common symptom is chest pain or discomfort which may travel into the shoulder, arm, back, neck or jaw [7].

Acute myocardial infarction (commonly called a heart attack) remains a leading cause of morbidity and mortality worldwide, despite substantial improvements in prognosis over the past decade [8].

The statistical characteristics of MI also speaks volumes: as *Heart Disease and Stroke Statistics* reports [1], the estimated annual incidence of MI is 605,000 new attacks and 200,000 recurrent attacks in the US. The overall prevalence for MI is 3.1% in US adults (>19 years of age). Males have a higher prevalence of MI than females for all age groups except 20 to 39 years of age. MI prevalence is 4.3% for males and 2.1% for females.

Since the following terms are not just present in the current study, but they also appear in the dataset of HUMIR, here I quote a definition and some descriptive information for each registry field:

1. *Heart failure* Heart failure means your heart isn’t able to pump blood as well as it should. When your heart has less pumping power, that can damage your

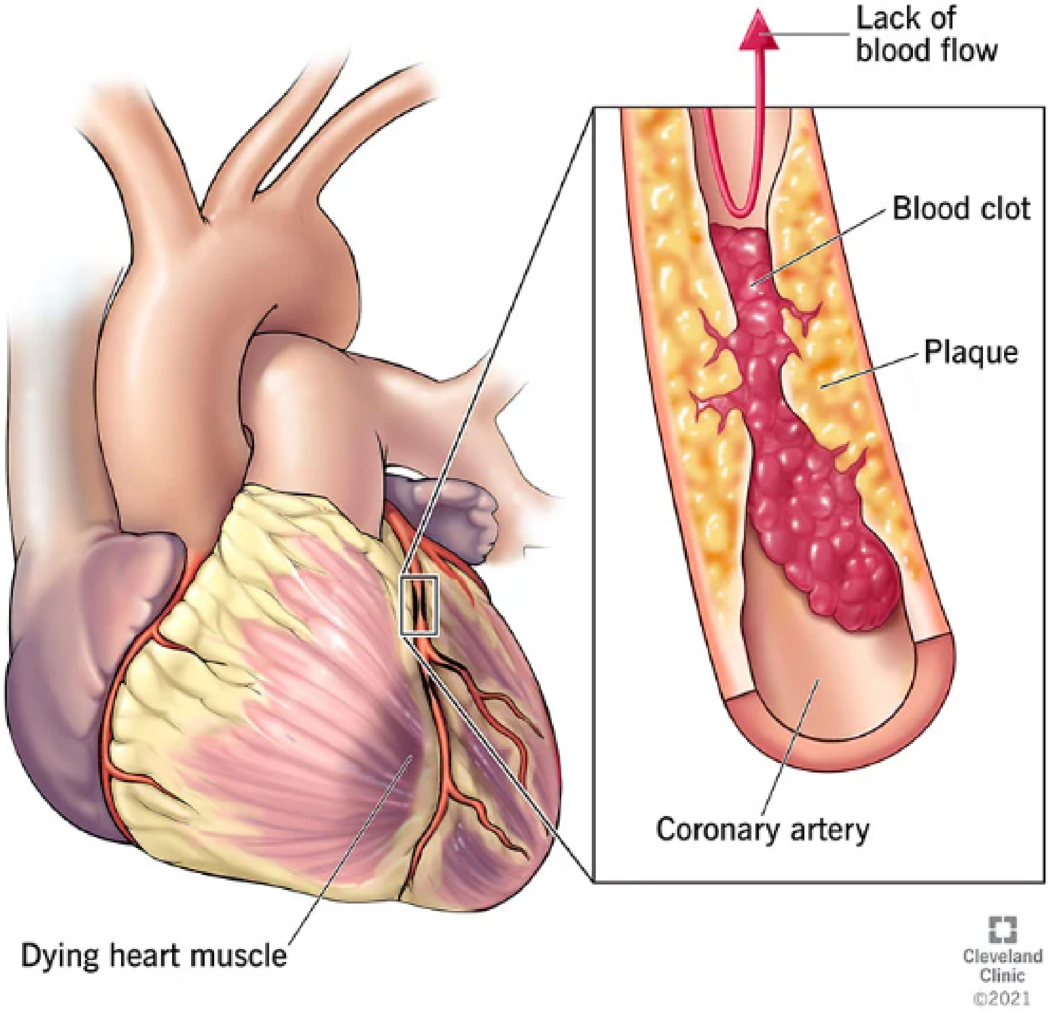


Figure 1.1: Heart attack - Illustration, Source: Cleveland Clinic

organs and fluid can collect in your lungs. Heart failure is the leading cause of hospitalization in people older than 65. [9]

2. *Hypertension* High blood pressure (hypertension) is the measurement of the pressure or force of blood pushing against blood vessel walls. When one has hypertension (high blood pressure), it means the pressure against the blood vessel walls is consistently too high. [10]

3. *Stroke* A stroke happens when part of one's brain doesn't have enough blood flow. This most commonly happens because of a blocked artery or bleeding in the brain. Without a steady supply of blood, the brain cells in that area start to die from a lack of oxygen. [11]

4. *Diabetes mellitus* Diabetes happens when your body isn't able to take up sugar (glucose) into its cells and use it for energy. This results in a build up of extra sugar in your bloodstream. [12]

5. *Peripheral artery disease* Peripheral artery disease (PAD, peripheral vascular disease or peripheral arterial disease) is plaque buildup in one's leg arteries, so the leg arteries cannot carry oxygen and nutrient-rich blood from the heart to the arms and legs. [13]

6. *Hyperlipidaemia* Hyperlipidemia, also known as dyslipidemia or high cholesterol, means you have too many lipids (fats) in your blood. Your liver creates cholesterol to help you digest food and make things like hormones. But you also eat cholesterol in foods from the meat and dairy aisles. As your liver can make as much cholesterol as you need, the cholesterol in foods you eat is extra. [14]

7. *Cardiogenic shock* Cardiogenic shock is a serious condition that happens when one's heart can't supply enough oxygen-rich blood to the body to meet its needs. It can be fatal when a lack of oxygen causes the organs to fail. [15]

2 Methods and tools of investigation

2.1 Methodologies

In the area of data mining, a well-known methodology called *CRISP-DM* exists which summarizes the main stages and questions of a given project, hence, serves as a base for the whole process. Although this process model was released decades ago, the aim of the fives developers (Integral Solutions Ltd, Teradata, Daimler AG, NCR Corporation and OHRA) succeed: in the 2020s the process model is still used and became the base of other process models as well.

Although other methodologies exist, for example: TDSP (Team Data Science Process) and KDD (Knowledge Discovery in Databases), here I list two reasons next to my decision:

1. there are no essential differences between these and CRISP-DM: even the official documentation of TDSP declares too, that, ”at a high level, these different methodologies have much in common” [16]

2. usage statistics reports [17] that CRISP-DM is still 4-5 times popular and more often used than any other framework

CRISP-DM stands for Cross-industry standard process for data mining. It defines the six sequential phases of a data mining project and also describes the main questions and tasks that the developers have to ask and solve to have a bigger chance of a successful data mining project, regardless of the exact area we are working on.

As I used *CRISP-DM* as process framework for all our researches, here the main findings and phases of the methodology are listed. The fix phases, with some remarks as it was summarized by the authors of [18], are the follows:

1. *Business understanding*: understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem

2. *Data understanding*: initial data collection, get familiar with the data, to identify data quality problems, to discover first insights into the data

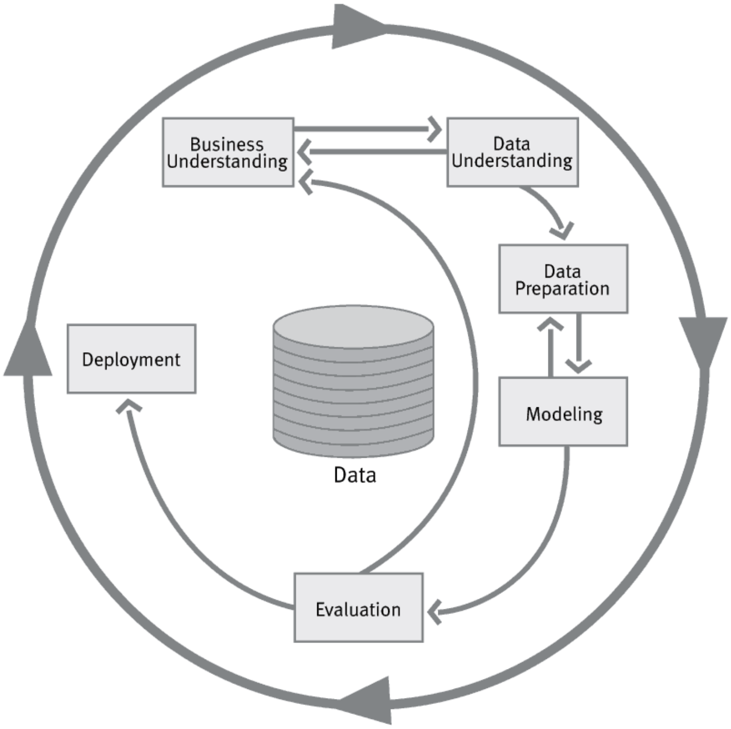


Figure 2.1: CRISP-DM methodology: Phases and connections

3. *Data preparation*: covers all activities to construct the final dataset, which will as the input of our modelling tools

4. *Modeling*: selecting the right modeling techniques, algorythms and their parameters

5. *Evaluation*: evaluate the model, review the steps executed to construct the model, to be certain it properly achieves the business objectives

6. *Deployment*: the knowledge gained from the dataset need to be organized, presented and built back into the original (business, health) environment

As Figure 2.1 shows, CRISP-DM clearly defines the sequential processes and its relations.

2.2 Measurement

In the studies, I used area under the Receiver Operating Characteristic (ROC) curve, or simply ROC AUC as a single-number measure for evaluating performance of a learning algorithm. Several other metrics could have been used, here are the two most important reasons behind the selection of ROC AUC:

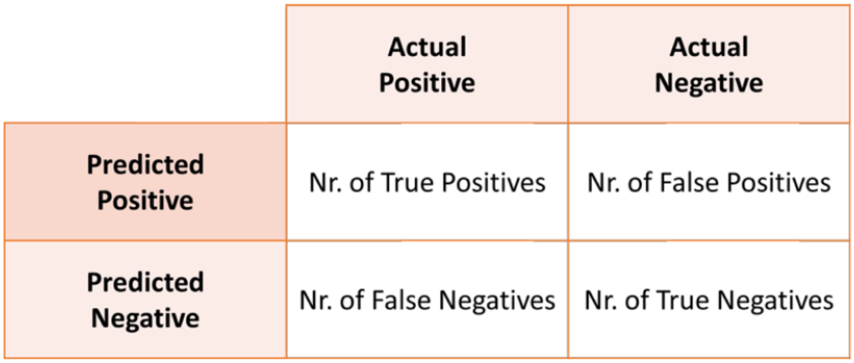
1. Literature recommendation: in several papers conclusions', ROC and ROC AUC have been presented as a recommended metric on real-world, imbalanced datasets. For example, Huang et al. [19] proved that AUC is – in general – a better measure than accuracy. Bradley [20] found that AUC has several desirable properties compared to accuracy.

2. Consistency: my earliest publication on the current topic was published with models described with ROC AUC. To be comparable with all the previous scientific results, I constantly applied this value.

F1-score, as the harmonic mean of precision and recall or even Precision-Recall AUC could have been used too through the whole research to describe the performance of each model - this could be a future research topic.

ROC curve is a graph which visually shows the performance of a classification model with binary target variable. The name *ROC* is coming from the history: originally, it was designed for operators of military radar receivers in World War II – but today it is highly used in almost any sectors from medicine to meteorology, including data mining researches.

As all the other classification metrics, a ROC curve can be generated by calculating the *Confusion matrix*. *Confusion matrix* is a table which summarizes the performance of a classification algorithm in a way that it shows the actual values against the predicted values. In a binary classification problem, the *Confusion matrix* is a 2x2 table as can be seen on Figure 2.2.



	Actual Positive	Actual Negative
Predicted Positive	Nr. of True Positives	Nr. of False Positives
Predicted Negative	Nr. of False Negatives	Nr. of True Negatives

Figure 2.2: Confusion matrix with a binary target variable, by definition

The *True Positive Rate* and the *False Positive Rate* at various threshold values build up ROC AUC. *True Positive Rate* can be called *sensitivity* or *recall* as well and can be calculated with dividing the the Nr. of True Positives with all the actual positives (the sum of Nr. of True Positives and Nr. of False Negatives). By definition, this value represents the *probability of detection*. *False Positive Rate* is also known as *probability of false alarm* because its calculation is the following: dividing the Nr. of False Positives with all the actual negatives (the sum of Nr. of False Positives and Nr. of True Negatives).

2.3 Software environment

Statistical analysis, data preparation, modelling and visualisation were all principally performed under the R statistical language [21], version 3.6.1. Table 2.1 summarises the used packages for given machine learning models.

For resampling and training the models, I used *Caret* [28] package; and to deal with missing data and imputation, *mice* [29] package was applied.

Table 2.1: R packages and their references used to train machine learning models

Decision tree	*Rpart* [22]
Neural network	*Nnet* [23]
Regression	*rms* [24]
Random forest	*randomForest* [25]
Generalized Boosted Models	*GBM* [26]
Ensembled	*caretEnsemble* [27]

R is a system for statistical computation and graphics. It provides, among other things, a programming language, graphics, interfaces to other languages and debugging facilities [30]. It is an open source project and available under GNU General Public License.

The R language was developed in the early 1990s as a dialect of the S statistical programming language. As the official manual writes, ”the language syntax has a superficial similarity with C, but the semantics are of the FPL (functional programming language) variety with stronger affinities with Lisp and APL”.

In R, the basic functionality is in-built, but to work on specific areas like on the machine learning domain, the functions come from *third-party* libraries called packages. In the January of 2023, over 19,000 packages are available from ’Analysis of Numerical Plankton Images’ to ’Analyze Text, Audio, and Video from ’Zoom’ Meetings’. These cover almost each and every area of modern statistics and data mining, including the most recent areas like the package which deals with ’Access to TikTok Ads via the ’Windsor.ai’ API’.

As by default, R has no graphical user interface, I used [31] as an integrated development environment (IDE). To reproduce our researches, it is not necessary to have an IDE, but with its features (syntax-highlighting, direct code execution, tools for plotting, history, debugging and workspace management) [31] makes development and source code management easier.

2.4 Hardware environments

During the investigation, I had two different hardware environments and configurations:

1. *Normal configuration* with the following resources: Intel Core i3 processor (i3-4030U CPU 1.90 GHz), 12 GB memory, no SSD. This environment is my personal laptop, and it was used in the early publications connected with Logistic regression, Decision tree and Neural network.

2. *Extended configuration* is a cloud-based architecture powered by *Amazon Web Services* with *EC2* instances[1]. It had the following parameters: 16 vCPU, 70 ECU, 64 GB memory (*m5.4xlarge* configuration). I used the *On-Demand* method, i.e. you pay for the compute capacity by the hours the instance runs. (For *m5.4xlarge*, it is $0.92 / hour in the *Europe / Frankfurt* region.) This environment was applied with the latest researches and models like Random forest, General Boosted Model and Ensembled.

During the investigation, I experienced that, on this size of dataset, the *normal configuration* was not suitable for latter models. For example, the neural network training times were around 5 hours with the *normal*, and 25 minutes with the *extended* configuration.

The different hardware configurations don't have effect on the models' performance or their prediction power (i.e. the hardware doesn't modify the underlying algorithms), only the training times got less with the cloud-based option.

[1] Amazon EC2: https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/concepts.html

3 Improving mortality prediction

In this section I present in detail the individual phases of the research and in this context I highlight the new scientific results. First, I review the the Hungarian Myocardial Infarction Registry in general; list the ongoing, official registries that collect cardiovascular data; then highlight the tasks related to data preparation of the dataset.

Then, in the second thesis, I list the machine learning models and their results, which I developed in order to achieve progress in the 30-day and 1-year mortality forecast compared to regression, which is considered a classic solution in this area.

And finally, in the third thesis another branch of the research is presented in which I investigated the differences between the resampling methods used to determine the tuning parameters in case of decision trees.

3.1 Registry overview & Data preparation (Thesis group 1)

3.1.1 Introduction

Nowadays, most of the countries have their own mortality and disease statistics based on International Classification of Diseases – however, these statistics never contain clinical informations, for example results of former examinations, comorbidity or smoking behavior of the patients. Several databases store information about patients and diseases, but only a few system exists that focuses directly on myocardial events and treatments.

The common vision behind such registries is the more specific the information we collect, the better quality control we can have. Thus, the quality of the treatment and then the prognosis of the patients improve.

In this section, I introduce Hungarian Myocardial Infarction Registry (HUMIR) registry; review the changing of the legal environment; show the completeness and validity, and, finally, overview some research results that were obtained using the data collected.

In Section 3.1.3, I list three other ongoing European myocardial projects as well.

3.1.2 Hungarian Myocardial Infarction Registry

The history

In the 1970s, World Health Organization (WHO) started a global investigation on acute myocardial infarction registries [32]. Hungary participated in the project with the South Pest Myocardial Infarction Registry [33], which was a paper-based system covering all patients who had acute myocardial infarction (AMI) in six districts of Budapest (affected 373 269 inhabitants [34]). The program made it possible to measure the incidence rates of AMI and the pre-hospital, in-hospital and 1-year mortality rates [35].

Based on the monitoring system, a new patient care system called Myocardial Infarction Patients' Care in CCU also started. The results were published almost a decade later [36]. The improvement on survival is significant, as shown in Table 2.1.

Table 3.1: Mortality rates in 1971 and 1979 in South Pest

	Results - 1971	Results - 1979
Admitted to Coronary Care Unit	7,8%	57,1%
Pre-hospital mortality	30,5%	25,1%
28-day-mortality - Pre- and in-hospital cases	51,7%	41,5%

In the next decades the diagnostic criterias, the clinical forms, and the optimal care strategies have significantly changed the physicians' knowledge about AMI. In addition, physicians faced some other challenges as well:

1. the current International Statistical Classification of Diseases and Related Health Problems (ICD-10) does not separate the two clinical forms of AMI: ST-segment (STEMI) and Non–ST-segment (NSTEMI) elevation myocardial infarction (unlike NSTEMI, in case of STEMI, a part of the electrocardiogram wave, the ”ST” segment, rises higher than normal)

2. therefore, the hospital and finance databases are not capable to investigate the treatment of AMI

3. the frequency of catheter-directed therapy of STEMI is not known

4. the 28-day and 1-year mortality information is not available

5. the frequency of drugs in the secundary prevention is also not known

As a consequence of the all these deficiencies, a need for a new infarction registry system arose. In January 1, 2010, a web-based system was introduced to collect AMI information from five districts of Budapest and the county of Szabolcs-Szatmár-Bereg [37].

Legal background

Until 2013, the transfer of the data was voluntary for the hospitals and was based on the patients' full, written informed consent. In Januar 1, 2014 a new legal environment changed this situation. As Hungarian Gazette writes [38], ”Diagnosing a myocardial infarction, the patient care doctor is forwarding the patient's identity and healthcare data, concerning the myocardial infarction, to the National Registry of Myocardial Infarction”. At this point, it became mandatory for hospitals to report each and every MI-case to the registry.

Table 3.2 summaries the essential stages of the Hungarian registry.

Table 3.2: Essential stages of the Hungarian Myocardial Infarction Registry

Stage	Data transfer	Date
Research plan		08.07.01 - 09.12.31
IR Pilot Investigation	voluntary	10.01.01 - 11.12.31
HUMIR	voluntary	12.01.01 - 13.02.28
HUMIR	obligatory	13.03.01 - 13.12.31
HUMIR	legal regulations	14.01.01 - present

In the recent years, around 15,000 new patients got registered per year and until December 2022, the 94 participating hospitals reported 157,724 cases in 142,439 patients.

Changing of completeness and data validation (2011-2016)

Table 3.3 contains the number of patients registered in HUMIR and the completeness of the stored data.

Completeness is a ratio and calculated on the proportion of two numbers: the numbers of patients treated with AMI originated from National Health Insurance Fund of Hungary (the central official organ of health insurance, supervised by the Government of Hungary; Hungarian acronym: OEP); and the numbers of patients registered by the hospitals in HUMIR.

Researches and results based on HUMIR

In the last few years, the data of HUMIR has been used in several researches to extract new results. In the followings, a few examples are listed.

A research reported that between 1st of January 2010 and 1st of May 2011 4293 patients were registered, among them 52.1% with STEMI, 42.1% with NSTEMI, while

Table 3.3: Completeness of the registered data in HUMIR

Year	Nr. of Patients (HUMIR)	Completeness of data (HUMIR)
2010	2407	below 30% (voluntary period)
2011	6877	below 30% (voluntary period)
2012	7550	appr. 30% (voluntary period)
2013	7828	51%
2014	10458	67%
2015	12536	82%
2016	13843	83,9%

3% of the patients had unstable angina (a type of acute coronary syndrome), and 2.8% of the cases had other diagnosis or the hospital diagnosis was missing [39].

In the year of 2010 and 2011, 4981 patients (3038 men) were included in the database. The target of the research [40] was to compare the clinical data and prognosis of patients with STEMI in that years. Women were significantly older (67.7±13.5 vs. 60.5±12.5 years; p<0.001). Hypertension, diabetes, and stroke were more frequent among women, whereas smoking and previous myocardial infarction were found more often among men. Percutaneous coronary intervention was significantly more frequently performed in men than in women (82.4% vs. 75.3%; p<0.001).

Based on the data of 8582 myocardial infarction patients (4981 with STEMI), a research found that the hospital, 30-day and 1-year mortality of patients with STEMI were 3.7%, 9.5% and 16.5%, respectively. In patients with NSTEMI these figures were 4%, 9.8% and 21.7%, respectively [41].

Another research based on information stored in HUMIR found that the mean age of STEMI patients was lower by 5.3 years than that of patients treated for NSTEMI. In the group of NSTEMI patients, the occurrence of diabetes, hypertension, peripheral vascular disease, and previous history of myocardial infarction and stroke were significantly more frequent. The in-hospital mortality rate of STEMI patients was 3.7%, and

30-day and 1-year mortality rates were 9.5 and 16.5%, respectively [42].

For 10 000 residents the incidence of myocardial infarction in Budapest was 28.63 in males and 16.21 in females, while in Szabolcs-Szatmár-Bereg county the mean incidence was 32.49 for males and 18.59 for females [37].

A research compared casemix, treatments and outcome for STEMI patients who are treated in Hungary or Sweden [43]. The Swedish data source was the SWEDEHEART registry. There were substantial differences in baseline characteristics between the two countries, with the Hungarian STEMI patients being younger and having more cardiovascular risk factors. More patients in Sweden received thrombolysis (5.4% versus 1.5%) or underwent primary PCI/subsequent coronary angiogram (91.2% versus 84.2%). The 30-day mortality was lower in Sweden than in Hungary (7.9% versus 9.5%; odds ratio 0.81, 95% confidence interval 0.72 to 0.93).

The aim of authors of another research was to obtain data on the significance of the culprit vessel in patients with STEMI treated successfully by primary percutaneous coronary intervention [44]. The culprit vessels were the left main artery, left anterior descendent artery, left circumflex artery, and right coronary artery. The majority of the culprit lesions were found in the left anterior descendent artery (44.3%), the right coronary artery (40.9%), and the left circumflex artery (13.7%). The culprit vessel was overall a highly significant ($p<0.0001$) factor of survival.

In a research [45], the frequency of the real-life usage of coronary intervention, its long-term efficacy and safety in elderly patients with AMI were investigated. A total of 8485 consecutive patients were enrolled; 65% of the patients were male (mean age was 65.1 ± 12.4); and 51% of all cases was STEMI. As a conclusion, the authors state that coronary intervention is underused among the elderly despite the mortality benefit.

In 2015, 12 681 patients had 12 941 acute myocardial infarctions. A research [46] in 2017 found that less than half of patients (44.4%) were treated with STEMI. 91.6% of the patients were treated in hospital with invasive facilities. Most of the patients

(94%) with positive coronary arteriography were treated with percutaneous coronary intervention. The 30 day mortality of the whole group was 12.8% vs. 8.6% of patients treated with an invasive procedure.

3.1.3 Existing myocardial registers in Europe

In this section, I list three ongoing European myocardial projects: Myocardial Ischaemia National Audit Project (MINAP) in England, Swedish Websystem for Enhancement and Development of Evidence-based care in Heart disease Evaluated According to Recommended Therapies (SWEDEHEART) and National Registry of Acute Myocardial Infarction in Switzerland (AMIS Plus). Where the information is available for the public, I discuss the validity and accuracy of the stored data.

Myocardial Ischaemia National Audit Project

The Myocardial Ischaemia National Audit Project is a national clinical audit of the management of heart attack in England, Wales and Northern Ireland. It is one of six national cardiac clinical audits that are managed by the National Institute for Cardiovascular Outcomes Research (NICOR), which is part of the Institute for Cardiovascular Science at University College London. MINAP was established in 1999 and data collection began in October 2000.

The aims of MINAP: to audit the quality of care of patients with acute coronary syndrome (ACS) and provide a resource for academic research [47].

Based on the 2015 Annual Report [48] 217 acute hospitals in England, Wales and Northern Ireland participate in the project and continuously send the encrypted data of 130 fields covering demographic factors, co-morbid conditions and treatment specifications in hospital.

80% of hospitals use MINAP software to enter the data into the system and the rest of the hospitals use a locally developed software or commercial applications for this purpose.

The data itself is available for research by application to NICOR.

Swedish Web-system for Enhancement and Development of Evidence-based care in Heart disease Evaluated According to Recommended Therapies (SWEDEHEART)

SWEDEHEART was launched in 2009 after merging four Swedish national registries on coronary artery disease. The project is supported by the Swedish Society of Cardiology, the Swedish Society of Thoracic Radiology, the Swedish Society of Thoracic Surgery, and the Swedish Heart Association. The registry is financed by the Swedish Association of Local Authorities and Regions, the Swedish state, and the Swedish Heart-Lung Foundation [49].

The primary purpose of SWEDEHEART is to support development of evidence-based therapies in acute and chronic coronary artery disease and in catheter-based or surgical valve intervention by providing continuous information on patient care needs, treatments, and treatment outcomes.

The number of participating Swedish hospitals is 74 in 2016, corresponding to 95% degree of coverage at hospital level. About the data fields [50]:

1. 106 variables for patients with ACS,
2. 75 variables regarding secondary prevention after 12-24 months,
3. 150 variables for patients undergoing coronary angiography/angioplasty and
4. 100 variables for patients undergoing heart surgery.

All data are registered on a web-based interface directly by the caregiver. The data itself is available for research by application to the SWEDEHEART steering group.

National Registry of Acute Myocardial Infarction in Switzerland (AMIS Plus)

The Swiss registry of acute coronary syndrome is called National Registry of Acute Myocardial Infarction in Switzerland (AMIS Plus). In the list of the aims of the registry, we find:

1. to understand the transfer, use and practicability of knowledge gained from randomised trials
2. to generate input for subsequent prospective and randomised studies.
3. to determine how adherence to guideline-based treatments in the ”real world” works.

The AMIS Project was initiated in 1997 [51]. AMIS Plus has been continuously collecting data since then on patients admitted to Swiss hospitals with acute coronary syndromes. Today it operates as an industry-sponsored project. The treating doctor or trained study nurse enter the information online or through a paper-based questionnaire. Out of 106 hospitals treating ACS in Switzerland, 76 hospitals temporarily or continuously send data into the registry (coverage of 72%).

The data itself is available for research with the approval of the AMIS Plus Steering Committee.

Auditing & Quality of data

One of the key questions about this large amount of data is its accuracy. All of the registers use solutions based on some kind of randomisation technique to check and improve the quality.

MINAP has a specially designed data validation tool. Every year, the system requires every hospital to re-enter 20 data items from the medical records of 20 randomly selected patients [47]. Then, this re-entered data gets compared with the original ones and an agreement score is generated to every hospital. At the end of the process, the hospitals get their scores with advices how to improve performance. The median level of agreement between MINAP data and re-audit data (across all hospitals) was 72% in 2003 and has risen to 89.5% in 2008 [52].

SWEDEHEART uses personal validation solution. A monitor visits approximately 20 hospitals every year. In 2007, accuracy of 96% was reported [50].

AMIS Plus also uses random selection for auditing. Two large and three small hospitals, and about 5-10 patients are randomly selected each year. A summary from 2014 [53] reports there were 0.05% critical, another 0.05% major, and 2.2% minor findings.

3.1.4 Dataset & Data Preparation

The dataset of my research was an extract from HUMIR from the year 2014 to 2016. It contained 47,391 patients with AMI. Less than half of the patients (42.8%) were treated with STEMI; and 40.3% of the patients were female. In the whole dataset the mean age was 67.06 year, with 67 as median and 12.8 as standard deviation.

Table 3.4 shows patient characteristics related to AMI in the dataset in case of STEMI and NSTEMI patients. The source of this information can be one of the followings: previously identified diseases diagnosed by the patent's own physician and based on a former hospital final report; or disease status identified in the current treatment (diabetes mellitus, hypertension, peripheral arterial disease).

The patient record contained 23 separate fields, which can be categorised into 3 groups. (The categorisation is made by the author of the current dissertation in order to make referencing easier.) The followings list the attributes of the dataset.

Group 1: General information about the patient

Table 3.4: Patient characteristics

	STEMI	NSTEMI
Gender: Male	n = 20 302, 61% (12 448)	n = 27 089, 58% (15 829)
Age	Mean: 65, St. dev: 13.08	Mean: 69.5, St. dev: 12.26
Prior myocardial infarction	n = 19 355, 20% (3 935)	n = 25 973, 34% (8 750)
Previous heart failure	n = 18 621, 11% (1 976)	n = 25 268, 23% (5709)
Hypertension	n = 19 623, 75% (14 760)	n = 26 568, 86% (22 839)
Prior stroke	n = 19 149, 8% (1 576)	n = 25 733, 13% (3 222)
Diabetes mellitus	n = 19 240, 29% (5 525)	n = 26 090, 39% (10 137)
Peripheral artery disease	n = 18 116, 11% (2 024)	n = 24 604, 19% (4 792)
Percutan coronary intervention	n = 20 306, 81% (16 490)	n = 27 091, 58% (15 620)
Hyperlipidaemia	n = 16 332, 31% (5056)	n = 22 335, 40% (8 868)

Smoking (Current + Former)	n = 13 171, 48% (6 297) +
	n = 13 171, 15% (1 914)
		n = 15 564, 34% (5 259) +
		n = 15 564, 22% (3 406)

1. *Event ID* is a 3 to 5 character-long identifier which uniquely identifies an event in the registry. It is an auto-incremented numeric field.

2. *Patient ID* is a 3 to 5 character-long identifier which uniquely identifies a patient in the registry. If two myocardial events belong to the same patient, the Event IDs are different but Patiend IDs are the same.

3. *Date of birth*

4. *Gender* is either *man* or *woman*

5. *ZIP code* is 4 character long geographic code giving the code of the area where the patient lives

6. *If the patient alives* as a binary field

7. *Date of death*. It is empty unless *If the patient alives* field is *false*.

Group 2: Previous medical history

This group of fields holds information about the patient related to previously reported diseases, prehospital treatment and smoking. Except smoking, these are all simple numeric fields with two possible values: *yes* or *no*. Of course, missing values are present here as well.

1. *Myocardial Infarction*
2. *Heart failure*
3. *Hypertension*
4. *Stroke*
5. *Diabetes mellitus*
6. *Peripheral artery disease*
7. *Hyperlipidaemia*
8. *Cardiogenic shock*
9. *Smoking*

If set, field *Smoking* can have three values: The patient is an active smoker, never tried or stopped it.

Group 3: Information about the pre- and in-hospital treatment

The last group consists of 7 fields which is a set describing the properties of the pre- and in-hospital treatment.

1. *Prehospital reanimation*
2. *Percutaneous Coronary Intervention* (PCI) during hospital stay

3. *Level of creatinine* is an almost all-empty field with some numeric value (range in the dataset: 0 to 698)
4. *Diagnosis* is the type of the heart attack: STEMI (MI with ST-elevation on the ECG) or NSTEMI (MI without ST-elevation on the ECG)
5. *Treatment ID*
6. *Date of admission*
7. *Creatinine* is a categorical field which value can be normal or abnormal

Target variables

Two target variables were used in the predictive models, namely 30-day and 1-year mortality.

”1-year mortality” is a binary result for each patient and AMI event; means the fact that if the given patient dies in a 1-year period from the date of submission. As can be seen in this section above, this information is not implicitly included in the list of fields – but can be calculated with a single operation based on two other fields.

This one operation checks if the *If the patient alives* field is not true and subtracts the *Date of admission* from *Date of death*. If this number (number of days between these two dates) is lower than 365, then the value of the target variable is 1 (true), otherwise it is 0 (false).

The other target variable is 30-day mortality, whose definition and method of calculation only differs from 1-year mortality in the length of the period investigated.

From the full dataset, a training and a validation sets were created with maintaining the original distribution of the target variables. The models were trained on the training set and then the predicting power was validated on the validation set. Proportion of 7:3 was used in the process of constructing the training and validation sets.

Data Preparation

As it was discussed in Section 2.1., data preparation is a multi-step process and it covers all activities to construct the final dataset from the initial raw data.

Data preparation tasks usually have to be performed multiple times and not in any prescribed order. Tasks include table, record and attribute selection as well as transformation and cleaning of data. The aim of this activity is to produce a final dataset in a format that can be used as the direct input of predictive models.

Collection, assessment, consolidation, cleaning, data selection and transformations are all tasks, together with missing data handling which builds up the process of data preparation.

In this section, I list the most important data preparation tasks applied in our study.

Formatting

Most of the fields needed some adjustment during the process of data preparation. After importing the initial dataset, the dates were simple character-based values instead of date objects. In order to be able to make operations between dates, I had to convert them into the appropriate format.

In order to generate a uniform input to all of the models, I have changed the set of values in case of almost each field.

Constructing

With the process detailed in the subsection *Formatting*, I have created 15 new, binary fields. In addition to that, *Age* of patients was added with a simple calculation from *Date of admission* and *Date of birth*.

Figure 3.1 shows the histogram of this new field *Age* in the full dataset (transparent bars), and only in case of patients who died within one year (grey bars).

Eliminating

There were 9 fields which were eliminated because either they didn't hold relevant information, either they were used already in the construction - but could have

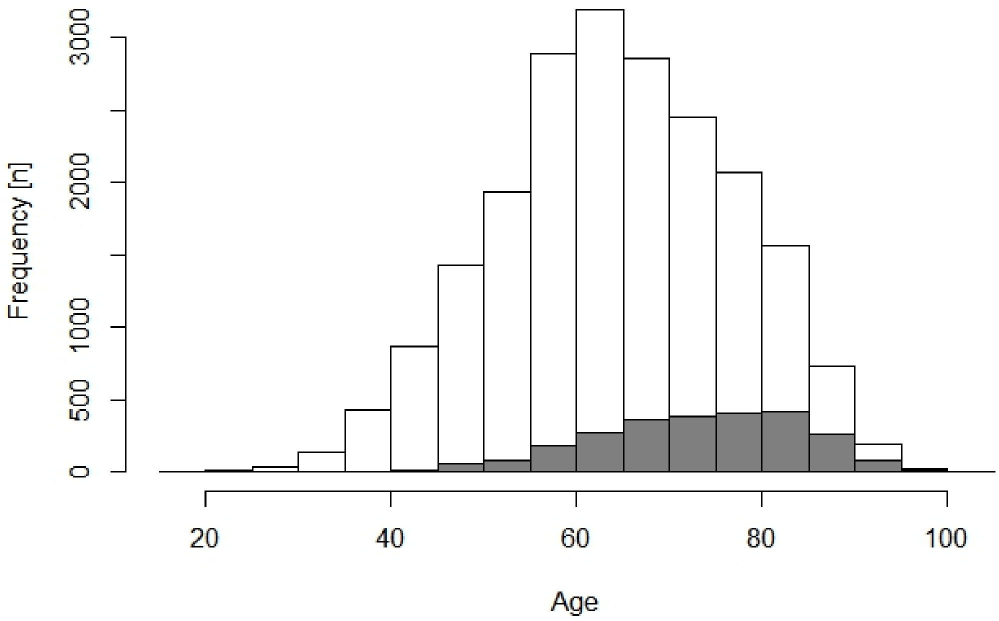


Figure 3.1: Histogram of age in the full dataset and in case of patients with 1-year mortality

slowed down the process of model training. For this reason, *Event ID*, *Patient ID*, *ZIP code* and *Treatment ID* were deleted with absolutely no usage; *Level of creatinine* was eliminated because only in case of 2.8% of the rows were filled in; and although *Date of admission*, *Date of death*, *Date of birth* and *If the patient alives* have been used in the process of constructing new fields (see *Target variables subsection*), they were eliminated as individual fields.

Imputation

Addressing missing values proved to be essential in the researches. The missing values of HUMIR dataset in percentage for each field are shown in Table 3.5 (table contains only the attributes where at least one missing value is present):

Table 3.5: Presence of missing values in percent

Event ID	0	Myocardial infarction	4.3	Pre-hospital reanimation	5.8
Patient ID	0	Heart failure	7.4	Cardiogenic shock	8.3
If the patient alives	0	Hypertension	2.5	PCI	0
Date of death	0	Stroke	5.3	Level of creatinine	0
Gender	0	Diabetes mellitus	4.4	Diagnosis	0
Date of birth	0	Peripheral vascular disease	9.9	Treatment ID	0
ZIP code	0	Hyperlipidaemia	18.4	Date of admission	0
		Smoking	39.4	Creatinine	6.0

As [54] warns, ignoring incompleteness or handling the data inappropriately may bias study results, reduce power and efficiency, and alter important risk/benefit relationships.

To deal with missing data multiple imputation using fully conditional specification (FCS) [29] and Bayesian linear regression was applied with 5 imputations and 5 iterations, leaving the final, prepared dataset size at $n = 47{,}391$.

As a result, 5 different sub-datasets were created, and on each I performed the full process of modelling for both the 30-day and 1-year mortality, as it will be detailed in *Section 3.2 (Performance of machine learning algorithms)*.

Although there are more approaches of addressing missing values from simple ones (filling in with the mean of the given variable) to more complex solutions like maximum likelihood [55] or Bayesian estimation [56], I decided to go with FCS-based multiple imputation, where imputations are generated sequentially by specifying an imputation model for each variable given the other variables. The reasons behind the selection of this technique, are the follows:

1. Multiply imputed datasets make possible to check the ”goodness” of the imputations, by comparing the prediction power of the given models trained on the different imputations

2. In the HUMIR-dataset, we have variables with different scales and complex relations between them, which would question the usability of another type of multiple imputation, Joint Modeling (see [29])

3. In addition, as [54] mentions, FCS multiple imputation is still rarely used in epidemiology, although it is ”a powerful and statistically valid method for creating imputations in large data sets with complex data structures”

Working with the imputed datasets

Applying multiple imputation, I have investigated to find a proper way to work with multiply imputed datasets. I found that although there are some tools dealing with the topic, a solution which can handle these imputations massively, simply does not exist:

1. *caret*'s *with.mids()* function performs a computation of each of imputed datasets in data – but it doesn't aggregate, just compare

2. *caretEnsemble* [27] is a package for making ensembles of different caret models – but it can be used only to the same dataset

3. as a part of *mice* package, *complete()* function generates a long matrix on a multiply imputed dataset – it's a handy tool but doesn't aggregate results

4. *pooling* [57] can be used to analyze multily imputed datasets in one step.

As a consequence, for each of the five imputation and two target variable, I always constructed several models. For example, when comparing decision tree, neural network and regression, I had 10 decision tree, 10 neural network and 10 regression models - and, finally, we had 15 models for 30-day mortality and another 15 models for 1-year mortality.

With fulfilling the previous points step-by-step, the raw data became suitable for participating in the implementation of the machine learning process as input for the predictive models.

Thesis group 1: In a methodological approach, I have discussed and analyzed the data preparation of artificial intelligence algorithms on the dataset of the Hungarian National Myocardial Infarction Register.

Thesis 1.1

At the international level, I have took a look at the official registries that collect cardiovascular data; within this, I have highlighted the uniqueness of the Hungarian National Myocardial Infarction Register and gave a ”literature recipe” for the use of artificial intelligence methods.

Thesis 1.2

For the dataset of the Hungarian Myocardial Infarction Registry, I have developed a data preparation procedure, with which the raw data became

suitable for participating in the implementation of the machine learning process as input for the predictive models.

Relevant own publications pertaining to this thesis group:
[P-1] [P-2] [P-3]

3.2 Performance of machine learning algorithms (Thesis 2)

After having the prepared dataset, the next phase is the modelling, as CRISP-DM methodology declares (see Section 2.1). In my research, in addition to the 'classical' regression model, I built models with Decision Trees, Neural Network, Random Forest, General Boosted Model algorithms and Ensembled techniques to predict the 30-day and 1-year mortality of patients who have suffered a heart attack.

The following considerations were the reasons for choosing the aforementioned algorithms:

1. Decision tree: despite its obvious disadvantage (very prone to overfitting), it is the basis of the most modern algorithms (among those included in the research: Random Forest, GBM). In addition, it is a white-box model which fact could be a help in the hand of phisicians.

2. Neural Network: according to the literature, even when applied alone, it brings acceptable results in this field of medicine ([58] and [59]). In addition, as a black-box model, it represents another type of model.

3. Random Forest and Boosting: they are obviously the most effective model types in many fields of medicine, including mortality prediction

4. Ensembled: with the ensembled approach, the goal was to exploit common advantages of the individual models

In the followings I present the context and the scientific results in the next order:

1. first of all, I review the challenging question of ”Regression vs. Machine learning solutions” especially on the field of mortality prediction and AMI *(Section 3.2.1)*

2. then I list the key points and characteristics of the applied models and quote some important connected results from outer researches *(Section 3.2.2 - 3.2.6)*

3. then finally, I list my exact 30-day and 1-year mortality prediction results *(Section 3.2.7 - 3.2.9)*

3.2.1 Introduction - Regression vs. Machine learning solutions

Logistic regression is the most commonly adopted and trusted model in the field of mortality prediction. From the view of medicine, it is import to know if there is a competitive opponent for the mostly used and trusted regression. Several studies work with regression - in general, it represents the ”classical” statistical approach with fast computational time and high accuracy. Next to regression, researchers try to use machine learning-based solutions to build predictive models to reach higher accuracy. In this subsection, I list a few attempts and their results.

Lee et al. [60] developed a logistic regression model based on a dataset of patients diagnosed with heart failure at multiple hospitals in Ontario, Canada. In a derivation set of 2,624 patients, the mortality rates were 8.9% in-hospital, 10.7% at 30 days, and 32.9% at 1 year. While validating the model, the area under the Receiver Operating Characteristics (ROC) curve was 0.80 for 30-day mortality and 0.77 for 1-year mortality.

Based on a dataset of 52,616 patients, Jack et al. [61] developed logistic regression models to predict 30-day and one-year mortality after an AMI. They predicted mortality with an area under the ROC curve of 0.78 for 30-day mortality and 0.79 for one-year mortality. In two independent validation datasets, this model reached 0.77 and 0.78, respectively.

Chin et al. [62] developed (n = 65,668) and validated (n = 16, 336) a logistic regression model to predict the risk of in-hospital mortality of patients with AMI. They reported AUC of 0.85 and 0.84 in the derivation and validation cohorts, respectively.

Clermont et al. [63] compared the performance of logistic regression and artificial neural network (ANN) models while predicting hospital mortality for patients in the intensive care unit. Seven intensive care units with 1,647 admissions were investigated, and finally they found that the two models have similar performance (0.80 and 0.84 as the area under the ROC curve).

Nilsson et al. [64] aimed to develop a method to select risk variables and predict mortality after cardiac surgery by using artificial neural networks. They also used area under the ROC curve as performance indicator and found that area of artificial neural networks (0.81) was larger than the logistic model's (0.79).

Orr found [65] that implementing a probabilistic neural network model to estimate mortality risk following cardiac surgery is relatively rapid, and it is an alternative to standard statistical approaches. He got 0.72 and 0.81 as ROC AUC for the training and validation sets. The neural network model reached 0.74 on an independent dataset of the following year.

Voss et al. [66] investigated if neural networks improved on the risk estimate of the commonly used logistic regression. They used multi-layer perceptron (MLP) and probabilistic neural networks (PNN) to estimate the risk of MI or acute coronary death. As they reported, the AUC of the MLP was greater than that of the PNN (0.897 versus 0.872), and both exceeded the AUC for LR of 0.840. As a conclusion, the authors declare that use of the MLP to identify high-risk individuals as candidates for drug treatment would allow prevention of 25% of coronary events in middle-aged men.

Austin compared [58] the predictive power of logistic regression with that of regression trees for predicting mortality after hospitalization with an AMI. His study shows that regression trees (0.762 AUC) do not perform as well as logistic regression (0.845 AUC). Author used data on 9 484 patients admitted to hospital with an AMI in Ontario, Canada.

In another study, Austin et al. [59] used ensemble-based methods, including bootstrap aggregation (bagging) of regression trees, random forests, and boosted regression trees. They found that ensemble methods offered substantial improvement in predicting cardiovascular mortality compared to conventional regression trees, but may not lead to clear advantages over conventional logistic regression models.

Convolutional neural network was applied by Acharya et al. [67] to automatize the diagnosis of congestive heart failure using ECG signals. They presented an 11-layer deep convolutional neural network model and out of four different datasets, one attained the highest accuracy of 98.97%, specificity and sensitivity of 99.01% and 98.87% respectively.

One group of quoted researches uses only regression; other ones compare regression with neural network, but in another field of medicine; some publishes only the results of neural networks or decision trees; but only a few of them investigated the differences on an AMI dataset and, the most important, none of them used the official Hungarian myocardial registry to predict short- and long-term mortality.

In this study, I used the results of the logistic regression model as a reference for comparison. I used regression without assuming any interaction between the variables or applying any penalization.

3.2.2 Decision tree

Decision tree is a graphical model that uses a tree-like structure of classifying examples. Classification trees (target variable with discrete values) and regression trees (target variable with continuous values) are one of the simplest tools in the field of machine learning, they are widely used because of their comprehensibility and their ”white-box” property (in the sense that the final knowledge of a model can be expressed in a readable form).

In each step, the algorithm splits the source set into two subsets based on a feature and a corresponding value. This operation gets repeated in a recursive manner, until a node has all the same values of the target variable, or another stopping criteria fires.

A few usual parameters to determine when to split are: the minimum number of observations that must exist in a node in order for a split to be attempted; the minimum number of observations in any terminal node; the maximum depth of any node of the final tree; a complexity parameter (*cp*) which is used like the following: ”any split that does not decrease the overall lack of fit by *cp* is not attempted” [22].

Although DT itself usually doesn't deliver acceptable prediction result, it has an important role in Machine Learning as it serves as the basis for many other good-performing algorithms like Random Forest or Boosting techniques.

3.2.3 Neural network

Artificial neural networks are based on a collection of connected neurons (nodes), where the connections can transmit a signal from one neuron to the other. This theory is inspired by biological neural networks. Neural networks are usually referenced as ”black box” models, as they have no ability of explaining their answers and presenting the acquired knowledge in a comprehensible way [68].

A basic neural network has interconnected neurons in three different types of layers (layers are listed with their primary roles): Input Layer (processes and analyzes the data, then passes it on the next layer); Hidden Layer (receives data from the input layer or from another hidden layer; passes it on to the next layer; a network can contain any number of hidden layers); Output Layer (receives data from the previous layer and gives the final result). Each connection has a default weight and the learning itself is the process of changing the weights after each piece of data is processed. Activation function is responsible for calculating the output of the node which is transmitted to the next layer.

A high number of neural network-based applications can be found in the last few years ([69] [70] [71] [72] [73]). In this study, a feed-forward neural network with a single hidden layer was used.

3.2.4 Random Forest

The basic idea of Random Forest algorithm is building many small, weak, less-correlated trees in parallel. Then, with averaging (regression) or majority vote (classification) we can combine the weak trees to form a strong learner.

The RF algorithm works as follows: first, we select a bootstrap sample (S(i)) from the original dataset (S), where S(i) denotes the *i*th sample). Second, on each bootstrap sample and on each node, a decision tree-based learning method gets performed, but with only a randomly selected, (very) small subset (f) of features (F).

The following two pseudocodes illustrate the basic behaviour of Random Forest algorithm. The first one contains the selection of bootstrap sample and the combination (averaging or majority voting) of the small trees.

```
Require: S: Initial Dataset, F: Features, B: number of weak trees in the forest
Output: H: the final tree
  RandomForest :
  $H = 0$
  for $i = l$ to $B$ do
      S(i) = a bootstrap sample from S
      h(i) = ModelWeakTree(S(i), F)
      H = H U h(i)
  end for
```

```
return $H$
```

The second part of the algorithm demonstrates how the weak trees are created:

```
Input: S, F
Output: TREE
  ModelWeakTree :
for all node do
    f = small subset of F
    Split on best feature and value in f
end for
return $TREE$
```

The main advantages of RF are the small trees: by reducing the number of potentially eligible features to a small number (usually: 2, 3, 5) for each splitting, the training speed gets faster, furthermore, the correlation between the trees get decreased. From the view of system architecture, the process of training can be run in a parallel way: the bigger number of cores in the processor has good impact on the training time.

As the algorithm above depicts, RF is an ensemble method: while creating the final, strong learner, it uses other techniques as well.

The applied implementation of Random Forest algorithm is based on the original publication published by L. Breiman [74]. In the process of RF-modelling, the two most important questions are the value of the following parameters:

1. *ntree*: number of small, weak trees to grow

2. *mtry*: the number of variables randomly sampled as candidates at each split

In case of classification models, the default value for *mtry* in the R package is the square root of the number of predictor variables, while in case of regression, it is the number of predictor variables divided by 3 (rounded down, in both cases).

In one hand, the bigger number of weak trees (*ntree*) usually produces more accurate models, but can slow down the training process because of the bigger memory requirement.

3.2.5 General Boosted Model

Boosting is a heavily applied method in the wide spectrum of medicine. In the next examples researchers used this method to construct strong learners or found that it is one of best-performing method in the field of mortality prediction.

Saravanou, Antonia, et al. examined infant mortality prediction and established a boosted tree model which performs the best (0.85 AUC) in predicting the number of infants per 1000 that do not survive until their first birthday [75].

Du, Xinsong, et al. were working with prediction of in-hospital mortality of patients with febrile neutropenia using gradient boosting tree and other non-linear and linear models [76], and, as a result, they achieved 0.92 AUC.

One of the recent attempt is made by Yan, Li, et al.: they applied a gradient boosting-based algorithm to predict mortality for COVID-19 patients [77].

In addition, the application of a boosting algorithm can be found in many other areas (early hospital mortality prediction using vital signals [78], mortality prediction in patients admitted to an Intensive Care Unit with a diagnosis of sepsis [79], mortality prediction in transcatheter aortic valve implantation [80], prediction of in-hospital mortality after pancreatic resection in pancreatic cancer patient [81]).

Boosting and heart diseases

Angraal, Suveen, et al. applied five methods including RF, gradient descent boosting and support vector machine for predicting mortality and heart failure (HF) hospitalization for outpatients with HF with preserved ejection fraction [82]. They found that the RF was the best performing model with a mean C-statistic of 0.72 for predicting mortality.

Risk scores for prediction of mortality 30-days following a ST-segment elevation myocardial infarction (STEMI) have been developed by [83] Shouval, Roni, et al., while they compared six machine learning algorithms to the conventional validated risk scores. They observed the maximal predictive performance with the RF models (AUC = 0.91), performing similarly to Naïve Bayes (AUC = 0.87) and AdaBoost (AUC = 0.87) and significantly better that the other algorithms.

Investigating a database containg patient information from 2005 to 2012 from a cardiac surgical center, Allyn Jérôme, et al. found that out of the different machine learning techniques, RF and Gradient Boosting Machine had the best AUC value. Both achieved 0.768 when feature filtering was skipped; and there were only 0.004 difference with enabled filtering. [84].

Sherazi, Syed Waseem Abbas, et al. developed multiple mortality prediction models using gradient boosting machine, generalized linear model (GLM), RF and deep neural network (DNN) [85]. The goal was to propose a machine learning–based 1-year mortality prediction model after discharge in clinical patients with acute coronary syndrome. The best AUC values were reached by gradient boosting machine and DNN models (0.898), then RF (0.883), then GLM (0.873). Overall, gradient boosting machine was superior to other approaches in the aspect of AUC, recall, accuracy, and F-score.

Fundamentals of GBM

Instead of using a single model, Boosting represents the idea that the final model could be more powerful if we continuously add weak models (e.g. decision trees) to our system, each compensating the weaknesses of its predecessors. In a more detailed way, as it is visualized in Figure 3.2., boosting is the process of iteratively adding basis functions in a greedy fashion so that each additional basis function further reduces the loss function. Boosting has two major algorithms: Adaptive Boosting and Grandient Boosting.

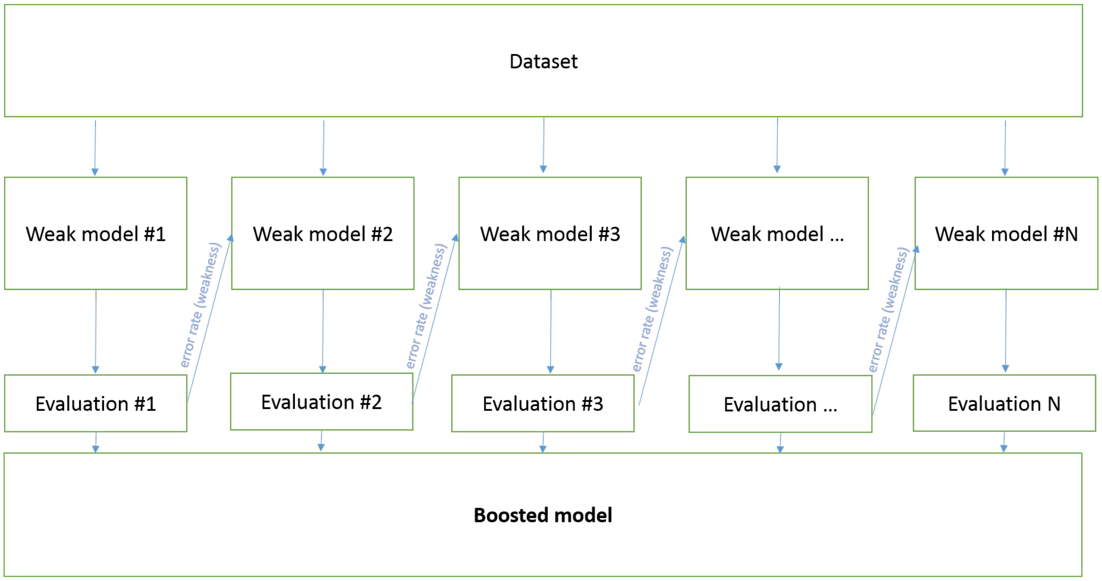


Figure 3.2: Structural overview of Boosting method

Adaptive Boosting (known as AdaBoost) is developed mainly for classification problems, so in this case, the weakness of each learner is the set of misclassified data points [86]. AdaBoost solves this issue with adding increased weights to these points (while decreasing the weight of well-classified items) so that the next weak learner will pay extra attention to putting it to the right class.

Gradient Boosting also adds more weak learners to the system, but does the correction in another way: instead of adding sample weights and tuning them based on the success of classification, it compares the difference between the predicted and the real value coming from the dataset. Originated from this behaviour, Gradient Boosting can be used for both regression and classification problems.

The implementation of R's generalized boosted modeling framework closely follows Friedman's Gradient Boosting Machine [87].

Settings

In my research, for each parameter-combination, a bootstrap based validation with 100 replicates was used on the training set to obtain a reliable estimate of model performance. The tuning parameters of the corresponding package are the following:

- *n.trees*, specifying the total number of trees to fit;
- *interaction.depth*, meaning the maximum depth of each tree;
- *n.minobsinnode*, declaring the minimum number of observations in the terminal nodes of the trees;
- and the *shrinkage* parameter which is applied to each tree in the expansion.

3.2.6 Ensembled modelling

Ensembled modelling is one of the most promising area of machine learning-based predicting. In different domains researchers try to combine the advantages of individual classifiers to produce a strong learner. In the current subsection I summarize the results of some of the most-related articles.

Latha et al. [88] used ensembled modelling on the Cleveland Heart Disease Database to improve the accuracy of prediction of heart disease risk. They used weak classifiers like decision tree (C4.5), Bayesian network, Naïve Bayes, Random forest and neural networks to combine them in different ensembled-based modelling techniques like Boosting, Bagging, Stacking and Majority vote. This comparative analytical approach was done to determine how the ensemble technique can be applied for improving prediction accuracy in heart disease. As a result, a comparison of the various ensembling strategies revealed that the accuracy of the weak classifiers could be increased by a maximum of 7.26%.

Austin el al. found [89] that improvements in the misclassification rate using boosted classification trees were at best minor compared to when conventional classification trees were used. They analysed short-term (30-day) mortality in two cohorts of patients hospitalized with either acute myocardial infarction (N = 16,230) or congestive heart failure (N = 15,848). They observed minor to modest improvements to sensitivity, with only a negligible reduction in specificity.

In another study [90] on the same datasets, Austin el al. evaluated the improvement that is achieved by using ensemble-based methods, including bootstrap aggregation (bagging) of regression trees, random forests, and boosted regression trees. They found that ensemble methods offered substantial improvement in predicting cardiovascular mortality compared to conventional regression trees; but conventional logistic regression models that incorporated restricted cubic smoothing splines had even better performance. An example of ROC AUC values from their study: on the "EFFECT Follow-up" database, their models achieved the following results by ROC AUC: regression tree: 0.767, bagged trees: 0.820, random forest: 0.843, Boosted trees (depth four): 0.852, Logistic regression: 0.852, Logistic regression—Splines: 0.858, Logistic regression—GRACE score: 0.826.

A neural network ensemble method was proposed [91] by Das et al.. Three independent neural networks models were used (Levenberg–Marquardt, scaled conjugate gradient and Pola–Ribiere conjugate gradient algorithms) as primary learners, and the final, ensembling layer combined their results with averaging. The investigated database contained 303 complete samples. Although they didn't published the predictive performance of the individual models, the final model gained 89.01% classification accuracy, 80.95% sensitivity and 95.91% specificity values on the validation dataset.

Subramanian et al. were also focused on heart failure mortality and used partial patient data from the dataset of Vesnarinone Evaluation of Survival Trial [92]. On the data of 963 patients, they established three logistic regression models to predict survival and an ensemble model learned by boosting. On of the major finding of the study is that their ensemble model performed significantly better than the standard approach of logistic regression. As authors discuss, the reason for this significant increase in predictive accuracy is that ”an ensemble of models adjusts better for the biological variability inherent in clinical studies that are derived from patient data.”

Although the previous examples were focusing on heart failure and mortality prediction, researchers gain advantages of ensemble modelling in various fields: Bagging, Random Forests and Extra Trees were used by [93] Petkovic et al. when they addressed the task of feature ranking for hierarchical multi-label classification. Extra Tree is similar to RF, with two main differences: instead of using bootstrap replicas, Extra Trees use the whole original sample; and the selection of cut points is random and not an optimum split, like in RF [94]. Three feature ranking scores like Symbolic, Genie3 and the Random Forest Score were investigated and authors found the first two scores yield relevant feature ranking. In the domain of medical image processing, Tóth et al. [95] described an efficient 3D visualization framework in connection with an ensemble-based decision support system.

Fundamentals of Ensembled Modelling

Ensembled modelling as a strategy based on the idea that if we combine the predictive performance of different classifiers, it can produce a stronger learner. *Bagging* also known as *Bootstrap aggregation*, *Boosting* and *Stacking* are the main classes of ensemble learning methods.

1. In *Bagging*, from the original dataset new datasets (called *bootstrap samples* or *bootstrap replicates*) are selected with replacement; we train the models on each of them; and finally the outcome is calculated with averaging (in case of regression) or majority vote (in case of classification).

2. In *Boosting*, simple, 2-3 level depth trees are used and we build the models trying to predict based on the prediction error of the previous tree. The two types are: ADA Boosting and Gradient Boosting.

3. In *Stacking* different types of individual machine learning models are applied

(*1-st level learner*) and trained on the same, original dataset, then we combine the prediction results of them in an upper level (*meta-learner* or *second-level learner*).

In the list of previously published models, my RF models is similar to *Bagging* category (although there are some differences between RF and bagged models); and our GBM models belong to *Boosting* category: a given number of decision trees were used to construct a final, better learner.

With GBM, I was focusing on *Stacking*, as different types of first-levels learners were used, then I tried to exploit the common predictive power of them in an upper level.

The schematic overview of Stacking is depicted on Figure 3.3. As it shows, there can be any number of 1st-level learners, they are trained on the full, original dataset and produce their ”local predictions”. These different predictions serve as inputs for the meta-learner who attempts to combine these predictions to have the best possible final outcome. As can be seen, the 1-st level learners have to be fully trained and the local predictions have to be made *before* the Meta-learner starts to operate.

In the current study, the 1-st level learners are RF, GBM and NN, while the Meta-learner is Generalized Linear Model, so the ensembled model is a combination of machine learning algorithms and regression models. The modelling structure of the current research is explained and visualized in details in the followings.

Settings

After generating the imputations, training and validation sets were created on each imputations with maintaining the original distribution of the target variables. The trainings were used as the input data of the models (on these, the algorithm performed boosting to find the optimal hyperparameters for the given model); while the validation datasets were used to manually measure the prediction performance in ROC AUC.

ROC AUC was applied to select the optimal parameters using the largest value.

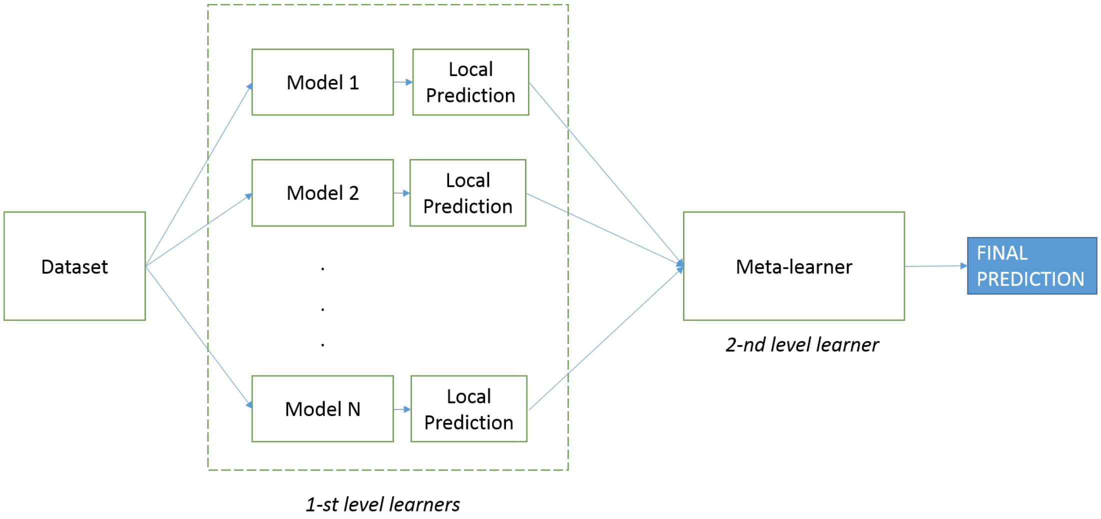


Figure 3.3: Schematic overview of Stacking

For each parameter-combination, a bootstrap based validation with 10 resampling iterations were used on the training set to obtain a reliable estimate of model performance.

The Modelling structure can be visualized in three figures: on Figure 3.4, the full modelling structure is visualized, while the next two figures focus on the separate sections in a more detailed way.

Figure 3.5 depicts the first step: the connection between the original dataset, the imputations, the target variables and the models as inputs of the ensembled models. It contains only one case (RF model) out of the three, but the same processes were performed for GBM and NN as well.

After I finally had all the 5 (number of imputations) * 2 (number of target variables) * 3 (number of model types) = 30 models, I could go on with the ensembling phase. Figure 3.6 depicts the connection between the initial models and the ensembed ones.

3.2.7 Results - 30-day mortality

The 30-day mortality rate for the whole dataset (n = 47,391) was 11.16%.

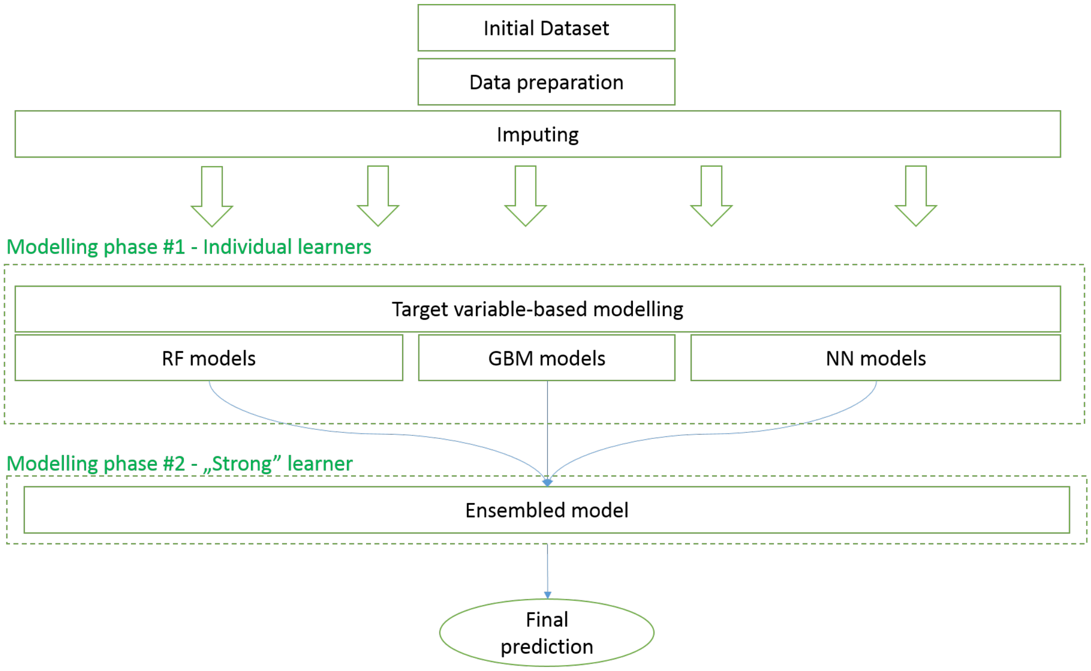

Initial Dataset
Data preparation
Imputing
Modelling phase #1 - Individual learners
Target variable-based modelling
RF models
GBM models
NN models
Modelling phase #2 - „Strong” learner
Ensembled model
Final prediction

Figure 3.4: Ensembled Modelling structure - Overview

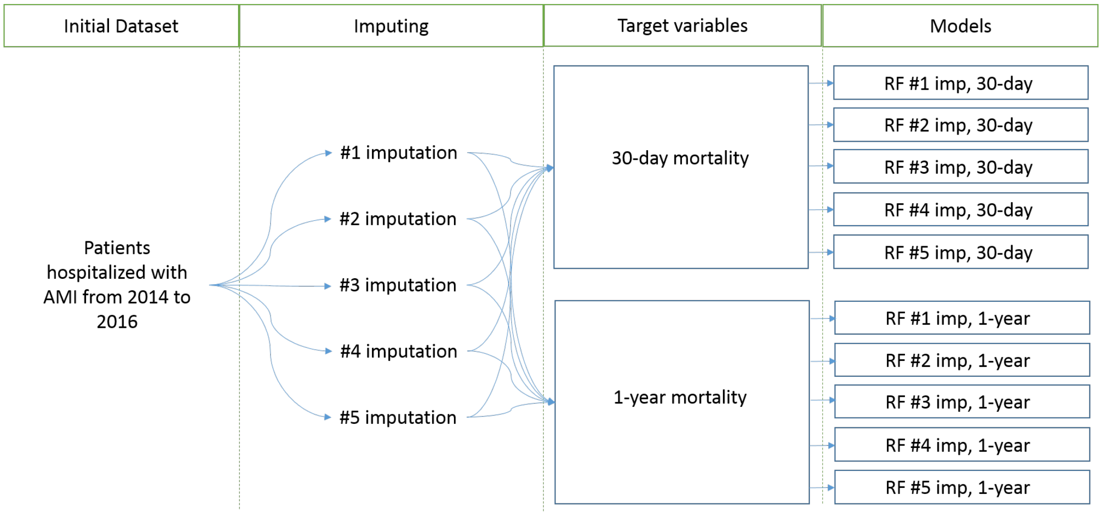


Figure 3.5: Ensembled Modelling structure - Step 1

Regression, Decision Tree, Neural network results

In the training sets, we achieved an average of 0.788 as ROC AUC for decision

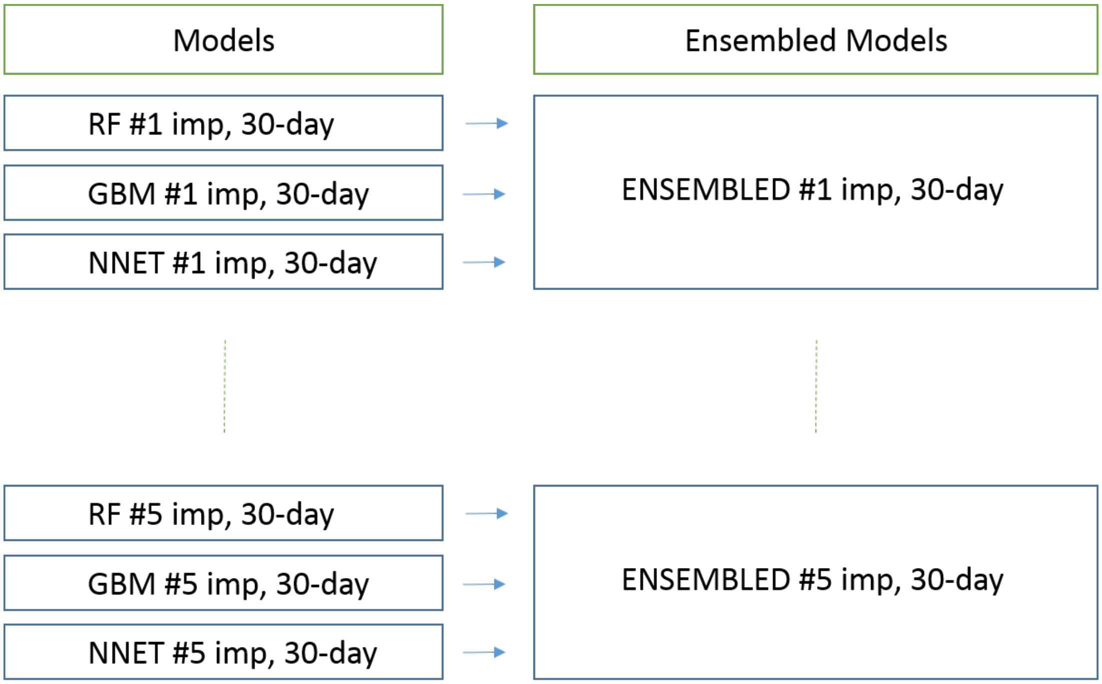


Figure 3.6: Ensembled Modelling structure - Step 2

tree models, 0.837 for neural network models and 0.836 for regression models. In case of neural network and regression models, the standard deviation of the AUC was negligible (<0.001), but the tree models' was 0.025.

Out of the five, neural network performed the best in the third imputation. It reached 0.840 as ROC AUC (95% Confidence Interval (CI), 0.834 - 0.845). In this same imputation, the result of the regression model was 0.836 (95% CI, 0.832 - 0.843) and the decision tree reached 0.783 (95% CI, 0.776 - 0.790).

In the validation sets, an average of 0.774, 0.835 and 0.834 were experienced in case of decision tree, neural network and regression models, respectively.

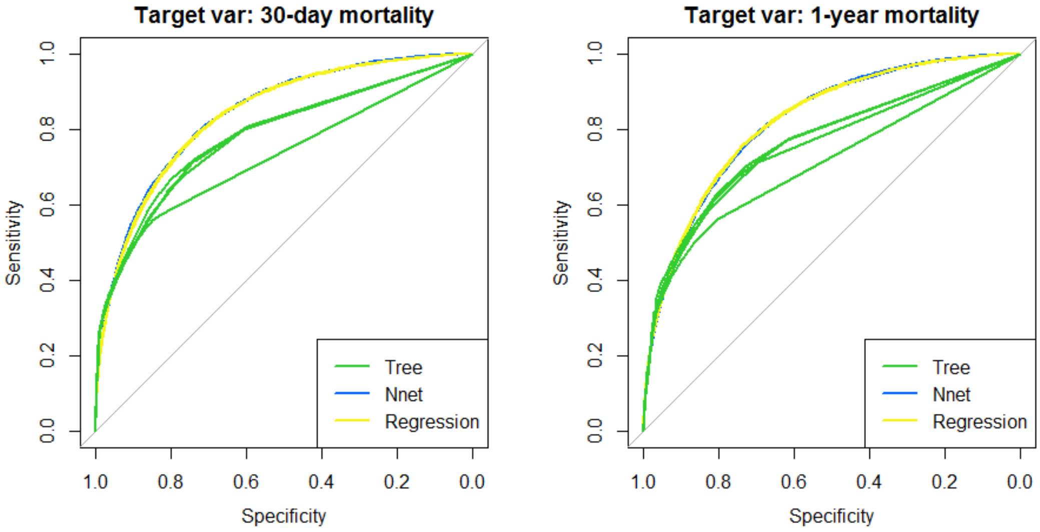


Figure 3.7: Performance of the decision tree, neural network and regression models. Training set. Left: 30-day mortality. Right: 1-year mortality.

Table 3.6: ROC AUC values of the 30-days models, training set

	Imp. #1	Imp. #2	Imp. #3	Imp. #4	Imp. #5
Regression	0.8361	0.8369	0.8356	0.8340	0.8376
Decision tree	0.7810	0.7993	0.7826	0.7856	0.7921
Neural net	0.8369	0.8378	0.8389	0.8348	0.8384

Table 3.6 contains all AUC values for the 30-day mortality models on the training set. Values of Table 3.7 shows the results on the validation set.

At 5% significance level, the differences were non-significant between regression and neural network, but they were significant between both and decision trees (in all imputations). To assess the differences between the methods, pairwise differences of performance measures are calculated across resamplings and checked if they're equal with zero in expected value using a Welch corrected *t*-test with Bonferroni adjustment [96] for multiplicity.

Table 3.7: ROC AUC values of the 30-days models, validation set

	Imp. #1	Imp. #2	Imp. #3	Imp. #4	Imp. #5
Regression	0.8347	0.8328	0.8350	0.8388	0.8319
Decision tree	0.7703	0.7762	0.7771	0.7752	0.7745
Neural net	0.8357	0.8335	0.8337	0.8398	0.8326

Random Forest results

Table 3.8 shows the resulted ROC AUC values of Random Forest models for each imputations in case of 30-day mortality as target variable.

Table 3.8: ROC AUC values of Random Forest models, 30-day mortality

Model Nr.	***Training set***	***Validation set***
#1	0.845	0.850
#2	0.844	0.847
#3	0.846	0.842
#4	0.839	0.854
#5	0.842	0.844

The average for 30-day mortality is 0.843 for training set and 0.847 for validation set. In case of the highest AUC (Model #3) on training set, the best treshold is 0.090 (Specificity: 0.756, Sensitivity: 0.789); while it is 0.103 (Specificity: 0.790, Sensitivity: 0.765) for the highest AUC value on the validation set (Model #4). ROC curves of these best models on training and validation sets are displayed on Figure 3.8.

We can state that, there is no significant difference exists between the predictive power of models trained on different imputations. The RF models represent a stable learner as the standard deviation for the 30-day models are 0.0029 (training) and 0.0047 (validation). These numbers are 0.0021 and 0.0036 for the 1-year models.

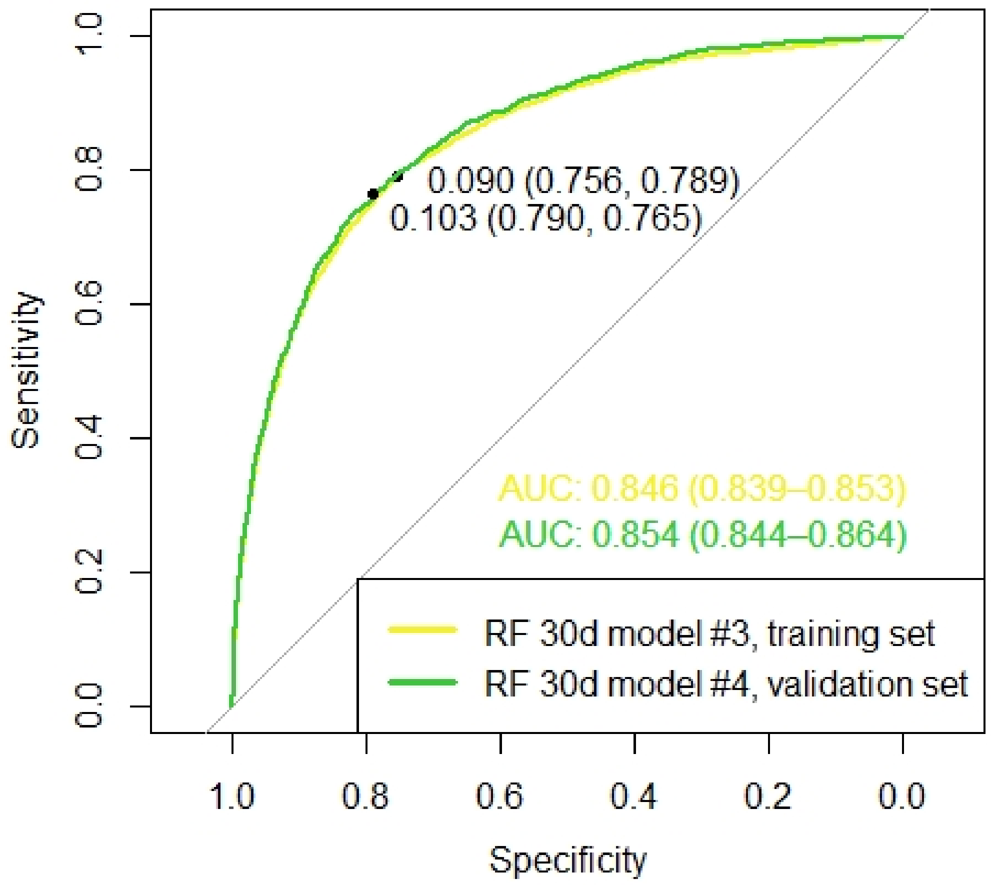


Figure 3.8: ROC curves of the best models on training and validation sets, 30-day

For all 10 models, *3* proved to be the best value for *mtry*, the number of variables randomly sampled as candidates at each split. Next to *3*, with *2* and *5* almost the same result (¡1% difference) was achieved. During parameter-tuning, *ntree* was held constantly at 500, as the default number of small, weak trees to grow. (Here as well, ROC was used to select the optimal model using the largest value.)

In comparison with the decision tree models, we can state that, significant improvement can be reached with Random Forest on the same dataset. Figure 3.9 and 3.10 plots the ROC curves of both models; the AUC on the figures are the highest

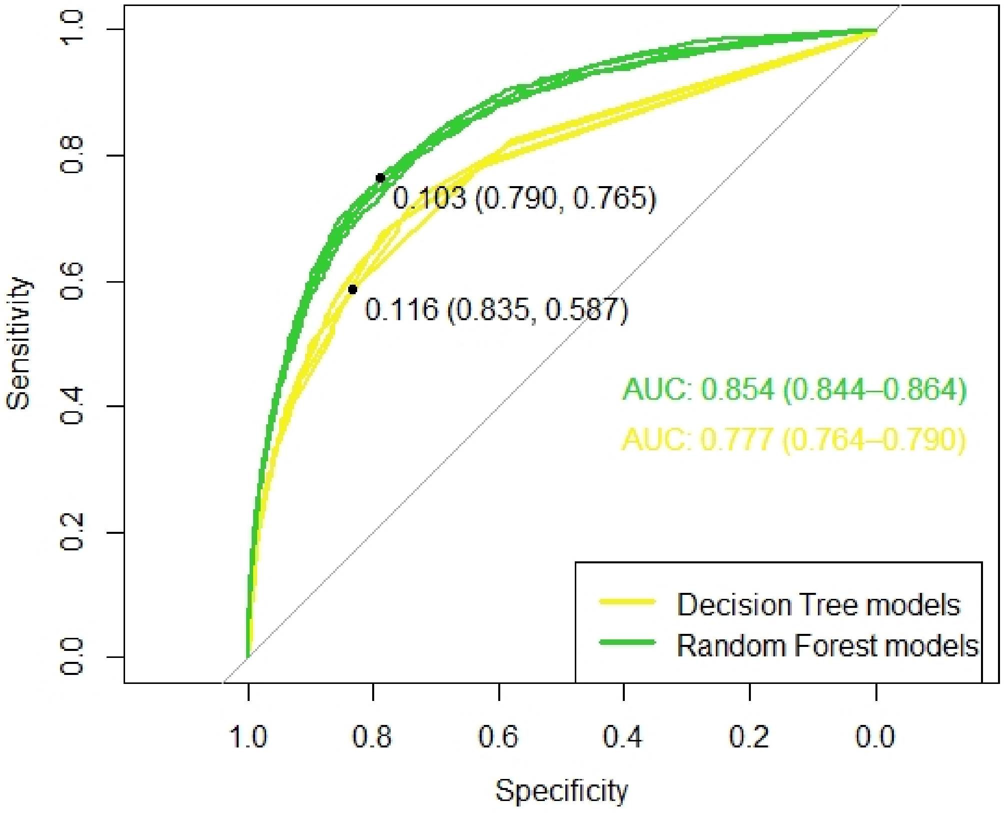


Figure 3.9: ROC curves all Decision Tree and Random Forest models, validation set, 30-day

values from DT and RF models.

The numerical differences between the DT and RF models are the following: improvement of 5.5% and 7.3% (training and validation) for 30-day models; 8.1% and 9.2% (training and validation) when predicting 1-year mortality.

GBM results

Table 3.9 shows the resulted ROC AUC values of GBM models for each imputation in case of 30-day mortality as the target variable.

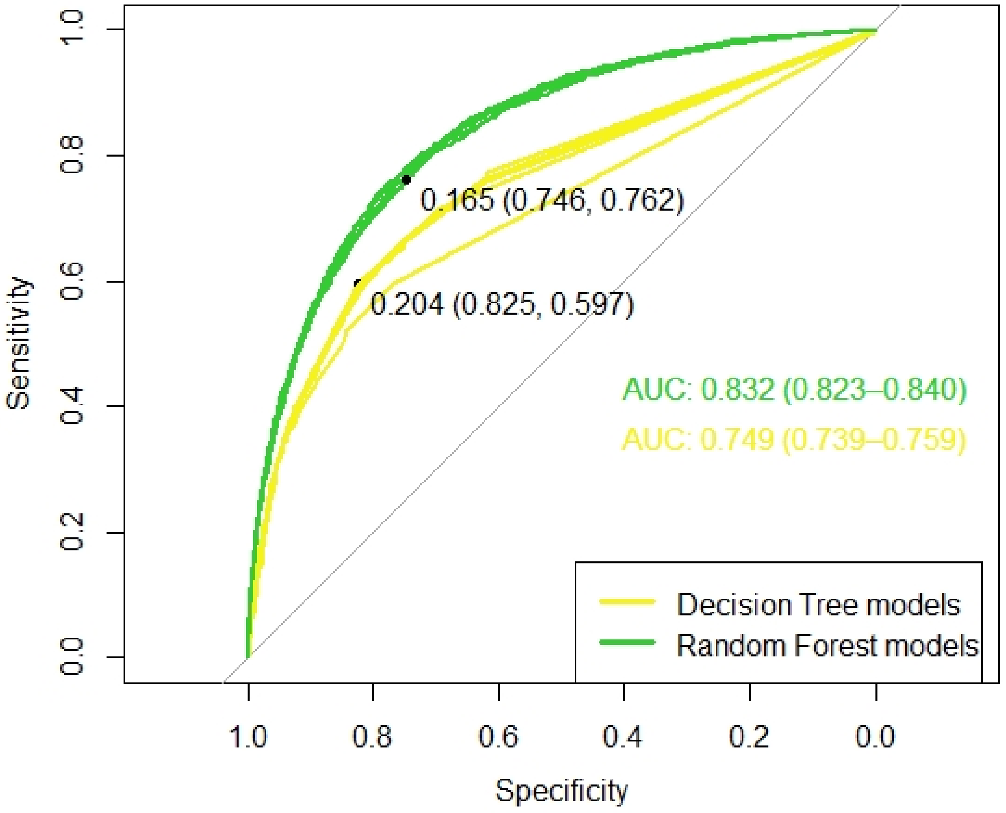


Figure 3.10: ROC curves all Decision Tree and Random Forest models, validation set, 1-year

The average for 30-day mortality is 0.847 for training set and 0.839 for validation set. The highest AUC on the training set is reached by Model #5, here the best threshold is 0.118 (Specificity: 0.800, Sensitivity: 0.734); while it is 0.102 (Specificity: 0.759, Sensitivity: 0.778) on the best-performing model on the validation set (Model #4). ROC curves of these best models on training and validation sets are displayed in Figure 3.11.

Ensembled results

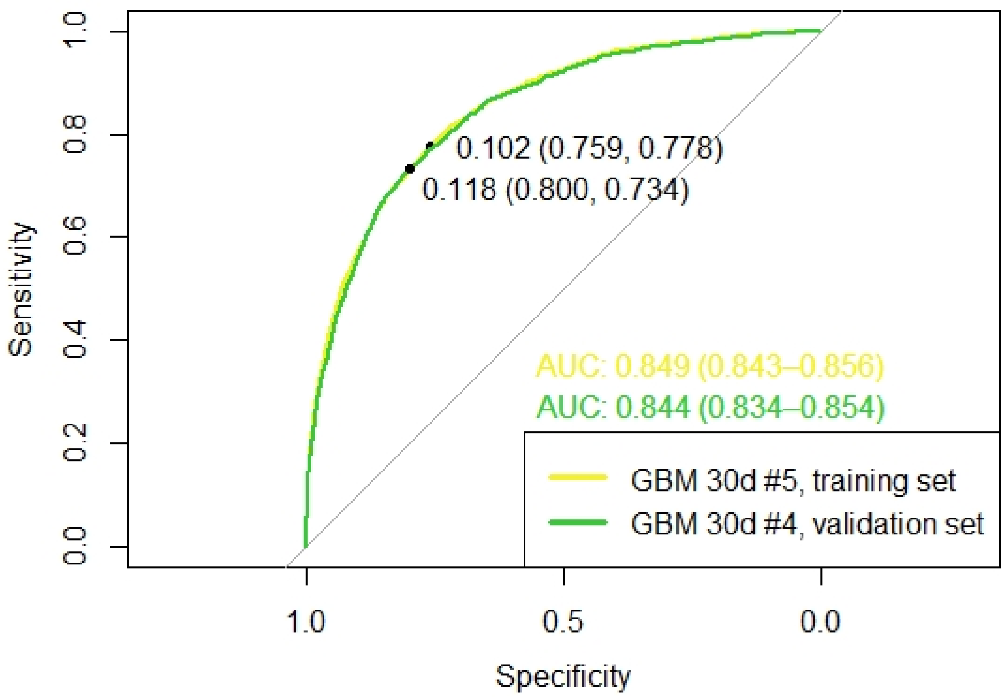


Figure 3.11: ROC curves of the best models on training and validation sets, 30-day

In Table 3.10 I summarized the ROC AUC values of the individual and ensembled models for 30-day mortality. All values were calculated on the corresponding

Table 3.9: ROC AUC values of GBM models, 30-day mortality

Model Nr.	***Training set***	***Validation set***
#1	0.847	0.841
#2	0.848	0.838
#3	0.847	0.837
#4	0.844	0.844
#5	0.849	0.835

validation datasets.

Table 3.10: ROC AUC values of the 30-days models, validation set.

	Imp. #1	Imp. #2	Imp. #3	Imp. #4	Imp. #5	Avg
GBM	0.8411	0.8381	0.8375	0.8443	0.8346	0.8391
RF	0.8499	0.8472	0.8416	0.8528	0.8436	0.8470
NN	0.8358	0.8334	0.8353	0.8398	0.8326	0.8354
Ensembled	0.8592	0.8542	0.8517	0.8602	0.8522	0.8555

Table 3.11: ROC AUC values of the 1-year models, validation set.

	Imp. #1	Imp. #2	Imp. #3	Imp. #4	Imp. #5	Avg
GBM	0.8169	0.8202	0.8251	0.8178	0.8246	0.8209
RF	0.8323	0.8332	0.8392	0.8312	0.8384	0.8349
NN	0.8134	0.8166	0.8234	0.8140	0.8224	0.8180
Ensembled	0.8358	0.8371	0.8439	0.8349	0.8432	0.8390

Figure 3.12 depicts the performance of all the four models in a ROC curve while numerical differences between the methods with 99.2% confidence intervals are shown on Figure 3.13, both for a randomly selected case (30-day mortality as target variable and the first imputation was selected). Table 3.12 reports the standard deviation between the ROC AUC values of the separate models trained on the different imputations.

Table 3.12: Standard deviation of the ROC AUC values of imputations per model type and target variable.

	GBM	RF	NN	Ensembled
30-day models	0.0037	0.0045	0.0028	0.0040
1-year models	0.0038	0.0037	0.0047	0.0043

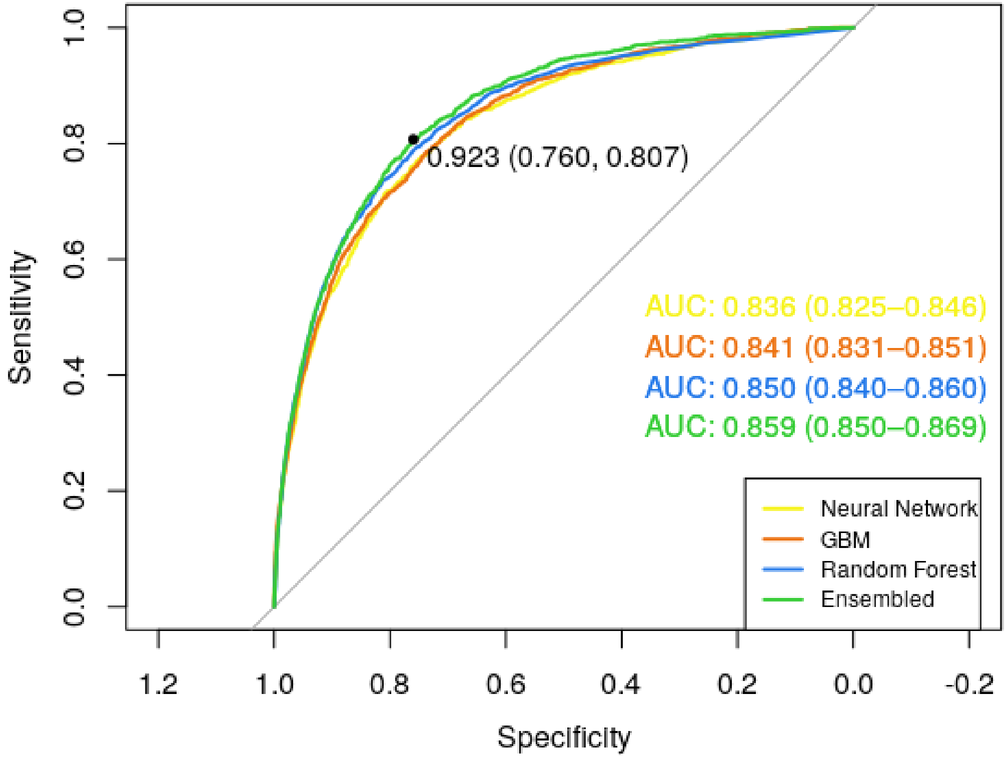


Figure 3.12: Performance of our Neural Network, Random Forest, Generalized Boosted and Ensembled models.

Target: 30-day mortality, dataset: Imputation #1, validation set.

3.2.8 Results - 1-year mortality

The 1-year mortality rate for the whole dataset (*n* = 47,391) was 19.74%.

Regression, Decision Tree, Neural network results

In the training sets, I achieved an average of 0.754 as ROC AUC for decision tree models, 0.8194 for neural net models and 0.8109 for regression models. The standard deviation of AUC was the same as in case of 30-day mortality: for neural net and regression models, it was negligible (<= 0.001), but the tree models' was 0.025.

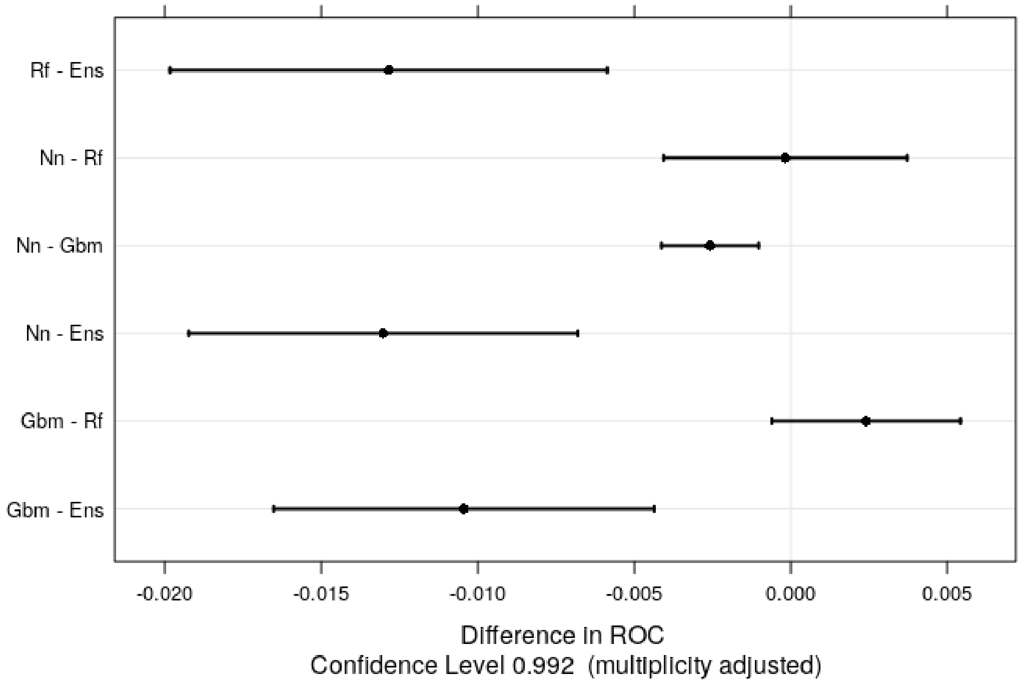


Figure 3.13: Numerical differences between our Neural Network, Random Forest, Generalized Boosted and Ensembled models. Target: 30-day mortality, dataset: Imputation #1, validation set.

Figure 3.14 depicts the performance of the models in all imputations and both target variables, while Table 3.13 contains the corresponding AUC values for the 1-year mortality models on the training set. Values of Table 3.14 shows the results on the validation set.

Again, the differences were non-significant between regression and neural network, but they were significant between both and decision trees.

Table 3.13: ROC AUC values of the 1-year models, training set

	Imp. #1	Imp. #2	Imp. #3	Imp. #4	Imp. #5
Regression	0,8207	0,8200	0,8167	0,8212	0,8167
Decision tree	0,7208	0,7666	0,7512	0,7716	0,7615
Neural net	0,8210	0,8204	0,8170	0,8215	0,8171

Table 3.14: ROC AUC values of the 1-year models, validation set

	Imp. #1	Imp. #2	Imp. #3	Imp. #4	Imp. #5
Regression	0,8128	0,8163	0,8229	0,8138	0,8221
Decision tree	0,7082	0,7525	0,7490	0,7502	0,7567
Neural net	0,8134	0,8166	0,8234	0,8140	0,8224

Random Forest results

Table 3.15 shows the resulted ROC AUC values of Random Forest models for each imputations in case of 1-year mortality as target variable. This means an average of 0.835 on training set and 0.836 on validation set.

Table 3.15: ROC AUC values of Random Forest models, 1-year mortality

Model Nr.	***Training set***	***Validation set***
#1	0.837	0.833
#2	0.837	0.834
#3	0.832	0.839
#4	0.836	0.832
#5	0.834	0.840

In case of the highest AUC (Model #1) on validation set, the best threshold is 0.157 (Specificity: 0.730, Sensitivity: 0.797); while it is 0.165 (Specificity: 0.747, Sensitivity: 0.780) for the highest AUC value on the validation set (Model #5). ROC

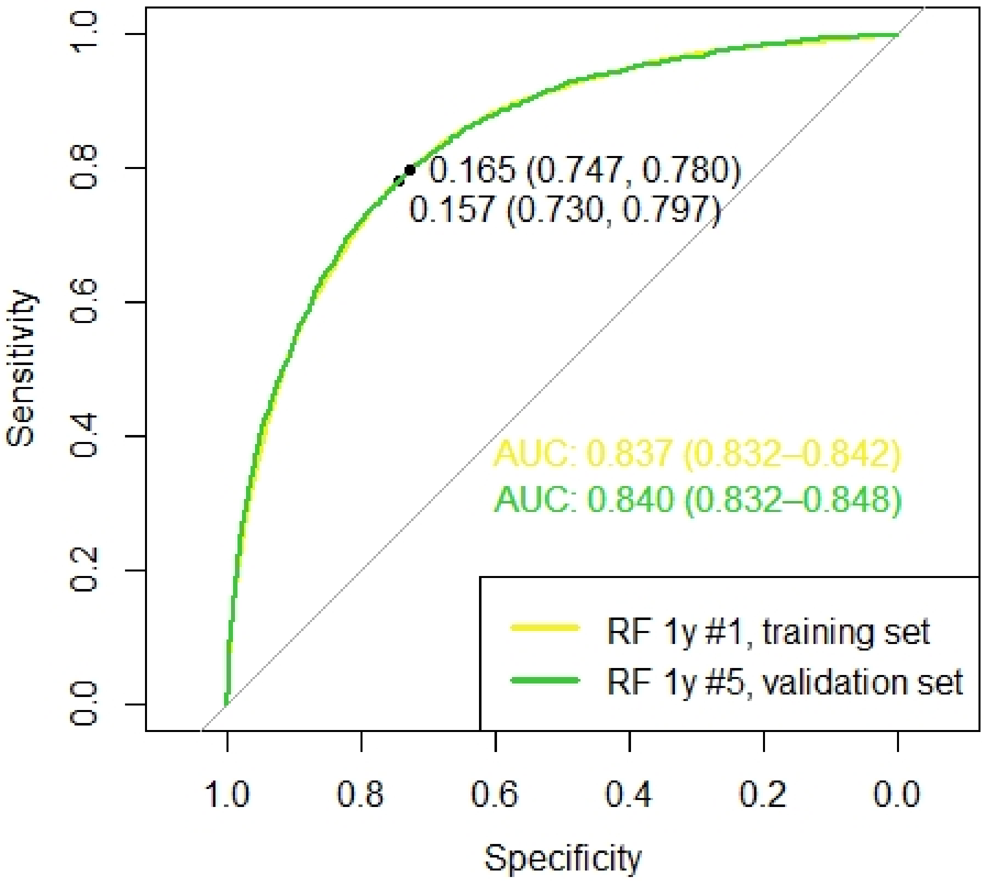


Figure 3.14: ROC curves of the best models on training and validation sets, 1-year

curves of these best models on training and validation sets are displayed on Figure 3.14.

GBM results

Table 3.16 shows the resulted ROC AUC values of GBM models for each imputation in case of 1-year mortality as the target variable. This means an average of 0.828 on the training set and 0.821 on the validation set.

The highest AUC on the training set is reached by Model #1, here the best threshold

Table 3.16: ROC AUC values of GBM models, 1-year mortality

Model Nr.	***Training set***	***Validation set***
#1	0.829	0.817
#2	0.829	0.820
#3	0.826	0.825
#4	0.829	0.818
#5	0.825	0.825

is 0.167 (Specificity: 0.708, Sensitivity: 0.791); while it is 0.200 (Specificity: 0.760, Sensitivity: 0.742) on the best-performing model on the validation set (Model #5). ROC curves of these best models on training and validation sets are displayed in Figure 3.15.

We used grid search to find the best possible values for the model's hyperparameters. Our final models reaching the best predictive power used *20.000* for the total number of trees to fit; *2* as the maximum depth of each tree; *0.001* as the shrinkage parameter; and the best value for the minimum number of observations in the terminal nodes of the trees proved to be *20*.

Ensembled results

In Table 3.11 I summarized the ROC AUC values of the individual and ensembled models for 1-year mortality.

Table 3.12 reports the standard deviation between the ROC AUC values of the separate models trained on the different imputations.

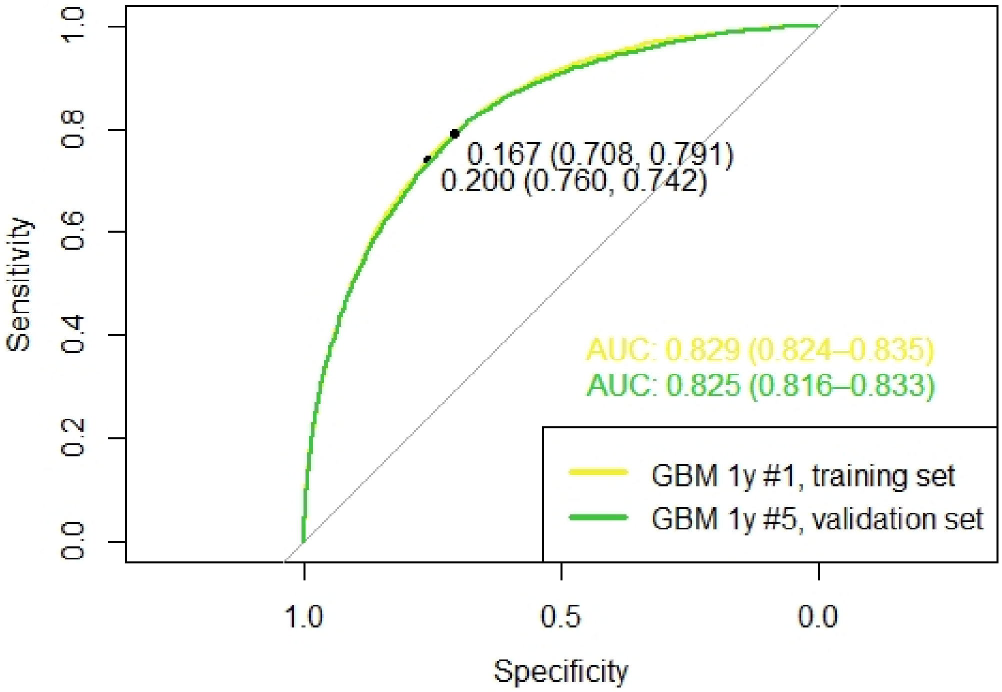


Figure 3.15: ROC curves of the best GBM models on training and validation sets, 1-year

3.2.9 Results - Variable Importance

Variable importance, in general, refers to a measure of how much a model uses a given variable to make accurate predictions. In this subsection, I deal with the variable importance values for each individual and the ensembled model.

Since the definitions and the methods of calculating the variable importance in separate model types differ, instead of listing the exact feature importance values for each model type, relative importance is used: the position of the given feature on the list of the most important fields. With using relative importance it becomes possible to compare the most important features of different models types, i.e. we can make a global order between the variables over the different models.

As multiple imputations were used, I aggregated variable importance values in the imputations in the following way: summed up all the relative importance values for each field for a given target variable and for a given model type, then divide this value by the number of imputations. The resulted value represents the relative importance of the given feature, and in this number, all the imputations added their effects.

The aggregated and relative values of feature importance in descending order for the 30-day models are the following:

1. GBM: Cardiogenic shock (36.3), Age (21.1), Abnormal level of creatinine (10.4), Percutaneous Coronary Intervention (6.7), Prehospital reanimation (6.6)

2. Random Forest: Age (31.1), Cardiogenic shock (14.2), Smoking = never (13.5), Smoking = quit (13.3), Hyperlipidaemia (6.6)

3. Neural net: Age (19.8), Cardiogenic shock (15.2), Percutaneous Coronary Intervention (9.6), Abnormal level of creatinine (9.1), Prehospital reanimation (7.4)

4. Ensembled: Age (26.1), Cardiogenic shock (15.7), Smoking = never (8.6), Smoking = quit (8.4), Abnormal level of creatinine (7.4)

The aggregated and relative values of feature importance in descending order for the 1-year models are the following:

1. GBM: Age (34.1), Cardiogenic shock (16.9), Abnormal level of creatinine (11.9), Percutaneous Coronary Intervention (10), Heart failure (7.8)

2. Random Forest: Age (36.6), Smoking = never (12.4), Smoking = quit (12), Cardiogenic shock (8.2), Abnormal level of creatinine (6.5),

3. Neural net: Age (23.6), Cardiogenic shock (10.9), Percutaneous Coronary Intervention (10.7), Abnormal level of creatinine (9.1), Prehospital reanimation (7)

4. Ensembled: Age (30.8), Cardiogenic shock (10.6), Abnormal level of creatinine (8.3), Percutaneous Coronary Intervention (7.3), Smoking = never (7.3), Smoking = quit (8.4), Abnormal level of creatinine (7.4)

3.2.10 Conclusions

In the current thesis, I investigated if there is a competitive opponent for the mostly used and trusted regression in mortality prediction on a realworld, unfiltered dataset with AMI patients. The target variables were the 30-day and the 1-year mortality. After applying several steps in the phase of data preparation and used 5 imputations to finalize the dataset - the average of the corresponding ROC AUC values of the selected models were compared against the others.

Result achieved with feed-forward Neural network, Random Forest, GBM and Ensembled techniques were similar to regression's; and in some cases they even exceeded the regression results by a few percent. Decision tree was not able to gain on this level and was in the last place in case of all imputations and both target variables. The best performer among all models was the Ensembled one, which uses the advantages of both the classical regression and artificial intelligence-based solutions.

All of the three models perform better in case of 30-day mortality than 1-year mortality. This difference can be partly accounted for the happenings between the 30 days and the 1 year: for instance, if the patient took the medications or not, if the patient had an operation during that period or not – these all can affect the prognosis.

There are minimal differences between the models of the same type and same target variable: the imputed data doesn't affect the power of models significantly.

Thesis 2:

> ***Thesis 2***
>
> **I have developed machine learning models for predicting 30-day and 1-**

year mortality that met and in some cases exceeded the predictive capabilities of regression.

Relevant own publications pertaining to this thesis:
[P-3] [P-4] [P-5] [P-6]

3.3 Comparing resampling methods (Thesis 3)

3.3.1 Introduction

In a comparative study I have investigated if a difference exists in the predictive power of decision tree models tuned with different resampling methods. K-fold cross validation, repeated cross validation and bootstrap were used to find the optimal parameters for each model on the dataset of Hungarian Myocardial Infarction Registry. The target variable was the 1-year mortality and the differences were measured in 10 different cases with different number of records on randomly selected, real-world datasets from our original HUMIR dataset.

3.3.2 Data structure

I defined 10 dataset sizes to simulate and examine the differences between datasets with variant number of records. The increasing number of records are 300, 500, 1 000, 1 500, 2 000, 4 000, 5 000, 10 000, 15 000 and 18 000.

After selecting 10 samples (without replacement) for each dataset size, multiple imputation was applied resulting in 5 differently imputed datasets for each sub-dataset. Then, each imputed dataset was used with each resampling method to train the decision tree model. Figure 3.16 shows the structure of sub-datasets, imputations, resampling methods and models.

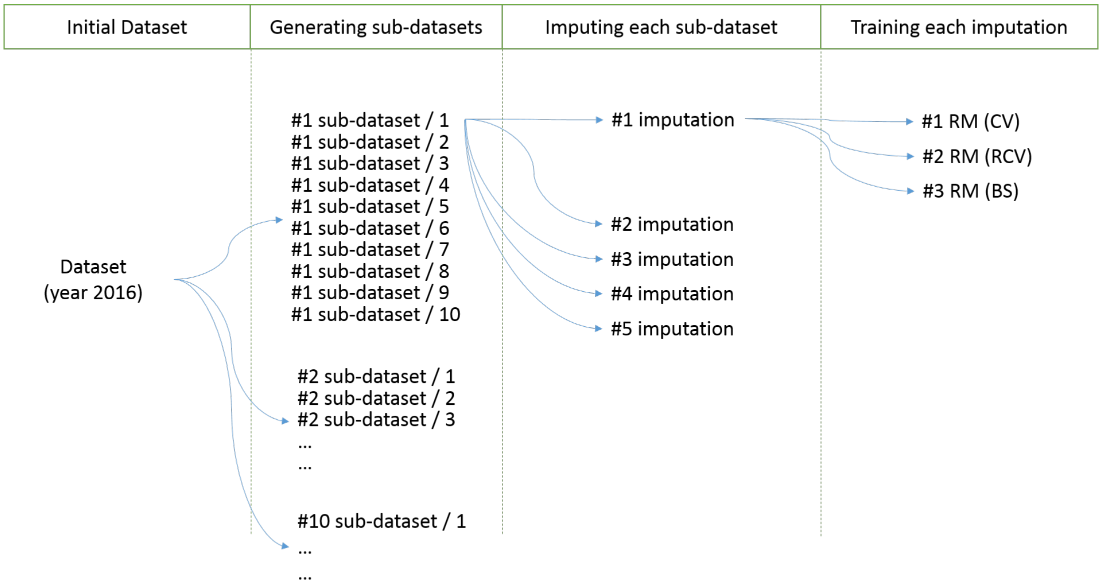


Figure 3.16: The structure of datasets and resampling methods.

I was using the metric of area under (AUC) Receiver Operating Characteristic (ROC) curve to find the best parameters for the decision tree model in case of each resampling method.

Each imputation was splitted into two parts: 80% of the data was used to generate and train the models using the different investigated resampling methods (training set) and 20% was used to check the generalization of the models (validation set). I used a method for splitting which keeps the distribution of the target variable the same as in the original dataset in case of each training and validation datasets.

3.3.3 Results

Figure 3.17 exactly shows how the models trained on different number of records perform in case of a given resampling method. For example, with the smallest sub-datasets (n = 300), the 10 different ROC AUC values (represented by dots) for BS, CV and RCV can be read from the first plot.

Mean is represented by a cross and it depicts a more general approach of the results: it shows the average of ROC AUC values of models trained on the sub-datasets with the same size. It can be read, that out of the 10 cases:

1. the RCV method is in the first place in all cases
2. the CV method is in the second place in 9 cases
3. the only case when the order between CV and BS is changed is the third ($n = 1000$)

Although the numerical differences between the performance of models with different resampling methods are small (between RCV and BS, the average of the differences is 0.0105, between RCV and CV, it's 0.0031), an order can be establised between them. In this dataset, RCV outperforms the other two resampling methods. CV was in the second, while BS was in the third place in 90% of the cases.

Figure 3.17 shows two other facts as well:

1. in case of all resampling methods, the ROC AUC averages are continuously growing – as we are having bigger datasets, the models get "better"
2. as we are heading to bigger datasets, the deviation of a given method's points are continuously getting smaller – as we are having bigger datasets, the models get more accurate and more reliable

As conclusion, we can state that, in the investigated dataset, repeated cross validation slightly outperforms cross validation and both have significantly better results than models trained with bootstrap method. And, as we are having bigger datasets, the predictive power of the models – regardless of the applied resampling method – improves: the ROC AUC values continuously get higher with a smaller deviation.

Thesis 3:

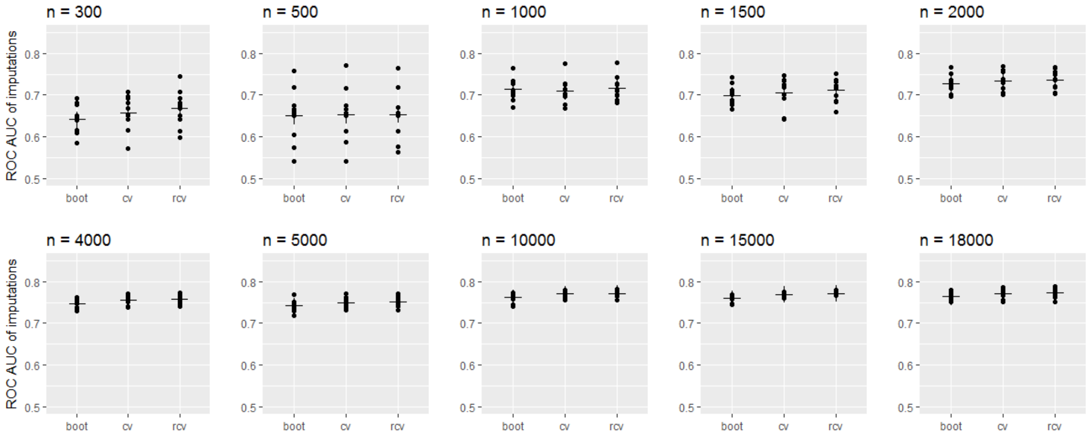


Figure 3.17: ROC AUC values and means of imputations (marked with crosses) for each sub-dataset

Thesis 3

I have showed that in the case of decision trees, there are minimal differences between the resampling methods used to determine the tuning parameters; and these differences disappear with a larger data set (n > 15000).

Relevant own publication pertaining to this thesis:

[P-7]

4 Application

4.1 Aim of the development

For demonstrating a possible clinical application of the research results, I have developed a web-based application where visitors can check the prediction capability of the given model. In short, through the application physicians can enter the patient data, click on a button, than the predicted possibilities are shown. In the background, in the process of prediction, the application is communicating with the original (R-environment-based) modelling infrastructure to use originally developed models to predict 30-days and 1-year mortality as outcomes.

4.2 Application structure

As it can be seen on Figure 4.1, the application structure consists of three layers:

1. User interface
2. Server-side application
3. Modelling environment

These three layers are in connection with each other to accomplish the whole mechanism from data collection until the display of the predicted values.

The role and function of the three layers are the follows:

1. The *User interface* is the public part of the software. It is a two-pager website where visitors can fill in all the fields which will be used to predict the outcome (called *Form screen*), and then see the predicted results (called *Result screen*). The full and detailed list of requested fields are shown in *Section 4.3*.

2. The *Server-side layer* receives the submitted field values, sends it to the third layer via *Command Line Interface*, waiting until it predicts the outcome, receives it, and finally displays the exact numbers

3. The prediction is made in the *Modelling environment*. It is our R-language based environment which processes the received fields, loads one of our previously developed predictive model, predicts the 30-day and 1-year mortality outcomes, then sends it to the Server-side layer

4.3 Requested Fields

On the *User interface* layer of the application, fields are separated based on the groups defined in *Section 3.1.4*. These fields are the predictors of the model, so they must have in the same format as they were used earlier in the modelling phase. For this reason, *dropdowns* (list of pre-defined values visitors can choose from) were applied everywhere where it was possible.

Application fields of *Group 1: General information about the patient* are shown on *Table 4.1*, fields of *Group 2: Previous medical history* are on *Table 4.2*, while *Table 4.3* lists the fields of *Group 3: Information about the pre- and in-hospital treatment*. For each field, next to the name, the field type, the possible values are shown as well, together with the information that if it is mandatory to fill in the given field to start the prediction.

Table 4.1: Fields of the developed application, Group 1

Field	**Field type**	**Possible values**	**Mandatory**
Age	Number	Integer only	Yes
Gender	Dropdown	Male or Female	Yes

From the full list of *Group 1* attributes, *Event ID*, *Patient ID* and *ZIP code* were eliminated in the process of data preparation with absolutely no usage; *If the patient alives* and *Date of death* are parts of mortality as target variable so they cannot be used in the modelling formula. As result, *Gender* and *Date of birth* (in the format of *Age*) were kept as input fields in the application on the *Form screen*.

Table 4.2: Fields of the developed application, Group 2

Field	**Field type**	**Possible values**	**Mandatory**
Myocardial Infarction	Dropdown	Yes or No	Yes
Heart failure	Dropdown	Yes or No	Yes
Hypertension	Dropdown	Yes or No	Yes
Stroke	Dropdown	Yes or No	Yes
Diabetes mellitus	Dropdown	Yes or No	Yes
Peripheral artery disease	Dropdown	Yes or No	Yes
Hyperlipidaemia	Dropdown	Yes or No	Yes
Cardiogenic shock	Dropdown	Yes or No	Yes
Smoked ever	Dropdown	Yes or No	Yes
Smoking at the moment	Dropdown	Yes or No	Only if Smoked ever is *Yes*

Each attribute from the full list of *Group 2* fields were built in as input fields in the application on the *Form screen*.

Since *Level of creatinine* and *Treatment ID* were eliminated from the attributes of *Group 3* in the process of data preparation; and the application automatically populates the current date for *Date of admission*, 4 fields left from the full list of *Group 3* as input fields on the *Form screen* of the application.

Table 4.3: Fields of the developed application, Group 3

Field	Field type	Possible values	Mandatory
Prehospital reanimation	Dropdown	Yes or No	Yes
Percutaneous Coronary Intervention	Dropdown	Yes or No	Yes
Diagnosis	Dropdown	STEMI or NSTEMI	Yes
Creatinine level	Dropdown	Normal or Abnormal	Yes

4.4 Methods & Tools

The application was developed in R and PHP programming languages: *User interface* and *Server-side application* was developed in PHP, HTML, Javascript and CSS; while the *Modelling environment* was the original R language-based environment which were used in the whole research to build the models. PHP was responsible for the *business logic*, data receiving and sending, and HTML, Javascript and CSS were used to build up the screens.

As the official documentation says, ”PHP (recursive acronym for PHP: Hypertext Preprocessor) is a widely-used open source general-purpose scripting language that is especially suited for web development and can be embedded into HTML” [97]. Although PHP is very popular (researches say that in February 2023, around 77.6% of all the websites use PHP as server-side programming language [98]), it could have been possible to use other languages as server-side language. Behind the selection of PHP, the main reason was author’ several-years experience with it.

The current version of the application uses Command Line Interface as communication layer between the Server-side and the Modelling layer. The following two facts made possible to interact between the two layers:

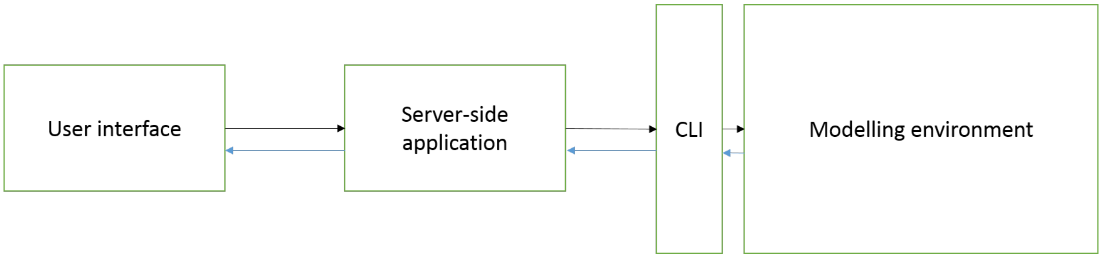


Figure 4.1: Structure of the developed application

1. R scripts can be called through command line with calling the previously installed *Rscript* binary file

2. PHP can execute commands with its *eval* or *shell_exec* function and can gather the returned value

4.5 Results

The application is fully developed and working as it is described in this section. It uses the previously published General Boosted Model to predict the 30-day and 1-year mortality for field values added on the User interface layer. *Figure 4.2* and *Figure 4.3* shows to two screenshots about the User interface: the first is the form and the second is the result screen.

4.6 Examples

To give some examples, here I list three scenarios, field values and the returned 30-day and 1-year mortality results.

4.6.1 Example #1

The entered values of each input field of Example #1 and the resulted probabilities for 30-day and 1-year mortality can be seen on Table 4.4.

Table 4.4: Application - Example #1

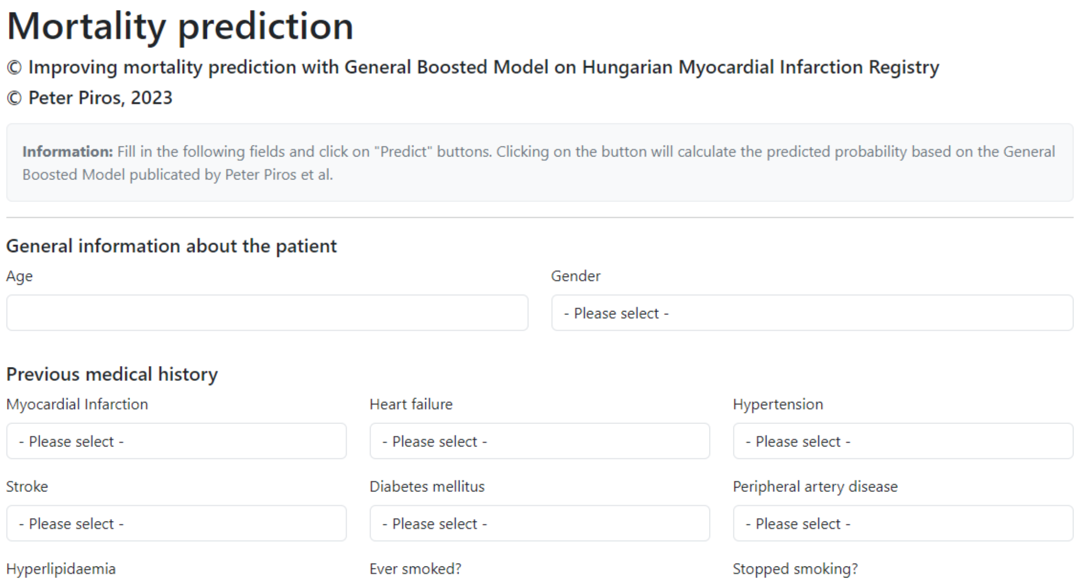
Field	**Field value**
Age	65
Gender	Male
Myocardial Infarction	Yes
Heart failure	Yes
Hypertension	No
Stroke	No
Diabetes mellitus	No
Peripheral artery disease	No
Hyperlipidaemia	No
Cardiogenic shock	No
Ever smoked?	Yes
Stopped smoking?	Yes
Percutaneous Coronary Intervention	Yes
Prehospital reanimation	Yes
Diagnosis	STEMI
Abnormal level of creatinin	No
Result - 30-day mortality	15.09%
Result - 1-year mortality	29.79%

4.6.2 Example #2

The entered values of each input field of Example #2 and the resulted probabilities for 30-day and 1-year mortality can be seen on Table 4.5.

Table 4.5: Application - Example #2

Field	**Field value**
Age	70
Gender	Male
Myocardial Infarction	Yes
Heart failure	Yes
Hypertension	Yes
Stroke	Yes
Diabetes mellitus	Yes
Peripheral artery disease	Yes
Hyperlipidaemia	Yes
Cardiogenic shock	Yes
Ever smoked?	No
Percutaneous Coronary Intervention	Yes
Prehospital reanimation	Yes
Diagnosis	STEMI
Abnormal level of creatinin	Yes
Result - 30-day mortality	51.65%
Result - 1-year mortality	63.49%


Mortality prediction

© Improving mortality prediction with General Boosted Model on Hungarian Myocardial Infarction Registry

© Peter Piros, 2023

Information: Fill in the following fields and click on "Predict" buttons. Clicking on the button will calculate the predicted probability based on the General Boosted Model publicated by Peter Piros et al.

General information about the patient

Age:

Gender: - Please select -

Previous medical history

Myocardial Infarction: - Please select -

Heart failure: - Please select -

Hypertension: - Please select -

Stroke: - Please select -

Diabetes mellitus: - Please select -

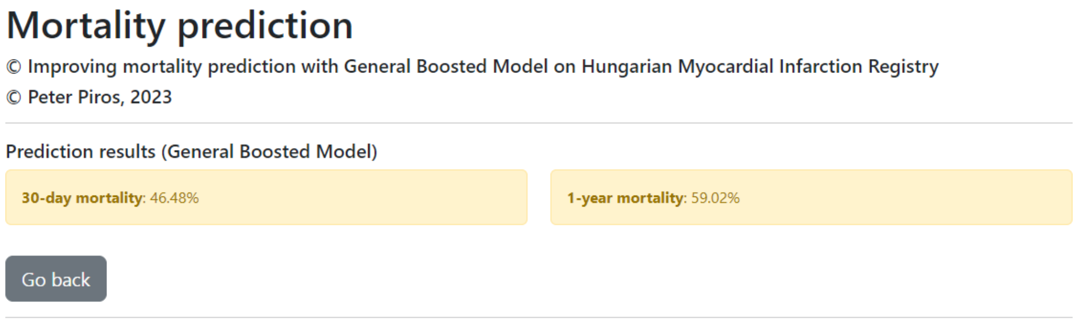
Peripheral artery disease: - Please select -

Hyperlipidaemia

Ever smoked?

Stopped smoking?

Figure 4.2: Screenshot of the developed application, Form screen


Mortality prediction

© Improving mortality prediction with General Boosted Model on Hungarian Myocardial Infarction Registry

© Peter Piros, 2023

Prediction results (General Boosted Model)

30-day mortality: 46.48%

1-year mortality: 59.02%

Figure 4.3: Screenshot of the developed application, Result screen

4.6.3 Example #3

The entered values of each input field of Example #3 and the resulted probabilities for 30-day and 1-year mortality can be seen on Table 4.6.

Table 4.6: Application - Example #3

Field	**Field value**
Age	80
Gender	Male
Myocardial Infarction	Yes
Heart failure	Yes
Hypertension	Yes
Stroke	Yes
Diabetes mellitus	Yes
Peripheral artery disease	Yes
Hyperlipidaemia	Yes
Cardiogenic shock	Yes
Ever smoked?	Yes
Stopped smoking?	No
Percutaneous Coronary Intervention	Yes
Prehospital reanimation	Yes
Diagnosis	STEMI
Abnormal level of creatinin	Yes
Result - 30-day mortality	68.3%
Result - 1-year mortality	83.66%

4.7 Limitations & Solutions

Although the web-based application demonstrates a possible clinical application of the research results in the current thesis, it is ready to use by physicians - the current version has some limitations. Solving these limitations would make the software more prepared to real-life situations.

One limitation is behind the communication of the *User interface* and the *Server-side* layers. In the current version, it is managed through *Command Line Interface* which solves this communication task - but only in the case if these two layers are on the same physical machine. Our server-side code uses *shell_exec* function to call the *Modelling environment*, but this function cannot be used to call other resources located on other physical server. To solve this issue, the communication could be made via HTTP REST API, another interface which makes it possible for two separated machines to securely communicate with each other via internet. Using HTTP REST API it would be possible to separate *Server-side* and *Modelling environment* layers, making the whole software infrastructure much more reliable. In this case, we would have a simple website for *User interface* and the *Modelling layer* could be a dedicated environment, like I had an Amazon instance in the process of model development (see *Section 2.4*).

Another limitation is with the usage of *Modelling layer*. In the R code, first the submitted data is received, the given model object is loaded, the prediction is performed, then we send the results back to the *Server-side layer*. In this process, the loading of the given model object takes the most time. In my testing environment (Intel Core i3 processor, 12 GB memory) loading a GBM model takes about 15 seconds, so to load both the 30-day and the 1-year model, it is around 30 seconds. The size of a GBM model object is around 8MB (this is the smallest, that's why I choosed this one for testing), but the Ensembled's is 386MB, and Regressions' is 243MB. This relatively large files sizes makes it extremely slow the loading process, so to transferring the whole software to a bigger instance would speed up the whole process.

Actually, these two mentioned limitations are in connection with each other: separating *User interface* and *Server-side* layers with transferring *Modelling layer* into a server with higher computation resources would make the software much more reliable and faster.

5 Thesis list

Thesis group 1: In a methodological approach, I have discussed and analyzed the data preparation of artificial intelligence algorithms on the dataset of the Hungarian National Myocardial Infarction Register.

Thesis 1.1

At the international level, I have took a look at the official registries that collect cardiovascular data; within this, I have highlighted the uniqueness of the Hungarian National Myocardial Infarction Register and gave a ”literature recipe” for the use of artificial intelligence methods.

Thesis 1.2

For the dataset of the Hungarian Myocardial Infarction Registry, I have developed a data preparation procedure, with which the raw data became suitable for participating in the implementation of the machine learning process as input for the predictive models.

Relevant own publications pertaining to this thesis group:

[P-1] [P-2] [P-3]

Thesis 2:

Thesis 2

I have developed machine learning models for predicting 30-day and 1-year mortality that met and in some cases exceeded the predictive capabilities of regression.

Relevant own publications pertaining to this thesis:

[P-3] [P-4] [P-5] [P-6]

Thesis 3:

Thesis 3

I have showed that in the case of decision trees, there are minimal differences between the resampling methods used to determine the tuning parameters; and these differences disappear with a larger data set (n $>$ 15000).

Relevant own publication pertaining to this thesis:

[P-7]

6 Conclusions

This dissertation has presented several practical approaches for the usage of machine learning models in medicine, especially in mortality prediction. During my research, I have been working continuously with a dataset originated from the Hungarian Myocardial Infarction Registry (HUMIR); on a full, unfiltered extract from 2014 to 2016, containing 47,412 patients hospitalized with acute myocardial infarction.

In the first thesis group, I have reviewed the Hungarian and three ongoing European myocardial projects, then developed a way how the Hungarian dataset can be transformed to be able to use it as inputs for machine learning algorithms.

Then, as stated in the second thesis, I have developed several machine-learning models based on Decision Tree, Neural Networks, Logistic Regression, Random Forest, Generalized Boosted Model and Ensembled algorithms to predict 30-day and 1-year mortality on the same HUMIR-dataset. The main challange I was facing was if there is a competitive opponent for the mostly used and trusted regression in the world of machine learning algorithms.

The published models achieved, and in some cases even exceeded, the predictive capabilities of regression generally (and rightfully) accepted in the field – thus demonstrating the raison d'être of machine learning solutions in the scientific field and in the Hungarian register.

The third thesis has presented a more specific question: I have stated that there is no significant difference in the predictive power of the decision tree models tuned with different resampling methods on the data of the Hungarian register.

In Section 4, I have presented how the research will continue: as an applied informatics area, it is important to build the knowledge back to medicine. For demonstrating a possible clinical application of the research results, I have developed a web-based application where visitors can check the prediction capability of the given model. With submitting the patient data, the predicted 30-day and 1-year mortality possibilities for the given models are shown in a few moments.

Although the developed application demonstrates a possible clinical use for the researches I have made, there is still room for improvements: separating *Server-side* and *Modelling environment* layers with HTTP REST API would be a desirable solution as it can make the whole software infrastructre more flexible and fault-tolerant. Another improvement would be a bigger hardware instance with more resources.

Bibliography

References

[1] Connie W Tsao, Aaron W Aday, Zaid I Almarzooq, Alvaro Alonso, Andrea Z Beaton, Marcio S Bittencourt, Amelia K Boehme, Alfred E Buxton, April P Carson, Yvonne Commodore-Mensah, et al. "Heart disease and stroke statistics—2022 update: a report from the American Heart Association". In: *Circulation* 145.8 (2022), e153–e639.

[2] Robert Beaglehole. "International trends in coronary heart disease mortality and incidence rates". In: *Journal of cardiovascular risk* 6.2 (1999), pp. 63–68.

[3] George A Kaplan and Julian E Keil. "Socioeconomic factors and cardiovascular disease: a review of the literature." In: *Circulation* 88.4 (1993), pp. 1973–1998.

[4] Johan Hallqvist, Michael Lundberg, Finn Did erichsen, and Anders Ahlbomb. "Socioeconomic differences in risk of myocardial infarction 1971–1994 in Sweden: time trends, relative risks and population attributable risks". In: *International Journal of Epidemiology* 27.3 (1998), pp. 410–415.

[5] M Pearson. "Lessons from the management of acute myocardial infarction". In: *Heart* 91.suppl 2 (2005), pp. ii28–ii30.

[6] *Cleveland Clinic website*. Accessed 2023.01.23 14:25. URL: `https://my.clevelandclinic.org/health/diseases/16818-heart-attack-myocardial-infarction`.

[7] Lung National Heart, Blood Institute, et al. "What are the signs and symptoms of coronary heart disease". In: *Bethesda: National Heart, Lung, and Blood Institute* (2016).

[8] Grant W Reed, Jeffrey E Rossi, and Christopher P Cannon. "Acute myocardial infarction". In: *The Lancet* 389.10065 (2017), pp. 197–210.

[9] *Cleveland Clinic website*. Accessed 2023.01.22 14:01. URL: https://my.clevelandclinic.org/health/diseases/17069-heart-failure-understanding-heart-failure.

[10] *Cleveland Clinic website*. Accessed 2023.01.20 12:05. URL: https://my.clevelandclinic.org/health/diseases/4314-hypertension-high-blood-pressure.

[11] *Cleveland Clinic website*. Accessed 2023.01.20 12:00. URL: https://my.clevelandclinic.org/health/diseases/5601-stroke.

[12] *Cleveland Clinic website*. Accessed 2023.01.18 19:45. URL: https://my.clevelandclinic.org/health/diseases/7104-diabetes-mellitus-an-overview.

[13] *Cleveland Clinic website*. Accessed 2023.01.18 19:40. URL: https://my.clevelandclinic.org/health/diseases/17357-peripheral-artery-disease-pad.

[14] *Cleveland Clinic website*. Accessed 2023.01.18 19:15. URL: https://my.clevelandclinic.org/health/diseases/21656-hyperlipidemia.

[15] *Cleveland Clinic website*. Accessed 2023.01.18 19:15. URL: https://my.clevelandclinic.org/health/diseases/17837-cardiogenic-shock.

[16] *What is the Team Data Science Process?* Accessed 2023.09.04 11:09. URL: https://learn.microsoft.com/en-us/azure/architecture/data-science-process/overview.

[17] *CRISP-DM is Still the Most Popular Framework for Executing Data Science Projects*. Accessed 2023.09.04 13:10. URL: https://www.datascience-pm.com/crisp-dm-still-most-popular/.

[18] Rüdiger Wirth and Jochen Hipp. "CRISP-DM: Towards a standard process model for data mining". In: *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. Vol. 1. Manchester. 2000, pp. 29–39.

[19] Jin Huang, Jingjing Lu, and Charles X Ling. "Comparing naive Bayes, decision trees, and SVM with AUC and accuracy". In: *Third IEEE International Conference on Data Mining*. IEEE. 2003, pp. 553–556.

[20] Andrew P Bradley. "The use of the area under the ROC curve in the evaluation of machine learning algorithms". In: *Pattern recognition* 30.7 (1997), pp. 1145–1159.

[21] R R Core Team et al. "R: A language and environment for statistical computing". In: (2017). URL: `https://www.R-project.org`.

[22] *Terry Therneau, Beth Atkinson and Brian Ripley (2017). rpart: Recursive Partitioning and Regression Trees. R. Package version 4.1-11.* URL: `https://CRAN.R-project.org/package=rpart`.

[23] WN Venables and B. D. Ripley. "Statistics complements to modern applied statistics with S. Fourth edition." In: *Springer, New York* (2002).

[24] Frank E Harrell. "Regression modeling strategies". In: *Bios* 330.2018 (2017), p. 14. URL: `https://CRAN.R-project.org/package=rms`.

[25] Andy Liaw, Matthew Wiener, et al. "Classification and regression by randomForest". In: *R news* 2.3 (2002), pp. 18–22.

[26] Brandon Greenwell, Bradley Boehmke, Jay Cunningham, and GBM Developers. "gbm: Generalized boosted models". In: *R package version* 2.5 (2019), pp. 37–40. URL: `https://CRAN.R-project.org/package=gbm`.

[27] DM Zachary and EK Jared. “caretEnsemble: Ensembles of caret models”. In: *R package version 2.0* 1 (2019). URL: `https://CRAN.R-project.org/package=caretEnsemble`.

[28] Max Kuhn. “Caret: classification and regression training”. In: *Astrophysics Source Code Library* (2017), ascl–1505. URL: `https://CRAN.R-project.org/package=caret`.

[29] Stef Van Buuren and Karin Groothuis-Oudshoorn. “mice: Multivariate imputation by chained equations in R”. In: *Journal of statistical software* 45 (2011), pp. 1–67. URL: `http://www.jstatsoft.org/v45/i03/..`

[30] R Core Team. “R language definition”. In: *Vienna, Austria: R foundation for statistical computing* 3.1 (2000).

[31] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC. Boston, MA, 2020. URL: `http://www.rstudio.com/`.

[32] Copenhagen 1971 WHO. “Ischaemic Heart Disease registers. Regional Office for Europe”. In: (1971).

[33] IVÁN GYÁRFÁS. “Az akut myocardialis infarctus Délpest lakosságában”. PhD thesis. 1977.

[34] Iván Gyárfás, András Jánosi, and Péter Ofner. “Myocardial Infarction Registry conducted forty years ago in Hungary”. In: *Orvosi hetilap* 152.20 (2011), pp. 793–796.

[35] A. Janosi, P. Ofner, A. Kovacs, and P. Brunner. “Infarctus Regiszter egykor es ma”. In: *Orvosi hetilap* (2010).

[36] A. Janosi, L Keller, Gyarfas I, Gara I., Zimmermann Zs, Panyko M., Peto A., Hernadi E., and I. Regula. “Infarctus Regiszter - egy evtized elemzese”. In: *Orvosi hetilap* (1981).

[37] András Jánosi, Péter Ofner, Gabriella Branyickiné Géczy, and Péter Polgár. "Incidence of myocardial infarction in Hungary. Population study in five districts of Budapest and Szabolcs-Szatmár-Bereg county". In: *Orvosi hetilap* 154.28 (2013), pp. 1106–1110.

[38] András Jánosi, Péter Ofner, Gabriella Branyickiné Géczy, and Péter Polgár. "The Healthcare Law and the Law on the Management and Protection of Personal Data". In: *Orvosi hetilap* 10 (2013), p. 218.

[39] András Jánosi, Péter Ofner, Béla Merkely, Péter Polgár, Péter Andréka, Károly Zámolyi, Róbert Gábor Kiss, János Tomcsányi, Zoltán László, András Vértes, et al. "Myocardial Infarction Registry–2010. Feasibility and first results". In: *Orvosi Hetilap* 152.32 (2011), pp. 1278–1283.

[40] András Jánosi, Péter Ofner, and László Voith. "Clinical presentation and hospital outcome of patients with ST-elevation myocardial infarction–Hungarian Myocardial Infarction Registry data". In: *Orvosi hetilap* 153.37 (2012), pp. 1465–1468.

[41] András Jánosi, Péter Ofner, Béla Merkely, Péter Polgár, Károly Zámolyi, Róbert Gábor Kiss, István Édes, Kálmán Csapó, Lajos Nagy, Géza Lupkovics, et al. "Short and long term prognosis of patients with myocardial infarction. Hungarian Myocardial Infarction Registry". In: *Orvosi hetilap* 154.33 (2013), pp. 1297–1302.

[42] András Jánosi, Péter Ofner, Tamás Forster, István Édes, Kálmán Tóth, and Béla Merkely. "Clinical characteristics, hospital care, and prognosis of patients with ST elevation myocardial infarction: Hungarian Myocardial Infarction Registry". In: *European Heart Journal Supplements* 16.suppl_A (2014), A12–A15.

[43] K Szummer, A Janosi, T Breuer, P Ofner, J Sundstrom, and T Jernberg. “Comparison of 30-day outcome in ST-elevation myocardial infarction patients treated in Sweden or Hungary: results from SWEDEHEART and the Hungarian myocardial infarction registry”. In: *European Heart Journal* 34.suppl_1 (2013).

[44] András Jánosi, Péter Ofner, Dániel Simkovits, and Tamás Ferenci. “Prognostic significance of the culprit vessel in patients with ST-elevation myocardial infarction treated with primary coronary intervention”. In: *Orvosi hetilap* 157.32 (2016), pp. 1282–1288.

[45] András Komócsi, Mihály Simon, Béla Merkely, Tibor Szűk, Róbert G Kiss, Dániel Aradi, Zoltán Ruzsa, Péter Andrássy, Lajos Nagy, Géza Lupkovics, et al. “Underuse of coronary intervention and its impact on mortality in the elderly with myocardial infarction. A propensity-matched analysis from the Hungarian Myocardial Infarction Registry”. In: *International journal of cardiology* 214 (2016), pp. 485–490.

[46] András Jánosi, Péter Ofner, Fahmi Al-Maktari, Erik Hajkó, Krisztina Hati, Zoltán Járai, Mihály Józan-Jilling, Róbert Gábor Kiss, Gerda Lóczi, Géza Lupkovics, et al. “Patient care of patients with myocardial infarction in Hungary. Analysis of National Myocardial Infarction Registry data collecting in 2015”. In: *Orvosi Hetilap* 158.3 (2017), pp. 90–93.

[47] Emily Herrett, Liam Smeeth, Lynne Walker, Clive Weston, MINAP Academic Group, et al. “The myocardial ischaemia national audit project (MINAP)”. In: *Heart* 96.16 (2010), p. 1264.

[48] *Myocardial Ischaemia National Audit Project - How the NHS cares for patients with heart attack (Annual Public Report - April 2014 – March 2015).* Accessed 2 April 2017. URL: `http://website-url.com`.

[49] *Myocardial Ischaemia National Audit Project - How the NHS cares for patients with heart attack (Annual Public Report - April 2014 – March 2015).* Accessed 2 April 2017. URL: `http://website-url.com`.

[50] Tomas Jernberg, Mona F Attebring, Kristina Hambraeus, Torbjorn Ivert, Stefan James, Anders Jeppsson, Bo Lagerqvist, Bertil Lindahl, Ulf Stenestrand, and Lars Wallentin. "The Swedish Web-system for enhancement and development of evidence-based care in heart disease evaluated according to recommended therapies (SWEDEHEART)". In: *Heart* 96.20 (2010), pp. 1617–1621.

[51] *Paul Erne. History and highlights of the AMIS Plus Registry. AMIS Plus Sponsors & Participants' Meeting. 5 March 2015, Berne.* Accessed May 1 2017. URL: `http://www.amis-plus.ch/pdf/presentations/History%5C_Highlights%5C_%5C_20150305%5C_Erne.pdf`.

[52] *Royal College of Physicians. The Clinical Effectiveness and Evaluation Unit of the Royal College of Physicians. National data quality assessment 2008: validation and data quality exercise.* 2008.

[53] *Dragana Radovanovic. Current Status of the AMIS Plus Project. 5 March 2015, Berne.* Accessed May 1, 2017. URL: `http://www.amis-plus.ch/pdf/presentations/Status%5C_20150305%5C_Radovanovic.pdf`.

[54] Yang Liu and Anindya De. "Multiple imputation by fully conditional specification for dealing with missing data in a large epidemiologic study". In: *International journal of statistics in medical research* 4.3 (2015), p. 287.

[55] Craig K Enders. "A primer on maximum likelihood algorithms available for use with missing data". In: *Structural Equation Modeling* 8.1 (2001), pp. 128–141.

[56] Shigeyuki Oba, Masa-aki Sato, Ichiro Takemasa, Morito Monden, Ken-ichi Matsubara, and Shin Ishii. “A Bayesian missing value estimation method for gene expression profile data”. In: *Bioinformatics* 19.16 (2003), pp. 2088–2096.

[57] Donald B Rubin. *Multiple imputation for nonresponse in surveys*. Vol. 81. John Wiley & Sons, 2004.

[58] Peter C Austin. “A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality”. In: *Statistics in medicine* 26.15 (2007), pp. 2937–2957.

[59] Peter C Austin, Douglas S Lee, Ewout W Steyerberg, and Jack V Tu. “Regression trees for predicting mortality in patients with cardiovascular disease: what improvement is achieved by using ensemble-based methods?” In: *Biometrical journal* 54.5 (2012), pp. 657–673.

[60] Douglas S Lee, Peter C Austin, Jean L Rouleau, Peter P Liu, David Naimark, and Jack V Tu. “Predicting mortality among patients hospitalized for heart failure: derivation and validation of a clinical model”. In: *Jama* 290.19 (2003), pp. 2581–2587.

[61] Jack V Tu, Peter C Austin, Randy Walld, Leslie Roos, Jean Agras, and Kathryn M McDonald. “Development and validation of the Ontario acute myocardial infarction mortality prediction rules”. In: *Journal of the American College of Cardiology* 37.4 (2001), pp. 992–997.

[62] Chee Tang Chin, Anita Y Chen, Tracy Y Wang, Karen P Alexander, Robin Mathews, John S Rumsfeld, Christopher P Cannon, Gregg C Fonarow, Eric D Peterson, and Matthew T Roe. “Risk adjustment for in-hospital mortality of contemporary patients with acute myocardial infarction: The Acute Coronary

Treatment and Intervention Outcomes Network (ACTION) Registry®–Get With The Guidelines (GWTG)™ acute myocardial infarction mortality model and risk score". In: *American heart journal* 161.1 (2011), pp. 113–122.

[63] Gilles Clermont, Derek C Angus, Stephen M DiRusso, Martin Griffin, and Walter T Linde-Zwirble. "Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models". In: *Critical care medicine* 29.2 (2001), pp. 291–296.

[64] Johan Nilsson, Mattias Ohlsson, Lars Thulin, Peter Höglund, Samer AM Nashef, and Johan Brandt. "Risk factor identification and mortality prediction in cardiac surgery using artificial neural networks". In: *The Journal of thoracic and cardiovascular surgery* 132.1 (2006), pp. 12–19.

[65] Richard K Orr. "Use of a probabilistic neural network to estimate the risk of mortality after cardiac surgery". In: *Medical Decision Making* 17.2 (1997), pp. 178–185.

[66] Reinhard Voss, Paul Cullen, Helmut Schulte, and Gerd Assmann. "Prediction of risk of coronary events in middle-aged men in the Prospective Cardiovascular Münster Study (PROCAM) using neural networks". In: *International journal of epidemiology* 31.6 (2002), pp. 1253–1262.

[67] U Rajendra Acharya, Hamido Fujita, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, Muhammad Adam, and Ru San Tan. "Deep convolutional neural network for the automated diagnosis of congestive heart failure using ECG signals". In: *Applied Intelligence* 49 (2019), pp. 16–27.

[68] Yaochu Jin. *Multi-objective machine learning*. Vol. 16. Springer Science & Business Media, 2006, pp. 336–337.

[69] Hamido Fujita and Dalibor Cimr. "Computer aided detection for fibrillations and flutters using deep convolutional neural network". In: *Information Sciences* 486 (2019), pp. 231–239.

[70] Shigeo Yamamura. "Clinical application of artificial neural network (ANN) modeling to predict pharmacokinetic parameters of severely ill patients". In: *Advanced Drug Delivery Reviews* 55.9 (2003), pp. 1233–1251.

[71] Roland N Dickerson, Darius L Mason, Martin A Croce, Gayle Minard, and Rex O Brown. "Evaluation of an artificial neural network to predict urea nitrogen appearance for critically ill multiple-trauma patients". In: *Journal of Parenteral and Enteral Nutrition* 29.6 (2005), pp. 429–435.

[72] Joseph Brian Adams and Yijin Wert. "Logistic and neural network models for predicting a hospital admission". In: *Journal of Applied Statistics* 32.8 (2005), pp. 861–869.

[73] Hsin-Yi Li, Rong-Guan Yeh, Yu-Che Lin, Lo-Yi Lin, Jing Zhao, Chih-Min Lin, and Imre J Rudas. "Medical sample classifier design using fuzzy cerebellar model neural networks". In: *Acta polytechnica Hungarica* 13.6 (2016), pp. 7–24.

[74] Leo Breiman. "Random forests". In: *Machine learning* 45 (2001), pp. 5–32.

[75] Antonia Saravanou, Clemens Noelke, Nicholas Huntington, Dolores Acevedo-Garcia, and Dimitrios Gunopulos. "Infant mortality prediction using birth certificate data". In: *arXiv preprint arXiv:1907.08968* (2019).

[76] Xinsong Du, Jae Min, Chintan P Shah, Rohit Bishnoi, William R Hogan, and Dominick J Lemas. "Predicting in-hospital mortality of patients with febrile neutropenia using machine learning models". In: *International journal of medical informatics* 139 (2020), p. 104140.

[77] Li Yan, Hai-Tao Zhang, Jorge Goncalves, Yang Xiao, Maolin Wang, Yuqi Guo, Chuan Sun, Xiuchuan Tang, Liang Jing, Mingyang Zhang, et al. "An interpretable mortality prediction model for COVID-19 patients". In: *Nature machine intelligence* 2.5 (2020), pp. 283–288.

[78] Reza Sadeghi, Tanvi Banerjee, and William Romine. "Early hospital mortality prediction using vital signals". In: *Smart Health* 9 (2018), pp. 265–274.

[79] Javier Esteban Garcıa-Gallo, NJ Fonseca-Ruiz, LA Celi, and JF Duitama-Muñoz. "A machine learning-based model for 1-year mortality prediction in patients admitted to an Intensive Care Unit with a diagnosis of sepsis". In: *Medicina intensiva* 44.3 (2020), pp. 160–170.

[80] Marco Mamprin, Svitlana Zinger, PHN de With, JM Zelis, and PAL Tonino. "Gradient boosting on decision trees for mortality prediction in transcatheter aortic valve implantation". In: *Proceedings of the 2020 10th International Conference on Biomedical Engineering and Technology*. 2020, pp. 325–329.

[81] Jose F Velez-Serrano, Daniel Velez-Serrano, Valentin Hernandez-Barrera, Rodrigo Jimenez-Garcia, Ana Lopez de Andres, Pilar Carrasco Garrido, and Alejandro Alvaro-Meca. "Prediction of in-hospital mortality after pancreatic resection in pancreatic cancer patients: A boosting approach via a population-based study using health administrative data". In: *PloS one* 12.6 (2017), e0178757.

[82] Suveen Angraal, Bobak J Mortazavi, Aakriti Gupta, Rohan Khera, Tariq Ahmad, Nihar R Desai, Daniel L Jacoby, Frederick A Masoudi, John A Spertus, and Harlan M Krumholz. "Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction". In: *JACC: Heart Failure* 8.1 (2020), pp. 12–21.

[83] Roni Shouval, Amir Hadanny, Nir Shlomo, Zaza Iakobishvili, Ron Unger, Doron Zahger, Ronny Alcalai, Shaul Atar, Shmuel Gottlieb, Shlomi Matet-

zky, et al. “Machine learning for prediction of 30-day mortality after ST elevation myocardial infraction: An Acute Coronary Syndrome Israeli Survey data mining study”. In: *International journal of cardiology* 246 (2017), pp. 7–13.

[84] Jérôme Allyn, Nicolas Allou, Pascal Augustin, Ivan Philip, Olivier Martinet, Myriem Belghiti, Sophie Provenchere, Philippe Montravers, and Cyril Ferdynus. “A comparison of a machine learning model with EuroSCORE II in predicting mortality after elective cardiac surgery: a decision curve analysis”. In: *PloS one* 12.1 (2017), e0169772.

[85] Syed Waseem Abbas Sherazi, Yu Jun Jeong, Moon Hyun Jae, Jang-Whan Bae, and Jong Yun Lee. “A machine learning–based 1-year mortality prediction model after hospital discharge for clinical patients with acute coronary syndrome”. In: *Health informatics journal* 26.2 (2020), pp. 1289–1304.

[86] Yoav Freund, Robert E Schapire, et al. “Experiments with a new boosting algorithm”. In: *icml*. Vol. 96. Citeseer. 1996, pp. 148–156.

[87] Jerome H Friedman. “Greedy function approximation: a gradient boosting machine”. In: *Annals of statistics* (2001), pp. 1189–1232.

[88] C Beulah Christalin Latha and S Carolin Jeeva. “Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques”. In: *Informatics in Medicine Unlocked* 16 (2019), p. 100203.

[89] Peter C Austin and Douglas S Lee. “Boosted classification trees result in minor to modest improvement in the accuracy in classifying cardiovascular outcomes compared to conventional classification trees”. In: *American journal of cardiovascular disease* 1.1 (2011), p. 1.

[90] Peter C Austin, Douglas S Lee, Ewout W Steyerberg, and Jack V Tu. “Regression trees for predicting mortality in patients with cardiovascular disease: what improvement is achieved by using ensemble-based methods?” In: *Biometrical journal* 54.5 (2012), pp. 657–673.

[91] Resul Das, Ibrahim Turkoglu, and Abdulkadir Sengur. “Effective diagnosis of heart disease through neural networks ensembles”. In: *Expert systems with applications* 36.4 (2009), pp. 7675–7680.

[92] Devika Subramanian, Venkataraman Subramanian, Anita Deswal, and Douglas L Mann. “New predictive models of heart failure mortality using time-series measurements and ensemble models”. In: *Circulation: Heart Failure* 4.4 (2011), pp. 456–462.

[93] Matej Petkovic, Saso Dzeroski, and Dragi Kocev. “Feature ranking for hierarchical multi-label classification with tree ensemble methods”. In: *Acta Polytechnica Hungarica* 17.10 (2020), pp. 129–148.

[94] Pierre Geurts, Damien Ernst, and Louis Wehenkel. “Extremely randomized trees”. In: *Machine learning* 63 (2006), pp. 3–42.

[95] János Tóth, Róbert Tornai, Imre Labancz, and András Hajdu. “Efficient visualization for an ensemble-based system”. In: *Acta Polytechnica Hungarica* 16.2 (2019), pp. 59–75.

[96] Bernard L Welch. “The generalization of ‘STUDENT’S’problem when several different population varlances are involved”. In: *Biometrika* 34.1-2 (1947), pp. 28–35.

[97] *What is PHP? Official page of PHP*. Accessed 2023.01.15 09:05. URL: https://www.php.net/manual/en/intro-whatis.php.

[98] *Usage statistics of PHP Version 7 for websites*. Accessed 2023.02.10 13:10. URL: https://w3techs.com/technologies/details/pl-php/7.

Own Publications Pertaining to Theses

[P-1] Péter Piros, Rita Fleiner, Tamás Ferenci, Péter Andréka, Hamido Fujita, Péter Ofner, Levente Kovács, and András Jánosi. "An overview of myocardial infarction registries and results from the Hungarian Myocardial Infarction Registry". In: *IOS Press* 297 (2017), p. 312. DOI: 10.3233/978-1-61499-800-6-312.

[P-2] László Beinrohr, Eszter Kail, Péter Piros, Erzsébet Tóth, Rita Fleiner, and Krasimir Kolev. "Anatomy of a Data Science Software Toolkit That Uses Machine Learning to Aid Bench-to-Bedside Medical Research—With Essential Concepts of Data Mining and Analysis Explained". In: *Applied Sciences* 11.24 (2021), p. 12135. DOI: 10.3390/app112412135.

[P-3] Péter Piros, Tamás Ferenci, Rita Fleiner, Péter Andréka, Hamido Fujita, László Főző, Levente Kovács, and András Jánosi. "Comparing machine learning and regression models for mortality prediction based on the Hungarian Myocardial Infarction Registry". In: *Knowledge-Based Systems* 179 (2019), pp. 1–7. DOI: 10.1016/j.knosys.2019.04.027.

[P-4] Péter Piros, Rita Fleiner, and Levente Kovács. "Random Forest-based predictive modelling on Hungarian Myocardial Infarction Registry". In: *2020 IEEE 15th International Conference of System of Systems Engineering (SoSE)*. IEEE. 2020, pp. 525–530. DOI: 10.1109/SoSE50414.2020.9130476.

[P-5] Péter Piros, Rita Fleiner, and Levente Kovács. "Finding improved predictive models with Generalized Boosted Models on Hungarian Myocardial Infarction Registry". In: *2020 IEEE 20th International Symposium on Computational Intelligence and Informatics (CINTI)*. IEEE. 2020, pp. 179–184. DOI: 10.1109/CINTI51262.2020.9305814.

[P-6] Péter Piros, Rita Fleiner, András Jánosi, and Levente Kovács. "Further evolution of mortality prediction with ensemble-based models on Hungarian Myocardial Infarction Registry". In: *Acta Polytechnica Hungarica* 20.4 (2023). DOI: 10.12700/APH.20.4.2023.4.7.

[P-7] Péter Piros, Rita Fleiner, Tamás Ferenci, Levente Kovács, and András Jánosi. "Comparing the predictive power of decision tree models with different tuning approaches on Hungarian Myocardial Infarction Registry". In: *2019 IEEE 13th International Symposium on Applied Computational Intelligence and Informatics (SACI)*. IEEE. 2019, pp. 326–331. DOI: 10.1109/SACI46893.2019.9111525.

Own Publications Not Pertaining to Theses

[Px-1] Péter Piros, Rita Fleiner, and Levente Kovács. "Linked data generation for courses and events at Óbuda University". In: *2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*. IEEE. 2017, pp. 000253–000258. DOI: 10.1109/SAMI.2017.7880313.

[Px-2] Rita Fleiner, Barnabás Szász, and Péter Piros. "Indoor navigation Linked Data at Obuda University". In: *2016 IEEE 11th International Symposium on Applied Computational Intelligence and Informatics (SACI)*. IEEE. 2016, pp. 25–30. DOI: 10.1109/SACI.2016.7507377.